KNOTS KNOTES

JUSTIN ROBERTS

Contents

Date: January 27, 2015.

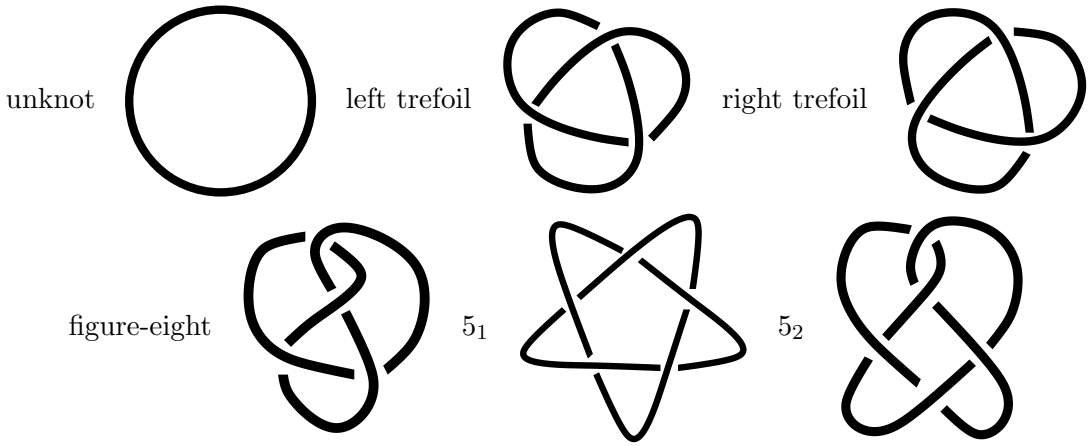
1. Motivation, basic definitions and questions

This section just attempts to give an outline of what is ahead: the objects of study, the natural questions (and some of their answers), some of the basic definitions and properties, and many examples of knots.

1.1. **Basic definitions.**

Definition 1.1.1 (Provisional)**.** A *knot* is a closed loop of string in $\mathbb{R}^3$; two knots are *equivalent* (the symbol $\cong$ is used) if one can be wiggled around, stretched, tangled and untangled until it coincides with the other. Cutting and rejoining is *not* allowed.

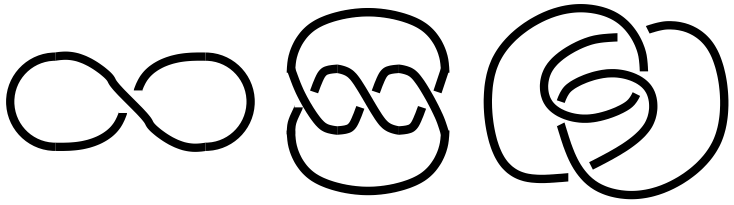
Example 1.1.2.



Remark 1.1.3. Some knots have historical or descriptive names, but most are referred to by their numbers in the standard tables. For example $5_1, 5_2$ refer to the first and second of the two 5-crossing knots, but this ordering is completely arbitrary, being inherited from the earliest tables compiled.

Remark 1.1.4. Actually the pictures above are *knot diagrams*, that is planar representations (projections) of the three-dimensional object, with additional information (over/under-crossing information) recorded by means of the breaks in the arcs. Such two-dimensional representations are much easier to work with, but they are in a sense artificial; knot theory is concerned primarily with three-dimensional topology.

Remark 1.1.5. Any knot may be represented by many different diagrams, for example here are alternative pictures of the unknot, right trefoil and figure-eight knot. (Convince yourself of the latter using string or careful redrawing of pictures!) (Warning: slightly confusingly, the picture on the left here is *not* what we mean when we say "figure-eight knot"!)



1.2. **Basic questions.**

Question 1.2.1. Mathematically, how do we go about formalising the definitions of knot and equivalence?

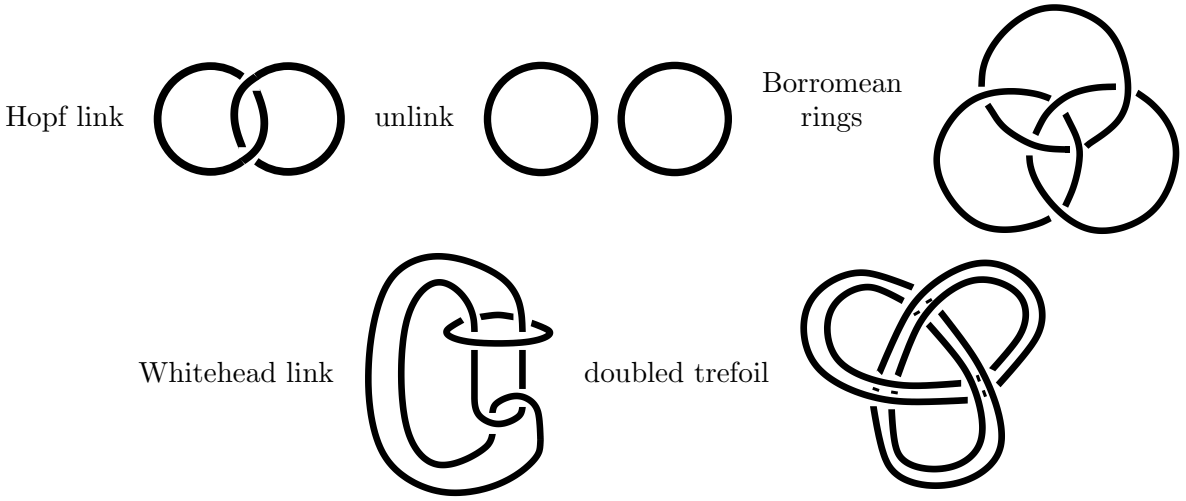
Question 1.2.2. How might we prove *inequivalence* of knots? To show two knots are equivalent, we can simply try wiggling one of them until we succeed in making it look like the other: this is a proof. On the other hand, wiggling a trefoil around for an hour or so and failing to make it look like the unknot is *not* a proof that they are distinct, merely inconclusive evidence. We need to work much harder to prove this. One of the first tasks in the course will be to show that the trefoil is inequivalent to the unknot (i.e. that it is *non-trivial* or *knotted*).

Question 1.2.3. Can one produce a table of the simplest knot types (a *knot type* means an equivalence class of knots, in other words a *topological* as opposed to *geometrical* knot: often we will simply call it "a knot"). "Simplest" is clearly something we will need to define: how should one measure the complexity of knots?

Although knots have a long history in Celtic and Islamic art, sailing etc., and were first studied mathematically by Gauss in the 1800s, it was not until the 1870s that there was a serious attempt to produce a knot table. James Clerk Maxwell, William Thompson (Lord Kelvin) and Peter Tait (the Professor of maths at Edinburgh, and inventor of the dimples in a golf ball) began to think that "knotted vortex tubes" might provide an explanation of the periodic table; Tait compiled some tables and gave names to many of the basic properties of knots, and so did Kirkman and Little. It was not until Poincaré had formalised the modern theory of topology around about 1900 that Reidemeister and Alexander (around about 1930) were able to make significant progress in knot theory. Knot theory was a respectable if not very dynamic branch of topology until the discovery of the Jones polynomial (1984) and its connections with physics (specifically, quantum field theory, via the work of Witten). Since then it has been "trendy" (this is a mixed blessing!) It even has some concrete applications in the study of enzymes acting on DNA strands. See Adams' "Knot book" for further historical information.

Definition 1.2.4. A *link* is simply a collection of (finitely-many) disjoint closed loops of string in $\mathbb{R}^3$; each loop is called a *component* of the link. Equivalence is defined in the obvious way. A knot is therefore just a one-component link.

Example 1.2.5. Some links. Note that the individual components may or may not be unknots. The Borromean rings have the interesting property that removing any one component means the remaining two separate: the entanglement of the rings is dependent on *all three components* at the same time.



Exercise 1.2.6. The Borromean rings are a 3-component example of a *Brunnian link*, which is a link such that deletion of any one component leaves the rest unlinked. Find a 4-component Brunnian link. (Slightly cryptic remark: it's quite possible that the first picture you try won't work. If so, can you actually prove that it won't work?)

Definition 1.2.7. The *crossing number* $c(K)$ of a knot K is the minimal number of crossings in any diagram of that knot. (This is a natural measure of complexity.) A *minimal* diagram of K is one with $c(K)$ crossings.

Example 1.2.8. The unknot has crossing number 0. There are no non-trivial knots with crossing numbers 1 or 2: one can prove this by enumerating all possible *diagrams* with one or two crossings, and seeing that they are either unknots or links with more than one component. Clearly the trefoil has crossing number less than or equal to 3, since we can draw it with three crossings. The question is whether it could be smaller than 3. If this were so it would have to be equivalent to an unknot. So proving that the crossing number really is 3 is equivalent to proving that the trefoil is non-trivial.

Exercise 1.2.9. Write a proof that there are indeed no knots with crossing number 1 or 2 (just draw the possible diagrams and check).

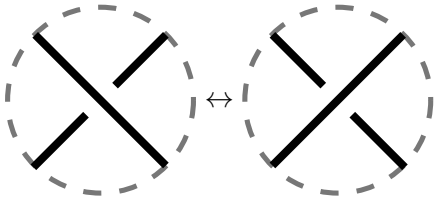
Exercise 1.2.10. Prove similarly that the only knots with crossing number 3 are the two trefoils (of course we don't know they are distinct yet!)

Remark 1.2.11. Nowadays there are tables of knots up to about 16 crossings (computer power is the only limit in computation). There are tens of thousands of these.

1.3. **Operations on knots.**

Much of what is discussed here applies to links of more than one component, but these generalisations should be obvious, and it is more convenient to talk primarily about knots.

Definition 1.3.1. The *mirror-image* $\bar{K}$ of a knot K is obtained by reflecting it in a plane in $\mathbb{R}^3$. (Convince yourself that all such reflections are equivalent!) It may also be defined given a diagram D of K: one simply exchanges all the crossings of D.



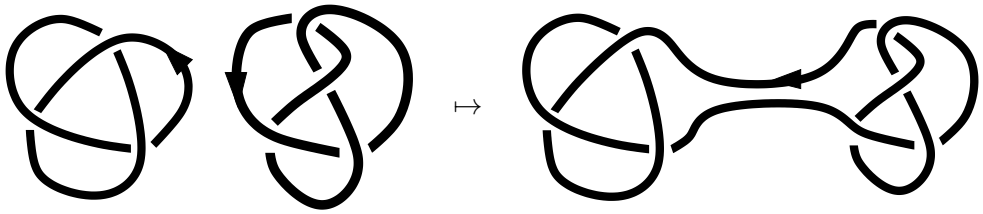
This is evident if one considers reflecting in the plane of the page.

Definition 1.3.2. A knot is called *amphichiral* if it is equivalent to its own mirror-image. How might one detect amphichirality? The trefoil is in fact not amphichiral (we will prove this later), whilst the figure-eight is (try this with string!).

Definition 1.3.3. An *oriented* knot is one with a chosen direction or "arrow" of circulation along the string. Under equivalence (wiggling) this direction is carried along as well, so one may talk about *equivalence* (meaning *orientation-preserving equivalence*) of oriented knots.

Definition 1.3.4. The *reverse* rK of an oriented knot K is simply the same knot with the opposite orientation. One may also define the *inverse* $r\bar{K}$ as the composition of reversal and mirror-image. By analogy with amphichirality, we have a notion of a knot being *reversible* or *invertible* if it is equivalent to its reverse or inverse. Reversibility is very difficult to detect; the knot 8_{17} is the first non-reversible one (discovered by Trotter in the 60s).

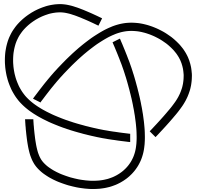
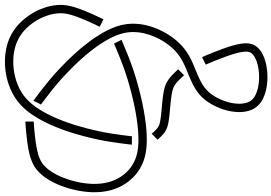
Definition 1.3.5. If K_1, K_2 are *oriented* knots, one may form their *connect-sum* $K_1 \# K_2$ by removing a little arc from each and splicing up the ends to get a single component, making sure the orientations glue to get a consistent orientation on the result. (If the knots aren't oriented, there is a choice of two ways of splicing, which may sometimes result in different knots!)



This operation behaves rather like multiplication on the positive integers. It is a commutative operation with the unknot as identity. A natural question is whether there is an inverse; could one somehow cancel out the knottedness of a knot K by connect-summing it with some other knot? This seems implausible, and we will prove it false. Thus knots form a *semigroup* under connect-sum. In this semigroup, just as in the postive integers under multiplication, there is a notion of *prime factorisation*, which we will study later.

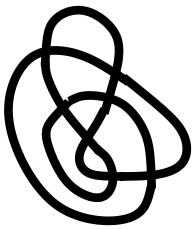
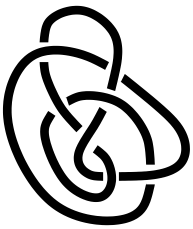
1.4. **Alternating knots.**

Definition 1.4.1. An *alternating diagram* D of a knot K is a diagram such passes alternately over and under crossings, when circling completely around the diagram from some arbitrary starting point. An *alternating knot* K is one which possesses *some* alternating diagram. (It will always possess non-alternating diagrams too, but this is irrelevant.) The trefoil is therefore alternating.

alternating diagram  non-alternating diagram 

Question 1.4.2. Hard research problem (nobody has any idea at present): give an intrinsically three-dimensional definition of an alternating knot (i.e. without mentioning diagrams)!

If one wants to draw a knot at random, the easiest method is simply to draw in pencil a random projection in the plane (just an immersion of the circle which intersects itself only in transverse double points) and then rub out a pair of little arcs near each double point to show which arc goes over at that point – clearly there is lots of choice of how to do this. A particularly "sensible" way of doing it is to start from some point on the curve and circle around it, *imposing* alternation of crossings.

projection  $\mapsto$ alternating diagram 

Exercise 1.4.3. Why does this "alternation" procedure actually work? As you move around the diagram, you pass through each crossing twice. When you pass it the first time, alternation forces you to choose which arc goes over and which goes under. When you return the second time, the existing choice always seems to be consistent with what you now need it to be. Why?

Alternating diagrams like this always seem to be "really knotted"; one may ask, as Tait did:

Question 1.4.4. Is every alternating diagram minimal? In particular, does every non-trivial alternating diagram represent a non-trivial knot?

The answer turns out to be (with a minor qualification) yes, as we will prove with the aid of the Jones polynomial (this was only proved in 1985).

Remark 1.4.5. All the simplest knots are alternating. The first non-alternating one is 8_{19} in the tables.

1.5. **Unknotting number.**

If one repeats the "random knot" construction above but puts in the crossings so that the first time one reaches any given crossing one goes *over* (one will eventually come back to it on the underpass), one produces mainly unknots. In fact there is always a way of assigning the crossings so that the result *is* an unknot. This means that given any knot diagram, it is possible to turn it into a diagram of the unknot simply by changing some of its crossings.

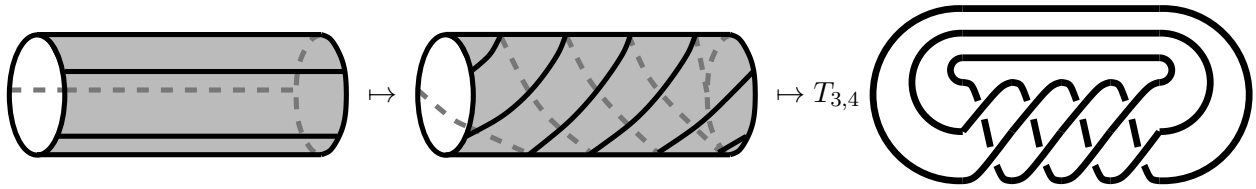
Definition 1.5.1. The *unknotting number* $u(K)$ of a knot K is the minimum, over all diagrams D of K, of the minimal number of crossing changes required to turn D into a diagram of the unknot.

In other words, if one is allowed to let the string of the knot pass through itself, one can clearly reduce it to the unknot: the question is how many times one needs to let it cross itself in this way. The unknot is clearly the only knot with unknotting number $u = 0$. In fact the trefoil has $u = 1$ and the knot 5_1 has $u = 2$. In each case one may obtain an *upper* bound simply by exhibiting a diagram and a set of unknotting crossings, but the *lower* bound is much harder. (Proving that the unknotting number of the trefoil is not zero is equivalent to proving it distinct from the unknot: proving that $u(5_1) > 1$ is even harder.)

1.6. **Further examples of knots and links.**

There are many ways of creating whole families of knots or links with similar properties. These can be useful as examples, counterexamples, tests of conjectures, and in connection with other topics.

Example 1.6.1. *Torus links* are produced by choosing a pair of integers p and q, with p positive; forming a cylinder with p strings running along it, twisting it up through "q/p full twists" (the sign of q determines the direction of twist) and gluing its ends together to form an unknotted torus in $\mathbb{R}^3$. The torus is irrelevant — one is only interested in the resulting link $T_{p,q}$ formed from the strands drawn on its surface — but it certainly helps in visualising the link.

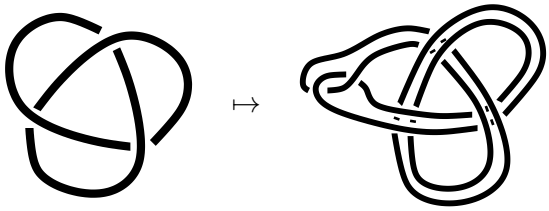


The trefoil can be seen as $T_{2,3}$ and the knot 5_1 as $T_{2,5}$. In fact $T_{3,4}$ is equivalent to the knot 8_{19}, which is the first non-alternating knot in the tables.

Exercise 1.6.2. How many components does the torus link $T_{p,q}$ have? Show in particular that it is a knot if and only if p, q are coprime.

Exercise 1.6.3. Give an upper bound for the crossing number of $T_{p,q}$. Give the best bounds you can on the crossing numbers and unknotting numbers of the family of $(2, q)$ torus links.

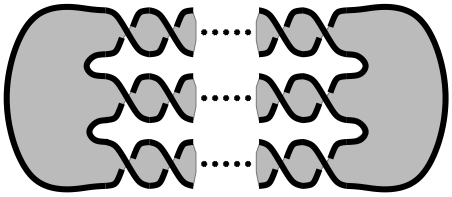
Example 1.6.4. Any knot may be *Whitehead* or *twisted doubled*: one replaces the knot by two parallel copies (there is a degree of freedom in how many times one twists around the other) and then adds a "clasp" to join the resulting two components together (in a non-unravelling way!).



Remark 1.6.5. A more general operation is the formation of a *satellite* knot by combining a knot and a *pattern*, a link in a solid torus. One simply replaces a neighbourhood of the knot by the pattern (again there is a "twisting" degree of freedom). Whitehead doubling is an example, whose pattern is shown below.

Example 1.6.6. The boundary of any "knotted surface" in $\mathbb{R}^3$ will be a knot or link. For example one may form the *pretzel links* $P_{p,q,r}$ by taking the boundary of the following surface (p, q, r denote the numbers of anticlockwise half-twists in the "bands" joining the left and right, so that negative

numbers give clockwise twists).



Exercise 1.6.7. How many components, as a function of the integers p, q, r, does the pretzel link $P_{p,q,r}$ have? In particular, when is it a knot?

1.7. **Methods.**

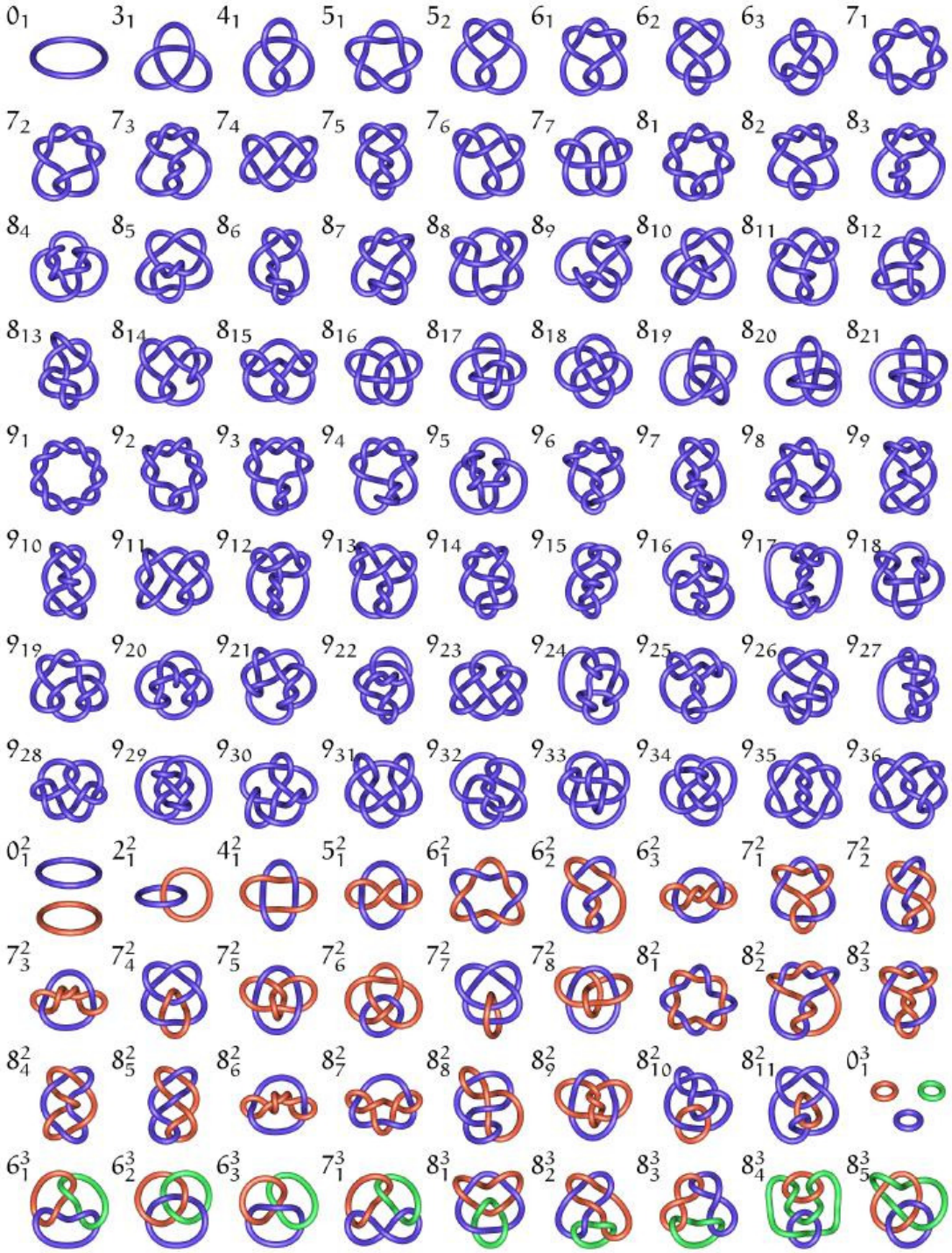
There are three main kinds of method which ware used to study knots. *Algebraic* methods are those coming from the theory of the fundamental group, algebraic topology, and so on (see section 7). *Geometric* methods are those coming from arguments that are essentially nothing more than careful and rigorous visual proofs (section 6). *Combinatorial* proofs (sections 3,4) are maybe the hardest to motivate in advance: many of them seem like miraculous tricks which just happen to work, and indeed some are very hard to explain in terms of topology. (The Jones polynomial is still a rather poorly-understood thing fifteen years after its discovery!)

Exercise 1.7.1. Imagine you wanted to write a computer program (an algorithm) to classify the different equivalence classes of knots, and produce a list extending the one on the main web page. Which bits are going to be the easy parts, and which will be the hardest parts? The idea is to try to get to grips with the fact that a computer can only perform tasks that will take a *finite amount of time and memory*. You don't have to make any attempt to write a program or even have any programming experience, you just should contemplate whether or not it seems possible! Perhaps the first thing to think about is how to represent a knot, using a finite amount of information. (The discussion of diagrams in the next section will help!)

Exercise 1.7.2. Come up with three interesting questions about knots and links to which you don't know the answer.

1.8. **A table of the simplest knots and links.**

This beautiful table came from Rob Scharein's wonderful Knotplot site. It shows the 72 simplest knots, and 36 of the simplest links.



What do we mean by simplest? Well, as you have seen already, there is only *one* knot which can be drawn using a diagram with 0, 1 or 2 crossings: the unknot. There is only *one* new knot which can be drawn using a diagram with 3 crossings: the trefoil. There is only *one* new knot which can be drawn using a 4-crossing diagram: the figure-right. But there are *two* new knots when we get to 5-crossing diagrams.... and so on.

There are several important remarks to make here:

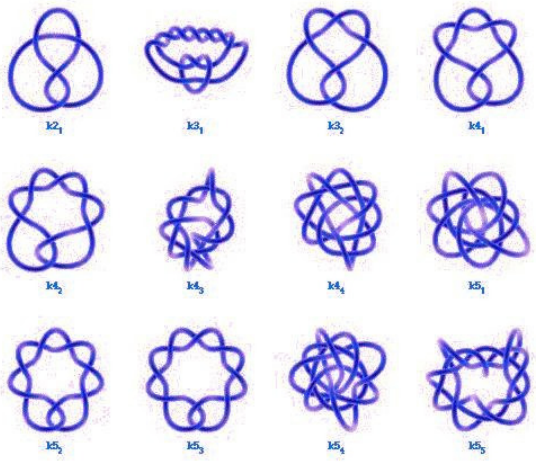
(1). The knots are given names like $5_1, 5_2$ which mean simply "the first 5-crossing knot" and "the second one". There's no sensible way to decide which one comes first! The ordering used in knot tables has been handed down over the years from tabulator to tabulator; it may not be very descriptive, but at least everybody agrees on which one is 5_1 and which 5_2. Whoever first tabulates the (let's say) 50-crossing knots (there might be millions of these) gets to choose the order and thereby "name" them $50_1, 50_2, \ldots$.

(2). The tables don't include knots which are connect-sums of simpler ones; only *prime* knots are listed. For example, the connect sum of a trefoil and a figure-eight can be drawn with 7 crossings, but we can refer to it as $3_1\#4_1$ and so there's no reason to list it separately.

(3). The tables also don't include mirror images. Most knots are distinct from their mirror images, but not all - as we mentioned, the figure-eight is an example of an amphichiral one, equivalent to its mirror. In many tables (but not this one), some kind of asterisk or symbol is placed next to the diagrams of the amphichiral knots to record their unusual property. All the other pictures therefore really describe a *pair* of distinct knots.

(4). The name 8_5^3 for a link indicates "the 5th 8-crossing link with 3 components". Similar remarks about primeness and mirror images apply to the tables for links.

(5). Although we normally take crossing number to measure the "complexity" of a knot or link, in the sense that "simple" means "drawable using a small number of crossings", there are many other quantities which can be used instead. All we really need is some kind of function taking values in an ordered set, having the property that the set of knots with complexity less than a given value is a finite set. (That way, we can imagine a finite table listing "all the knots with complexity less than c", a table which we can extend by simply increasing c.) The *hyperbolic volume* of a knot (a concept which will not be explained in these notes?) is perhaps the most interesting one. Look, on the Knotplot site, at the tables of knots ordered by hyperbolic volume: they look astonishingly different from the table above, indicating that there is not a very good correlation between crossing number and hyperbolic volume.



2. Formal definitions and Reidemeister moves

2.1. **Knots and equivalence.**

How should we formulate the definition of a knot?

Probably the most obvious thing to do is consider parametric curves in $\mathbb{R}^3$. Let $I = [0, 1]$ be the closed unit interval in $\mathbb{R}$. A continuous vector-valued function $\mathbf{x}(s)$ (that is, a triple of continuous real-valued functions $(x(s), y(s), z(s))$) with domain I defines such a curve; the continuity requirement makes sure that it is *unbroken*. If we also impose the condition $\mathbf{x}(0) = \mathbf{x}(1)$ then the initial and final point are made to coincide, so we have a parametric representation of a *closed loop*, rather than just an arc. If we require that the map $s \mapsto \mathbf{x}(s)$ is injective on the interval $[0, 1)$, then we enforce that the curve *does not intersect itself*. (We have to leave out the last point of the interval because the map is obviously is not injective if we leave it in!) These three conditions constitute a reasonable definition of a knot, which we now proceed to study further.

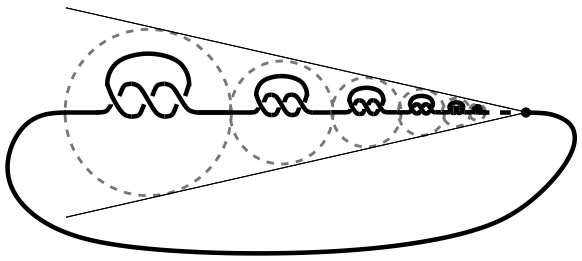
Next question: how should we formulate the notion of deformation of a knot?

Deformation is best visualised as a time-dependent process. Imagine starting at time $t = 0$ with a knot K_0, and deforming through a family of intermediate knots K_t to a final one K_1. We need to make sure this process of deformation is continuous in t. So we could consider continuous vector-valued functions of two variables $\mathbf{x}(s, t)$ for $(s, t) \in I \times I$, with the requirement that for each fixed value t, the function $s \mapsto \mathbf{x}(s, t)$ obeys the three conditions making it a knot K_t. (It's helpful to think of the parameter s as "space" or "arclength" along the knots, and t as "time" during the deformation).

Thus, we can define two knots K and K' to be equivalent if there exists a deformation running from $K_0 = K$ to $K_1 = K'$. It's easy to check that this does indeed define an equivalence relation.

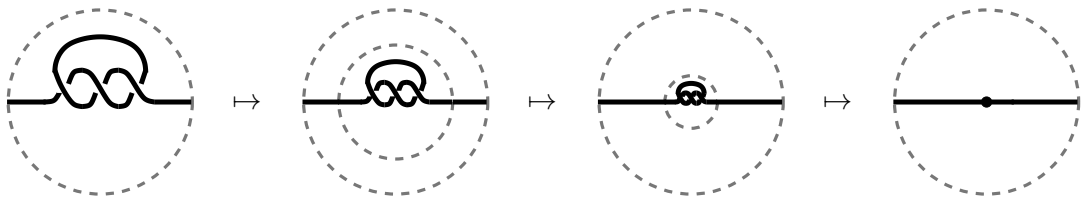
Success? Well, unfortunately no! There are two nasty problems.

Firstly, it allows "wild" knots like the one below, where we take an infinite string of trefoils, connected together and converging on a point.



It may not look as if a thing like this could be defined by continuous functions, but it can - we can use functions similar to $x \sin(\frac{1}{x})$, which is continuous, to achieve it. We can't really hope to understand things like this very well - they have "infinitely much knotting" and are just too complicated for elementary methods of study.

Secondly, surprisingly (and catastrophically), the way we have defined equivalence actually causes all knots to be equivalent to one another! "Gradually pulling the string tight" so that the knot shrinks to a point is a perfectly good continuous deformation between any knot and the unknot!



This may also be hard to believe. But you can get the idea as follows. Suppose $\mathbf{x}(s)$ is a continuous curve with domain $\mathbb{R}$, which satisfies

$$\mathbf{x}(s) = (x(s), y(s), z(s)) = (s, 0, 0) \quad \text{if } |s| \geq 1.$$

In other words it runs along the x-axis from $-\infty$ to -1, does a bit of knotting, then returns to the x-axis at $x = 1$ and runs off along it to $+\infty$. Then the function

$$(s,t) \mapsto \begin{cases} (tx(s/t,t), ty(s/t,t), tz(s/t,t)) & t \neq 0, |s| \leq t \\ (s,0,0) & |s| > t \\ (0,0,0) & t = 0, s = 0 \end{cases}$$

is a continuous deformation (check!) which at $t = 0$ is a straight line, and $t = 1$ is the original curve.

We can rectify this problem as follows. (Here I will use the language of point-set topology – don't worry if you don't understand this, as we won't use this definition either!) Forget about treating knots as *functions* (that is, parametrised loops): instead define a knot to be simply a *subset* of $\mathbb{R}^3$ which is homeomorphic to the circle S^1. We can then define two such knots to be equivalent if they are *ambient isotopic*, meaning that there exists an (orientation-preserving) homeomorphism $\mathbb{R}^3 \to \mathbb{R}^3$ carrying one to the other. This definition turns out to fix the second problem – now, not all knots are equivalent to one another – but it does not rule out wild knots.

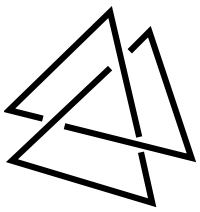
The easiest way to do that is to require that each knot should be representable as a *knotted polygon* in $\mathbb{R}^3$, that is a subset made up of *finitely many* straight-line arc segments. Obviously all "reasonable" knots can be approximated arbitrarily closely by polygonal subsets – we just have to use a very large number of tiny edges. On the other hand, the kind of "infinitely knotted" monster we saw earlier cannot be represented using a knotted polygon. We will call the class of knots representable by polygons *tame*, and from now on will only ever work with tame knots. Here is a formal definition. (I will use the standard notation $[a, b]$ for the arc-segment of $\mathbb{R}^3$ running from point a to point b.)

Definition 2.1.1. If K is a subset of $\mathbb{R}^3$ which can be written as a union of arc segments

$$K = [a_0, a_1] \cup [a_1, a_2] \cup \ldots \cup [a_{N-2}, a_{N-1}] \cup [a_{N-1}, a_0]$$

such that the segments are disjoint from one another except when consecutive (in which case they intersect in a single point) then we will say K is a *knot*.

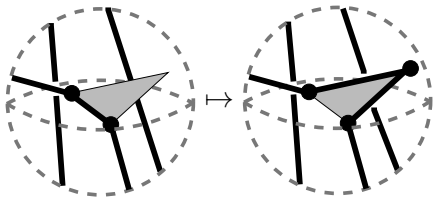
Example 2.1.2. A trefoil made from six straight edges.



Having decided that a knot should be a knotted polygon, we might as well use polygons to define the notion of equivalence too. This approach may seem a little clunky, but it does make things simple technically, and avoids any need to use point-set topology notions.

Definition 2.1.3. Suppose K is a knot in $\mathbb{R}^3$ having $[a_i, a_{i+1}]$ as one of its edges, and suppose $T = [a_i, x, a_{i+1}]$ is a closed solid triangle in $\mathbb{R}^3$ which intersects K only along the edge $[a_i, a_{i+1}]$. Then we may "slide the edge across the triangle without hitting anything": replace the edge

$[a_i, a_{i+1}]$ of K by two new edges $[a_i, x] \cup [x, a_{i+1}]$ so as to form a new knot K' with one more edge in total. Such a move is called a *Delta-*(Δ-)*move*.



A Δ-move is clearly an absolutely basic kind of deformation of knots which should be regarded as an equivalence. We will in fact define our equivalence relation to be the one "generated by Δ-moves" as follows:

Definition 2.1.4. Two knots K, J are *equivalent* (or *isotopic*) if there is a sequence of intermediate knots $K = K_0, K_1, K_2, \ldots, K_n = J$ of knots such that each pair K_i, K_{i+1} is related by a Δ-move or the reverse of a Δ-move.

This clearly defines an equivalence relation on knots (reflexive – use a sequence of length 0; symmetric – this is why we allowed the *reverse* of a *Delta*-move; transitive – compose two sequences in succession).

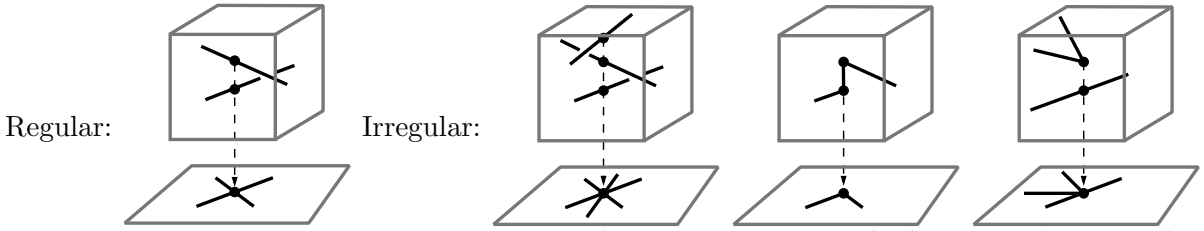
We will often confuse knots in $\mathbb{R}^3$ with their equivalence classes (or *knot types*), which are the things we are really interested in topologically. For example, the *unknot* is really the equivalence class of the boundary of a triangle (a knot with three edges), but we will often speak of "an unknot", suggesting a particular knot in $\mathbb{R}^3$ which lies in this equivalence class. Even though in this course we will always consider knots to be polygonal, we may as well carry on *drawing* them "smoothly", as long as we keep in mind that the pictures are approximations to polygonal subsets with very large numbers of tiny sides.

Exercise 2.1.5. Check that you can move one of the vertices of a knot a very small distance, keeping the rest fixed, by using two successive Δ-moves; also that you can effectively add a new vertex in the middle of any edge by using three Δ-moves.

Example 2.1.6. Any knot lying completely in a plane inside $\mathbb{R}^3$ is an unknot. This is a consequence of the "polygonal Jordan curve theorem", that any polygonal simple closed curve (that is, non-self-intersecting closed polygon) separates the plane into two pieces, one of which (the "inside") is topologically equivalent to a disc. (The proper Jordan curve theorem states that this holds for *any* (not necessarily polygonal) simple closed curve; this is a quite hard theorem of point-set topology. See Armstrong for some information about this.) By dividing the inside of the polygon into triangles in a suitable way, we can find a sequence of Δ-moves that shrinks the polygon down to a triangle. (More details on this, including a proof, when we come to talk about surfaces.)

2.2. **Projections and diagrams.**

Definition 2.2.1. If K is a knot in $\mathbb{R}^3$, its *projection* is $\pi(K) \subseteq \mathbb{R}^2$, where π is the projection along the z-axis onto the xy-plane. The projection is said to be *regular* if the preimage of a point of $\pi(K)$ consists of either one or two points of K, in the latter case neither being a vertex of K. Clearly a knot has an *irregular* projection if it has any edges parallel to the z-axis, if it has three or more points lying above each other, or any vertex lying above or below another point of K; on the other hand, a regular projection of a knot consists of a polygonal circle drawn in the plane with only "transverse double points" as self-intersections.



Definition 2.2.2. If K has a regular projection then we can define the corresponding *knot diagram* D by redrawing it with a "broken arc" near each *crossing* (place with two preimages in K) to incorporate the over/under information. If K had an irregular projection then we would not be able to easily reconstruct it from this sort of picture (consider the various cases mentioned above!) so it is important that we can find regular projections of knots easily.

Definition 2.2.3. Define an *ϵ-perturbation* of a knot K in $\mathbb{R}^3$ to be any knot K' obtained by moving each of the vertices of K a distance less than ϵ, and reconnecting them with straight edges in the same fashion as K.

Fact 2.2.4. If ϵ is chosen sufficiently small then all such ϵ-perturbations of K will be equivalent to it (though clearly very large perturbations could be utterly different!)

Fact 2.2.5. "Regular projections are generic". This means "knots which have regular projections form an open, dense set in the space of knots". Or, more precisely the following two properties:

(1). If K has an irregular projection then there exist *arbitrarily small* ϵ-perturbations K' (in particular, ones equivalent to K!) with regular projections.

(2). If K has a regular projection then *any* sufficiently small ϵ-perturbation also has a regular projection.

Thus, knots with irregular projections are very rare: the first proposition implies that if one constructed knots by randomly picking their vertices, they would be regular with probability 1. In particular, any knot with an irregular projection need only be wiggled a tiny amount in space to make its projection regular.

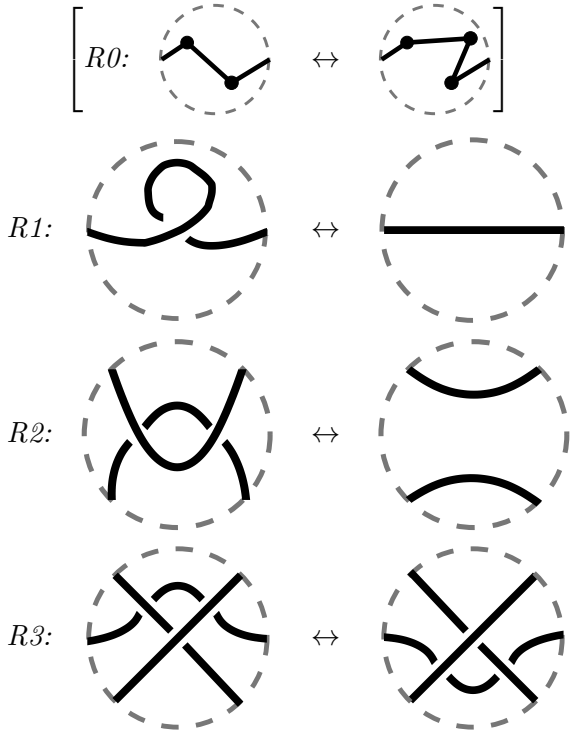
Corollary 2.2.6. *Any knot has a diagram. From a diagram one can reconstruct the knot up to equivalence. Any knot having a diagram with no crossings is an unknot.*

Proof. The first part just restates the fact above, that any knot is equivalent to one with a regular projection (and hence a diagram). The second part points out that a diagram does not reconstruct a knot in $\mathbb{R}^3$ uniquely (one doesn't know what the z-coordinates of its vertices should be, for example) but one does know the relative heights at crossings. It is a boring exercise to write a formal proof that any two knots in $\mathbb{R}^3$ having *exactly the same* diagram are equivalent by Δ-moves. The final part comes from the second and the example about the Jordan curve theorem. □

2.3. **Reidemeister moves.**

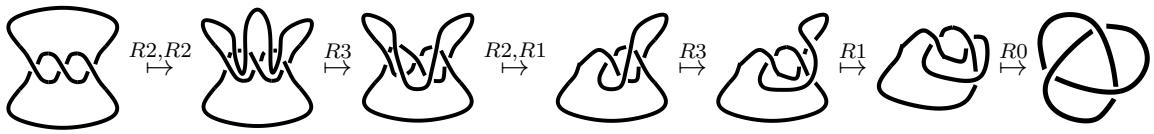
We now know how to represent any knot by a diagram. Unfortunately any knot can be represented by infinitely many different diagrams, which makes it unclear just how much of the information one can read off from a diagram (for example, its number of crossings or regions, its adjacency matrix when thought of as a planar graph, etc.) really has anything to do with the original knot, rather than just being an "artefact" of the diagrammatic representation. Fortunately, we can understand when two different diagrams can represent the same knot.

Theorem 2.3.1 (Reidemeister's theorem)**.** *Two knots K, K' with diagrams D, D' are equivalent if and only if their diagrams are related by a finite sequence $D = D_0, D_1, \ldots, D_n = D'$ of intermediate diagrams, such that each differs from its predecessor by one of the following three* (*really four, but we tend to take the zeroth for granted*) Reidemeister moves. (*The pictures indicate disc regions of the plane, and the portion of knot diagram contained: the "move" is a local replacement by a different portion of diagram, leaving everything else unchanged.*)



Before sketching the proof of this theorem it is best to make some further remarks about it and explain some of its consequences.

(1). The "if" direction is trivial. It's clear that sequences of Reidemeister moves don't change the equivalence class of knot represented by the diagram, because each such move can be realised 3-dimensionally by a small sequence of Δ-moves. But it is useful to state this fact clearly and be aware of it. It shows that, for example, exhibiting a sequence of moves between two diagrams constitutes a *proof* that they represent equivalent (i.e. "the same") knots. (In practice this is very tedious to do, and it's very easy to make a mistake, even if one uses chalk and keeps redrawing the same diagram, rather than drawing a sequence of separate ones.)



(2). The "zeroth" Reidemeister move is just a planar version of the Δ-move. The equivalence relation it (by itself) generates on planar diagrams is called *planar isotopy*. Thus, to say that two diagrams are planar isotopic amounts to saying that one can be wiggled, stretched and generally deformed into the other one, but without any change of combinatorial structure (i.e. how many crossings there are, which arcs connect which crossings, and so on.) We tend to take this move for granted (which is why I drew it smaller than the others!): it's usually understood that when we draw a knot diagram, we don't care what precise "shape" its edges have, and so on.

(3). A more highbrow way to state Reidemeister's theorem is that it proves that there is a bijective correspondence between the sets of equivalence classes

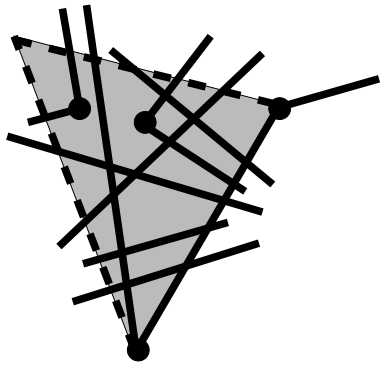
$$\left\{\frac{\text{polygonal knots in } \mathbb{R}^3}{\Delta\text{-move equivalence}}\right\} \leftrightarrow \left\{\frac{\text{polygonal knot diagrams in } \mathbb{R}^2}{\text{Reidemeister move equivalence}}\right\}.$$

At this point we see that in studying topological types of knots, it would be quite reasonable to *define* a knot type to be an equivalence class of diagrams under Reidemeister moves. If we did this we could ignore all the previous stuff about knots in $\mathbb{R}^3$, their projections, and so on. This is in fact what Gilbert and Porter do in their book, but it seems a bit artificial to start with that definition. When I first saw Reidemeister's theorem I found it quite easy to believe, but impossible to imagine how to actually prove it, or where these particular Reidemeister moves really "came from". I hope the explanation below will show that they emerge quite nicely from the proof.

(4). The main way we will use the theorem is to produce invariants of knots. We will construct functions, computable from knot diagrams, which take the same value on all diagrams of a given knot. The way to prove that a function of diagrams is a knot invariant is simply to check that it takes the same value on any diagrams differing by a single Reidemeister move: this is usually easy to do, if the function is in any way a locally-computable thing.

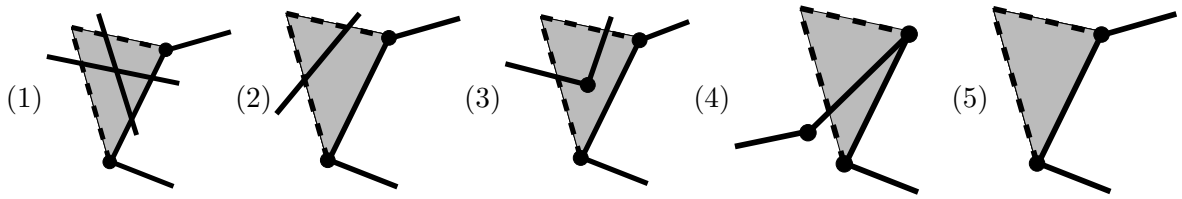
(5). Another application is to the problem of algorithmically classifying knots. This is described in the next subsection.

Proof of Reidemeister's theorem (sketch). As noted above, the "if" part is trivial, so we consider the "only if" part. Suppose that K and K' are knots with diagrams D and D'. If we know that K and K' are equivalent then for some n, there is a sequence of knots $K = K_0, K_1, \ldots, K_n = K'$, each differing from its predecessor by a Δ-move. We can assume that all of $K_1, \ldots, K_n$ have regular projections (by applying suitably tiny ϵ-perturbations) and hence diagrams $D = D_0, D_1, \ldots, D_n = D'$, each of which differs from its predecessor by the *projection of* a Δ-move. The projection of the triangle defining the Δ-move probably overlaps lots many pieces of the knot diagram, as shown below.



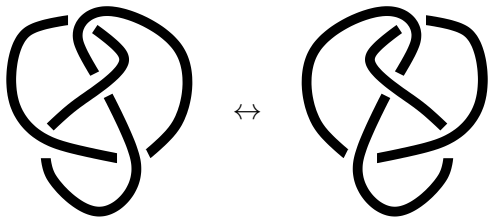
If this is so, it is possible to subdivide it into lots of smaller triangles (replacing the single "large" Δ-move by a sequence of "smaller" ones) so that each small triangle contains exactly one of the following elementary pieces of diagram: (1) a single crossing of the diagram and no vertices (2) a

single arc-segment and no vertex (3) two arc-segments joined at a vertex (4) a single arc segment sharing a vertex with an edge of the triangle (5) nothing, as shown below.



There are actually quite a few variations of each of these, depending on exactly which sides of the triangle the "contents" meet, and what the relative heights of the arcs are (i.e. over/under-ness of the crossings, which is not pictured). But with a little care, one can see that all of these "small" projections of Δ-moves can be realised by Reidemeister moves. (In the pictured versions, (1) is R3, (2) and (3) are R2, (4) is R1 and (5) is R0. But some variations of (1) require an R3 and an R2; some variations of (2) and (3) may be just R0.) See remark 2.3.5 below for a further relevant comment. □

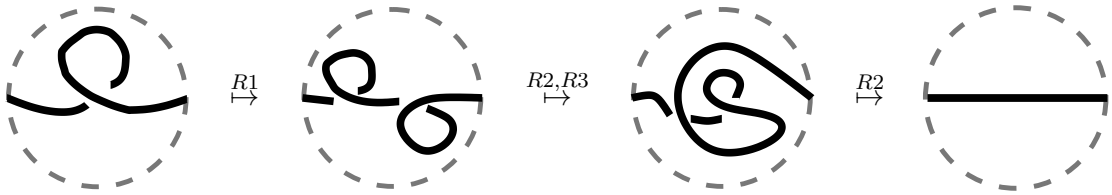
Exercise 2.3.2. Draw a sequence of Reidemeister moves which sends the diagram of the *figure-eight knot* below to its mirror image.



Exercise 2.3.3. Draw a sequence of Reidemeister moves which sends the Whitehead link to itself, but exchanges the two components. (Draw them in different colours to make it clear.)

Remark 2.3.4. The theorem is true without modification if one considers links of more than one component instead of knots.

Remark 2.3.5. The statement of Reidemeister's theorem given above is economical in its list of moves (this will be useful in the next chapter.) Suppose for example that one has a knot diagram containing a kink like the one shown above on the left of move R1 but with the crossing switched. Move R1 does *not* allow one to replace this by an unkinked strand in one go: it is quite simply a different local configuration, about which we have said nothing. However, it is possible (as it must be, given the theorem!) to remove this kind of kink using a combination of the existing moves R1, R2 and R3.

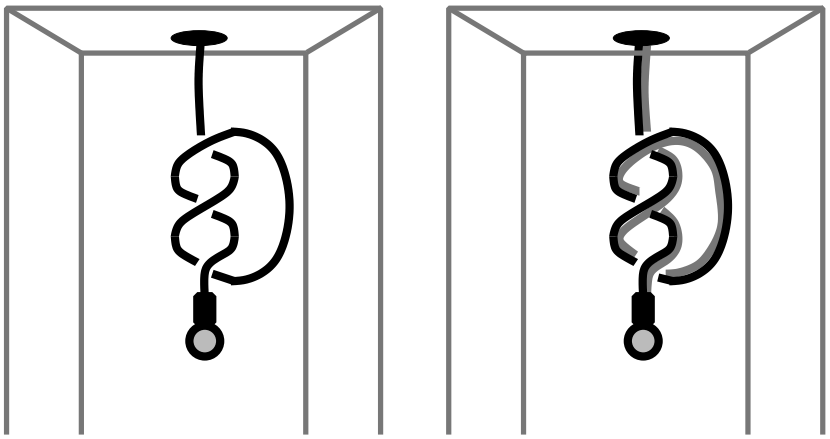


In fact R3 also has variants: the crossing might be switched, or the strand moved behind the crossing instead of in front. If one carries out a rigorous proof of the theorem, one will need all these configurations (two sorts of R1, one R2 and four R3's). But by similar compositions of the three official moves, these extra cases can be discarded.

Remark 2.3.6. If one wants to consider *oriented* knots or links, the Reidemeister moves have to be souped up a bit. We now need moves on oriented diagrams (every arc involved has an arrow

of direction, and these arrows are preserved by the moves in the obvious way), and in proving the theorem we seem to need even more versions of each move: there are two, four and eight possible orientations on each unoriented case of R1, R2, R3 respectively. The compositions just used to economise don't work quite so well, but they do reduce to the three standard unoriented configurations, each with all possible orientations. Thus there are two R1's, four R2's and eight R3's.

Exercise 2.3.7. Suppose a lightbulb cord is all tangled up. Can it be untangled without moving the bulb (or ceiling) during the process? Suppose there are *two* parallel cables (say a blue and a brown) going to the bulb, and blue is on the left-hand side at the fitting and the bulb - can you still do it without moving the bulb?



Exercise 2.3.8. Is it possible to design a computer program which classifies the different equivalence classes of knots and outputs a table like the one (of knots up to 9 crossings) I gave out? Consider the various kinds of tasks which the computer would have to carry out, and think about how feasible each one is. (This is a fairly subtle question! The idea is not to worry about the details of the program, but to realise that any task you ask the computer to perform must be limited to take a finite amount of time and storage space.)

Exercise 2.3.9. (2007M) State precisely Reidemeister's theorem on equivalence of (unoriented) links. (The reason for asking this apparently easy question is that even when you understand clearly what the theorem "says" and "means", actually writing it down in good grammatical English is harder than it sounds!)

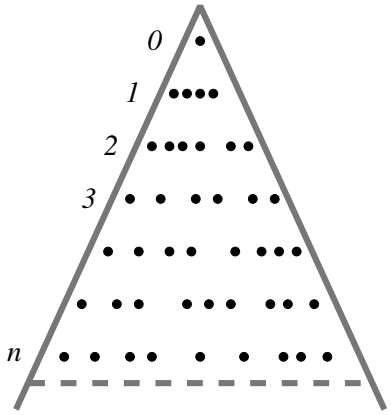
Exercise 2.3.10. Show that the number of regions in a diagram of a knot equals the number of crossings plus 2. (There are several possible arguments. You might consider how the number of regions changes as you physically draw the diagram with a continuous motion of the pen. Or you can think about Reidemeister moves.)

2.4. **Is there an algorithm for classifying and tabulating knots?**

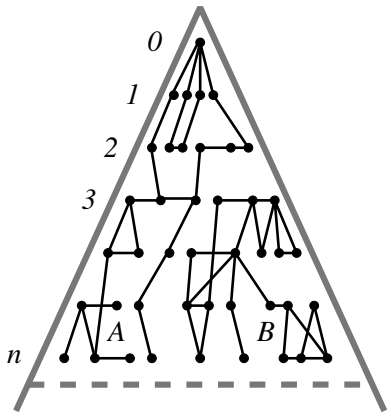
One might wonder whether Reidemeister's theorem makes the problem of generating knot tables – which really amounts to the *classification problem* of determining whether two different diagrams do or do not represent the same knot – into one which is algorithmically solvable by computer. Let's think about how this might be done.

Fix a positive integer n; we want to try to generate a table of all distinct knots which can be represented by a diagram with n or fewer crossings. A computer can certainly enumerate the finitely many *diagrams* with n crossings or fewer. (We can generate all n-crossing diagrams by drawing n crossings in a row, then pairing up their $4n$ free ends with arcs in all possible ways which don't introduce additional crossings.) But we need to decide which of these diagrams represent the same knot, and which represent distinct knots.

Imagine arranging these diagrams in a pyramid-like table of height n, with depth corresponding to crossing number. It will look something like this, where the dots in the nth row are supposed to correspond to the different n-crossing diagrams.



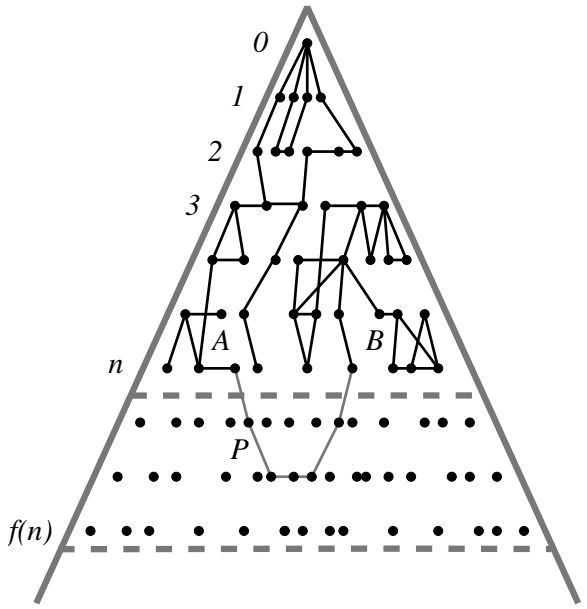
We could now add "edges" to the pyramid diagram, connecting all pairs of diagrams (with n or fewer crossings) which are related by a Reidemeister move. (Edges coming from R1 will join dots in adjacent vertical levels in the pyramid, because R1 alters the number of crossings in the diagram by 1; similarly R2-edges join dots 2 levels apart, and R3-edges connect dots within the same level). Again, generating all these edges is a finite procedure, since there are only finitely many places where an R-move could occur in each diagram.



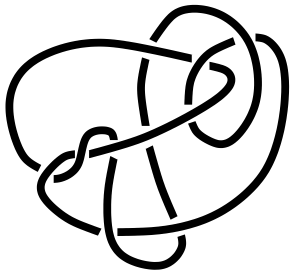
In our pyramid, the diagrams are now partioned into "clumps" consisting of diagrams which are connected to one another by these edges. All the diagrams in a given clump represent the same knot, and it's easy to conclude that the clumps correspond to the distinct knots with up to n

crossings. For example, in the above (not terribly realistic) picture, there are two clumps A and B, and so we might believe there are precisely two distinct knots represented by diagrams with n or fewer crossings.

But this is not true! Diagrams in different clumps need *not* represent different knots! Being in different clumps means only that there is no sequence of Reidemeister moves *which only involves diagrams with n or fewer crossings* joining them; it does not rule out the possibility that they are connected by a chain of R-moves which does involve at some point a diagram with *more* than n crossings. In the picture, this corresponds to the existence of a path P, going below the part of the pyramid we have constructed, joining A and B.



This really can happen. Here is an example: a 7-crossing diagram of the unknot (it's quite easy to see that it's an unknot), which cannot be directly reduced by a R1 or R2, and cannot be R3-ed at all! Consequently the first R-move in any sequence must be the *addition* of a kink or a R2 move *adding* two crossings. Therefore it cannot be transformed into the standard 0-crossing unknot diagram without using an 8- or 9-crossing diagram at some point!



So our finite algorithm (draw the n-level pyramid and its R-move edges, look at the clumps) doesn't quite work. Is all lost? Well, not necessarily! Suppose we knew that *whenever two $\leq n$-crossing diagrams could be related by a chain of R-moves, then they could be related by a chain utilising diagrams with no more than $2n$ crossings.* Then we could salvage the algorithm: to classify diagrams with n or fewer crossings into knot types, we could simply construct the $2n$-level pyramid and its clumps. It would now be true that diagrams *in the top n levels* lie in the same clump, if and only if they are connected by paths *in the whole $2n$-level pyramid.*

The quantity $2n$ here is not important: any function $f(n)$ of n would do. What is essential is that we *know that there is some bound* on how deep a sequence of R-moves P relating two $\leq n$-crossing diagrams might go, if it exists at all. The computer needs to be told "look for sequences of R-moves which don't go lower than $f(n)$" (we have to specify some actual number here) in order for its task to be a finite one. As a slight alternative, if we had a function $R(n)$ which we knew to be an upper bound on the *number* of R-moves needed to relate two $\leq n$-crossing diagrams of the same knot, this would also be enough: a chain P of length l between two n-crossing diagrams certainly can't go deeper than $n + 2l$, so bounding the length also bounds the depth.

Intuitively it seems obvious that such bounds should exist! For example, "clearly" we will never need to move through diagrams with 100 crossings, in order to relate two 10-crossing diagrams of the same knot. But it has been surprisingly hard to prove the existence of such bounds, and in fact this was only done very recently:

Theorem 2.4.1 (Coward & Lackenby, 2010)**.** *If D and D' are two diagrams of the same knot, both having less than n crossings, then there exists a sequence of Reidemeister moves relating them of length less than $R(n)$, where*

$$R(n) = 2^{2^{2^{\cdot^{\cdot^{\cdot^{2^n}}}}}}$$

with the number of 2s in the stack being $10^{1000,000n}$)

What can one say but "!!!!"? This is the most ridiculously enormous number I've ever seen in a paper that wasn't specifically devoted to ridiculously enormous numbers. Clearly (or perhaps "clearly") this is a titanic overestimate: probably the statement is still true with a much more reasonable function like $R(n) = 2^n$. But we don't know how to prove that, and we don't really care - the important point is that Coward and Lackenby managed to prove that *there is a finite value of $R(n)$*.

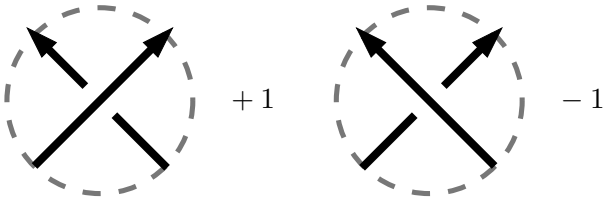
It's interesting to think about implementing this algorithm. If we applied our "generate the pyramid of depth $f(n)$ in order to group the $\leq n$-crossing diagrams into clumps", we would in practice *almost certainly* get a correct tabulation of $\leq n$-crossing knots even if we took something like $f(n) = 2n$. But we would only know *for certain* that it was accurate if we took $f(n) = R(n)$ (and unfortunately this version would always take longer than the age of the universe to run.)

3. Simple invariants

3.1. **Linking number.**

One of the simplest invariants that can actually be computed easily is the linking number of an oriented link. It is computed by using a *diagram* of the link, so we then have to use Reidemeister's theorem to prove that it is *independent* of this choice of diagram, and consequently really does depend only on the original link.

Definition 3.1.1. Let D be a diagram of an oriented link. Then the *total linking number* $Lk(D)$ is obtained by taking *half* the sum, over all crossings, of contributions from each given by

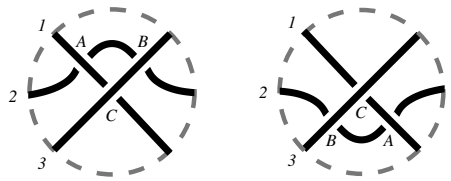


if the two arcs involved in the crossing belong to *different* components of the link, and 0 if they belong to the same one.

Remark 3.1.2. The sign of a crossing (*positive* or *negative* according to the above conventions) is only determined when the strings involved are *oriented*. This enables one to look at the crossing at an angle where both strings point "upwards", and then decide whether the SW-NE or SE-NW string is on top. If there are no arrows, one *cannot* distinguish between crossings in such a way, and this is why the linking number is only defined for oriented links.

Theorem 3.1.3. *If D, D' are two diagrams of an oriented link L then $Lk(D) = Lk(D')$, and hence this number is an invariant $Lk(L)$, the* total linking number of L.

Proof. The two diagrams differ by a sequence of (oriented - see remark 2.3.6) Reidemeister moves, so all we need to do is check that each of these preserves the linking number. Certainly planar isotopy preserves it. In all the other moves, we need only compare the local contributions from the pictures on each side, as all other crossings are common to both diagrams. In R1, one side has an extra crossing but it is a self-crossing, so contributes nothing extra. In R2, one side has two extra crossings: either the two strings involved belong to the same component (in which case both extra crossings are worth 0) or they belong to different components, in which case their contributions are equal and opposite, whatever the orientation on the strings (there are four cases). For R3, each of the three crossings on the left has a counterpart on the right which gives the same contribution, whatever the status of the strings involved or their orientation. Hence the sum of the three is the same on each side.



□

Example 3.1.4. The Hopf link has two possible orientations, one with $Lk = +1$ and one with $Lk = -1$: these are therefore distinct as oriented links. The 2-component unlink has $Lk = 0$. Hence this is distinct from the Hopf link even as unoriented links.

Since for knots, the total linking number is always 0 (all crossings are self-crossings) this invariant is totally useless as a knot invariant.

Exercise 3.1.5. Compute the linking number of the Borromean rings by first choosing an orientation for each component. Does the choice matter?

Exercise 3.1.6. Show that any (oriented) Brunnian link with three or more components has linking number zero. (Remark: the definition of Brunnian link given earlier does not really work for 2-component links. Instead we take a Brunnian 2-component link to be a link whose components are both unknots and whose linking number is zero.)

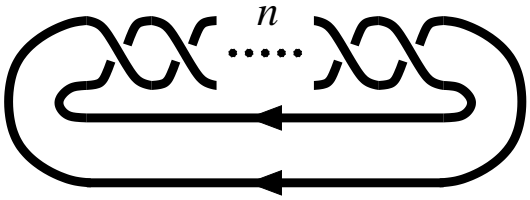
Exercise 3.1.7. Draw a picture to show that there exist oriented 2-component links with linking number n, for any integer n.

Exercise 3.1.8. Prove that if the orientation on one component of a two-component oriented link L is reversed then its linking number is negated. What is the linking number of the mirror-image link $\overline{L}$? Would either of these results still hold if L had three or more components?

Exercise 3.1.9. Suppose L is an oriented link whose components are numbered 1 to n. We can define a *pairwise linking number* L_{ij} by deleting all but the components numbered i and j and then computing their linking number in the usual way. Show that L_{ij} is an invariant of L, and that the total linking number $Lk(L)$ equals the sum $\sum_{i<j} L_{ij}$.

Exercise 3.1.10. Show that any diagram of a link can be changed into a diagram of the unlink by suitable crossing changes. Assume that the link is oriented: what is the effect of a crossing change on the linking number (hint: there are three possibilities)? Use this to prove that despite its initial factor of $\frac{1}{2}$, the linking number of any oriented link is always an integer.

Exercise 3.1.11. Show that by a combination of *self-crossing* changes and isotopy, any 2-component oriented link can be transformed into, and has the same linking number as, one of the links $L_n, n \in \mathbb{Z}$ shown below, where there are n full twists (pairs of crossings) at the top, so that one unknot winds n times about the other (the sign of n denotes the direction of winding). Thus the linking number has a nice visual interpretation as a *winding number* (c.f. complex analysis).



Exercise 3.1.12. (1997F) (i) Explain what is meant by a *knot*, an *equivalence of knots*, a *regular projection* of a knot, and a *knot diagram*. State Reidemeister's theorem characterising diagrams of equivalent oriented links.

(ii). Define the *linking number* of a two-component oriented link, and prove that it is an invariant of the link.

(iii). Prove that by changing some of the crossings, any diagram of a two-component oriented link may be turned into a diagram of a two-component unlink.

(iv). Prove that the unlinking number of a two-component oriented link (i.e. the minimal number of crossing changes needed to turn the link into an unlink) must be greater than or equal to the absolute value of the linking number.

3.2. **Invariants.**

Now that we have Reidemeister's theorem, we can at last construct some *invariants* and use them to prove that certain knots and links are inequivalent.

Definition 3.2.1. A *knot invariant* is any function i of knots which depends only on their equivalence classes.

Remark 3.2.2. We have not yet specified what kind of values an invariant should take. The most common invariants are integer-valued, but they might have values in the rationals $\mathbb{Q}$, a polynomial ring $\mathbb{Z}[x]$, a Laurent polynomial ring (negative powers of x allowed) $\mathbb{Z}[x^{\pm 1}]$, or even be functions which assign to any knot a group (thought of up to isomorphism).

Remark 3.2.3. The function of an invariant is to *distinguish* (i.e. prove inequivalent) knots. The definition says that if $K \cong K'$ then $i(K) = i(K')$. Therefore if $i(K) \neq i(K')$ then K, K' cannot be equivalent; they have been *distinguished by* i. (We may also say that i can *detect* that $K \ncong K'$.)

Remark 3.2.4. Warning: the definition does not work in reverse: if two knots have equal invariants then they are *not* necessarily equivalent. As a trivial example, the function i which takes the value 0 on all knots is a valid invariant but which is totally useless! Better examples will be given below.

Remark 3.2.5. Link invariants, oriented link invariants, and so on (for all the different types of knotty things we might consider) are defined and used similarly.

Example 3.2.6. The *crossing number* $c(K)$ is the minimal number of crossings occurring in *any* diagram of the knot K. This is an invariant by definition, but at this stage the *only* crossing number we can actually compute is that of the unknot, namely zero!

Example 3.2.7. The *number of components* $\mu(L)$ of a link L is an invariant (since wiggling via Δ-moves does not change it, it does depend only on the equivalence class of link).

Exercise 3.2.8. Define the *stick number* of a knot to be the minimal number of arc segments with which it can be built. Show that the only knots with 4 or 5 arcs are unknots, and show thus that the trefoil has stick number 6. Define the *human number* (!) of a knot to be the minimal number of people it takes (holding hands in a chain) to make the knot - what is it for the trefoil and figure-eight?

We have now seen several examples of invariants. One (the number of link components), was obvious and not interesting. Two (the linking number and τ) were computable from diagrams and proved to be invariant under Reidemeister moves. But we didn't know what they meant in any intrinsic sense. The others (the crossing number and the frankly silly stick and human numbers) were invariants by definition but for which there seems no clear way of computing them at all. The best we seem to be able with these is produce upper bounds (by just drawing examples), and with a lot more work, maybe lower bounds.

This is a basic dichotomy exhibited by the knot invariants one commonly encounters. One type is easily computable but must be proved to be invariant. Such invariants tend not to have a clear topological interpretation (we don't really know what topological information τ is measuring, for example). The other type is obviously invariant (anything defined in terms of "the minimal number of ..." tends to be of this form) but very hard to compute. It is often clear what these invariants "mean", but when we want to evaluate them we have to work very hard. The interplay between these two kinds of invariants, attempting to use "computable" invariants to deduce facts about "non-computable" ones, forms a large part of knot theory.

The unknotting number is another example in this second class. Here is the definition again.

Definition 3.2.9. The *unknotting number* $u(K)$ of a knot K is the minimum, over all diagrams D of K, of the minimal number of crossing changes required to turn D into a diagram of the unknot.

It seems *intuitively* clear that any diagram can be changed into a diagram of the unknot simply by switching some of the crossings. The unknotting number is then the minimal number of such changes necessary (over all diagrams of the knot). But we really should give a proof of this fact, because otherwise we don't even know that the unknotting number is always finite!

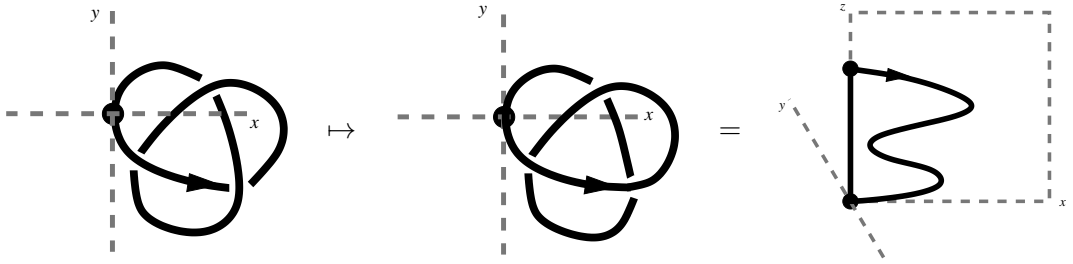
Experience shows that if one draws a knot diagram by hand, only lifting the pen from the page when one is about to hit the line already drawn (and consequently going "under" but never "over" a line already drawn), the result is an unknot. All we do is formalise this idea.

Lemma 3.2.10. *Any knot diagram can be changed to a diagram of the unknot by switching some of its crossings.*

Proof. Take a knot K in $\mathbb{R}^3$ with diagram D. Take a line L parallel to the y-axis and tangent to the knot at its left-hand side, so that the whole diagram is "to the right" of this line in $\mathbb{R}^2$. Parametrise the knot in $\mathbb{R}^3$, starting over p, by a map $t \mapsto (x(t), y(t), z(t))$, for $0 \leq t \leq 1$. (This map will be injective except for the fact that the $t = 0, t = 1$ both map to a point above p).

Now make a new knot K' by gluing the image of $t \mapsto (x(t), y(t), 1 - t)$ to a vertical arc-segment connecting its endpoints $(p, 1)$ and $(p, 0)$. This knot has the same xy-projection as K (technically it's irregular, because of the vertical edge, but this is irrelevant here), but its diagram D' is obtained from D by changing some of the crossings: you move around K' starting from p, the monotonically descending z-coordinate $1 - t$ means that you always go *over* the first time you reach a crossing, and *under* the second time. The resulting diagram is called the *standard descending projection*. (I call it the "lazy man's diagram", because when you draw it, you only take your pencil off the page when you have to!)

But now "look along L", meaning project K' into the xz plane, orthogonal to L: we get a projection with *no crossings*, because the z-coordinate was monotonic and the whole knot lies on one side of L. Therefore K' is an unknot, and D' is a *diagram* of this unknot.



(It doesn't really matter whether the basepoints are on the edge of the diagram here – this just gives the clearest picture that the resulting side view is indeed a diagram of an unknot.) □

Corollary 3.2.11. *For any knot K, $u(K) \leq c(K)/2$.*

Proof. Applying the above procedure to a diagram with the minimal crossing number $c(K)$, we use at most $c(K)$ crossing changes to obtain an unknot K'. If we actually take more than $c(K)/2$, change K instead to the unknot K'' whose z-coordinate is t instead of $1 - t$. This is achieved by changing exactly the crossings we *didn't* change to get K', so takes at most $c(K)/2$. □

Exercise 3.2.12. Prove that unknotting number and crossing number are examples of *subadditive* invariants, satisfying $i(K_1 \# K_2) \leq i(K_1) + i(K_2)$. (It has long been thought that both of these should be equalities, but nobody has ever been able to prove or find a counterexample for either statement!)

3.3. **3-colourings.**

The simplest useful, computable knot invariant is the *number of* 3*-colourings* $\tau(K)$, which we will now study. It is defined in a purely combinatorial way, which cannot be given a reasonable motivation at this stage in the course. It seems to spring from nothing and have no intrinsic geometric definition or meaning. However, in the final chapter we will be able to give a proper explanation of it.

Definition 3.3.1. Pick three colours. If D is an unoriented link diagram, one can consider colouring each of the connected arcs of D with one of the three colours. Suppose there are m arcs. Then there are 3^m such assignments, but we are only interested in the subset $T(D)$, called the *set of* 3*-colourings*, consisting of those satisfying the rule:

(*) at every crossing of D, the three incident arcs are either all the same same colour or are all different.

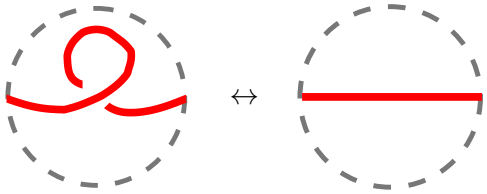
Let $\tau(D)$ be the number of elements of $T(D)$: this is the *number of* 3*-colourings* of the diagram.

Example 3.3.2. The standard diagrams of the unknot and of the trefoil have 3 and 9 3-colourings respectively. The standard diagrams of the two-component unlink and of the Hopf link have 9 and 3 respectively. (Note that the number of 3-colourings works for links as well as knots.)

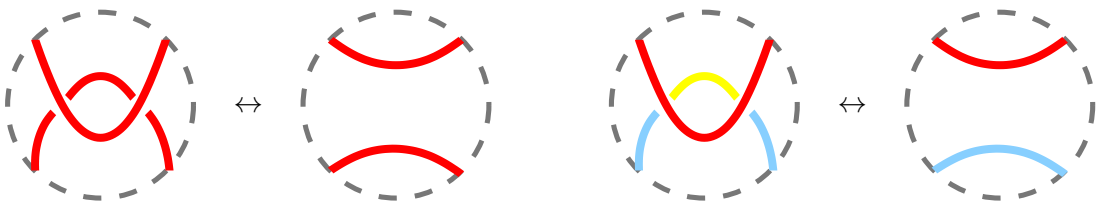
Remark 3.3.3. Obviously any diagram has at least three 3-colourings, because the monochromatic colourings satisfy (*).

Theorem 3.3.4. *The number of* 3*-colourings is a link invariant* $\tau(L)$*.*

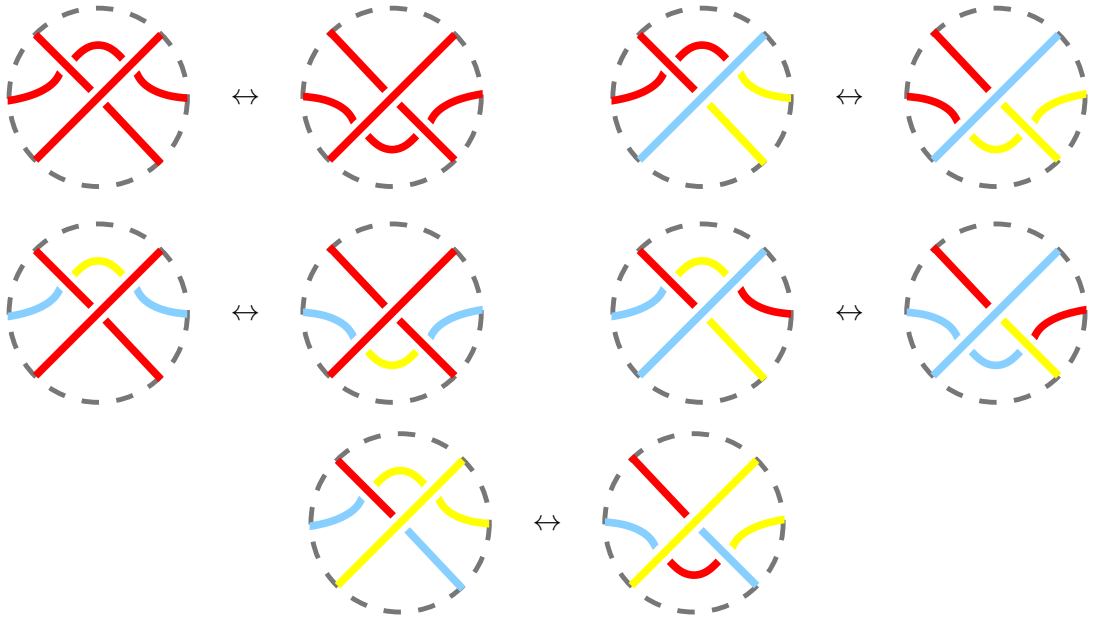
Proof. This theorem is the analogue of theorem 3.1.3 on invariance of linking number. The slightly more compressed statement is intended to mean exactly the same thing: any two diagrams related by Reidemeister moves have the same numbers of 3-colourings, and hence one can consider this number as a function of the *link*, independent of the choice of diagram. To prove it we actually produce explicit bijections between the sets $T(D), T(D')$ whenever D, D' differ by a Reidemeister move (obviously this makes $\tau(D) = \tau(D')$). Once again, planar isotopy clearly doesn't change anything. For R1, any 3-colouring of the left picture must have the same colour c on the two ends, because of the constraint at the crossing. Such a colouring immediately defines a colouring of the right picture: use the same colours everywhere outside this small pictured region, and extend the colour c across the single arc. One can map right to left by exactly the same process and obtain mutually inverse maps $T(D) \leftrightarrow T(D')$, thus a bijection.



For R2, a similar technique is used. Applying the constraints on the left-hand picture, one sees that the top two ends are the same colour a, and the bottom two are the same colour b (if $a = b$ then the middle arc is also this colour; if $a \neq b$ then it is the third colour c). Therefore this defines a colouring of the right-hand picture, and vice versa.



For R3 one has five cases to consider, based on consideration of the colours of the three left-hand ends. They could be all the same; they could all be different; or two could be the same, the third different (three cases according to which end is the odd one out). One has in each case to extend these "input" colours across the picture, and then see that there is a colouring of the right-hand picture with the same colours on the ends to which it corresponds.



□

Exercise 3.3.5. Compute $\tau(5_1)$ from its usual diagram. Observe that this invariant does not distinguish it from the unknot.

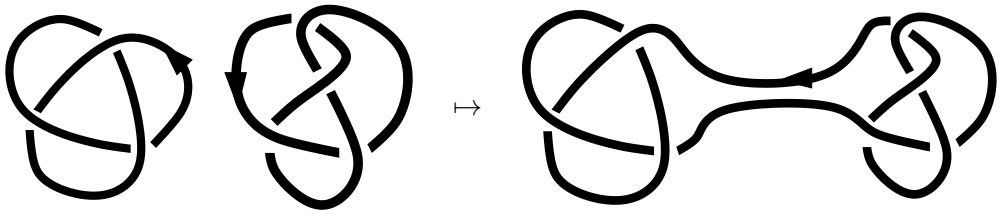
Exercise 3.3.6. Compute the number of 3-colourings τ for the figure-eight knot and the two five-crossing knots.

Exercise 3.3.7. Show that linking number fails to distinguish the Whitehead link from the unlink, but that τ succeeds.

Exercise 3.3.8. Try computing for other knots in the tables. Can you explain why the answer is always divisible by three? Can you explain why it is always a power of three?

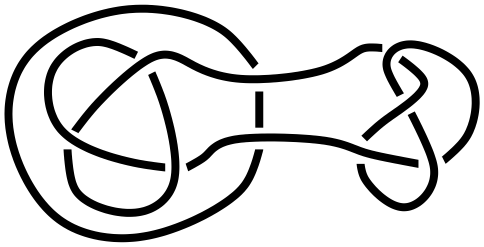
Exercise 3.3.9. Show that the 3-colouring number cannot distinguish a knot K from its mirror-image $\bar{K}$ (which is the knot defined by reflecting K in any plane in $\mathbb{R}^3$).

Exercise 3.3.10. The *connect-sum* $K_1 \# K_2$ of two oriented knots K_1, K_2 may be defined by a diagrammatic example like the one below.

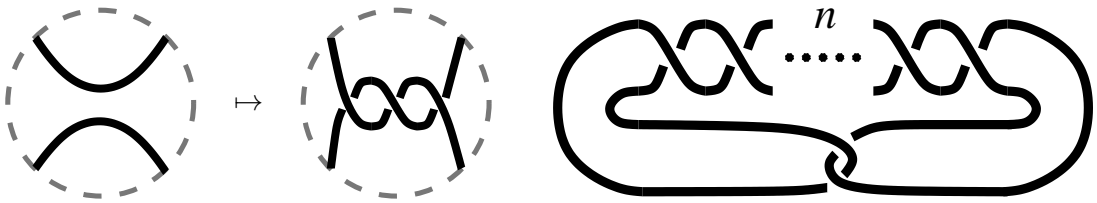


Prove that $\tau(K_1 \# K_2) = \frac{1}{3}\tau(K_1)\tau(K_2)$. (Trick/hint: consider two different ways of computing τ of the diagram D shown below.) Deduce by using repeated connect-sums of trefoils that there are

infinitely many distinct knots.



Exercise 3.3.11. Show that if a link L is changed into a new link L' by the local insertion of three *half-twists* as shown, then $\tau(L) = \tau(L')$. Calculate the number of 3-colourings of the *n-twisted double of the unknot*, shown below (there are n full twists, that is $2n$ crossings, on top).



So far we only have naive methods for computing $\tau(D)$, essentially based on careful enumeration of all cases. With a bit of thought, one can reduce the whole problem of computation to one of linear algebra (which means it is easy, and doable on computers even for very large numbers of crossings).

Start by calling the colours $0, 1, 2$. Let us consider a diagram D with m arcs $A_1, A_2, \ldots, A_m$, and l crossings $C_1, C_2, \ldots, C_l$.

Exercise 3.3.12. Looking at the knot table one sees that $m = l$ for most diagrams: when are they not equal?

Consider the set of all assignments of colours $x_i \in \{0, 1, 2\}$ to the arcs A_i. When does such an assignment constitute an honest 3-colouring? At a crossing where one sees three arcs A_i, A_j, A_k (two ending there and one going over; note that the arcs need not be distinct, for example in a 1- or 2-crossing unknot diagram), the three colours (x_i, x_j, x_k) must form one of the triples $(0, 0, 0), (1, 1, 1), (2, 2, 2)$ or any permutation of $(0, 1, 2)$, if they are to satisfy the condition (*). These nine triples are precisely those $(x_i, x_j, x_k) \in \{0, 1, 2\}^3$ satisfying $x_i + x_j + x_k = 0 \mod 3$ (check: this equation has nine solutions, and we have written them all down). So it makes sense to think of the colours as elements of the *field of three elements* $\mathbb{F}_3$. Then we can write

$$T(D) = \{(x_1, x_2, \ldots, x_k) \in \mathbb{F}_3^m : x_i + x_j + x_k = 0 \text{ at each crossing involving arcs } A_i, A_j, A_k\}.$$

Thus, $T(D)$ is the set of solutions of l homogeneous linear equations in m unknowns over the field $\mathbb{F}_3$.

Theorem 3.3.13. *$T(D)$ is an $\mathbb{F}_3$-vector space. Therefore $\tau(D) = 3^{\dim(T(D))}$ is a power of three.*

Proof. Solutions of homogeneous linear equations form a vector space, and the number of elements in a vector space over $\mathbb{F}_3$ equals 3 to the power of its dimension. □

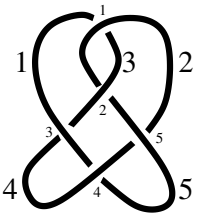
To calculate $\tau(D)$, we therefore associate to a diagram an $l \times m$ matrix A (over $\mathbb{F}_3$) encoding the l equations from crossings, and want to find the dimension of the space of solutions of $Ax = 0$ ($x \in \mathbb{F}_3^m$). This space is just the kernel, its dimension is just the nullity of the matrix, and we can calculate it by Gaussian elimination.

Remark 3.3.14. You've forgotten how do to Gaussian elimination? Oy vey! Briefly:

1. You have a rectangular matrix
2. Find the first non-zero column from the left
3. Find a non-zero entry in that column
4. Using a row swap, move the row containing that non-zero entry to the top
5. Subtract multiples of this new top row from all the ones below it so as to make all their entries in this column zero
6. Go back to stage 1 and repeat, but working now on the smaller rectangular submatrix you get by ignoring the top row, the column we used, and everything to the left of it
7. You finish when this matrix contains only zeroes (or is now a 0×0 matrix).

When you're done, look again at the whole matrix, which will now be in echelon form. The rank of the matrix is the number of non-zero rows; the nullity is the number of columns of the matrix minus its rank. If we began with a *square* matrix therefore, the nullity is therefore just the number of zero rows.

Example 3.3.15. For the knot 5_2 and a suitable numbering of the crossings and arcs, the matrix is



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

Just applying row operations one can reduce this to

$$\begin{pmatrix} 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 2 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

Hence the nullity is 1, and $\tau(5_2) = 3$.

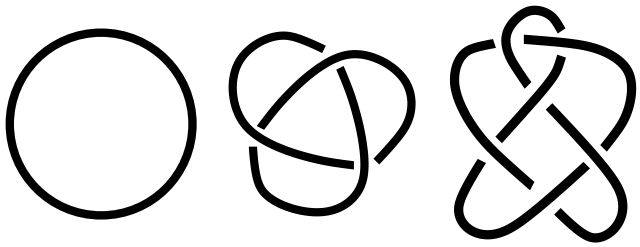
Remark 3.3.16. The most common mistake in computing using this method is to forget that *everything is performed mod 3!* Your matrix should only ever contain the numbers $0, 1, 2$. (I have to admit that I quite like to allow '-1's while I'm doing row subtractions, but I rewrite them all as '2's periodically so as not to forget. '3' is really zero though, so this is very confusing.)

Remark 3.3.17. We know that the monochromatic colourings are always solutions, and hence that the vector $x = (1, 1, \ldots, 1)$ is in the kernel of the matrix. This means that the sum of the entries in each row is $0 \in \mathbb{F}_3$. In fact, we can say more than that: each row of the matrix consists entirely of zeroes apart from three '1's (if the row corresponds to a crossing with three distinct arcs incident), or one '1' and one '2' (if it's a "kink" crossing with two distinct arcs); or, in the exceptional case of there being a disjoint 1-crossing unknot diagram somewhere in D, a complete row of zeroes occurs.

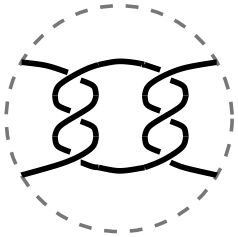
Remark 3.3.18. Livingston talks not about the number of 3-colourings but about "3-colourability of a knot". His definition of a 3-colourable knot is, in our language, one with more than three 3-colourings. Actually counting the number gives more information though, so we will not use his definition.

Exercise 3.3.19. (1999F). (i). Define the *number of* 3*-colourings* $\tau(D)$ of a knot diagram D. Explain briefly but clearly how one shows that it is an invariant of knots.

(ii). Compute the numbers of 3-colourings for the unknot, trefoil and the knot 6_2 shown below.



(iii). Suppose a knot diagram D contains a portion like the one shown below. By colouring this portion so that the four outgoing strings have the same colour, show that the knot represented by D cannot be the unknot or 6_2.

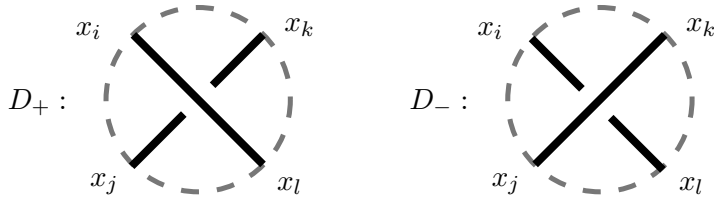


(iv). Show, in contrast, that the trefoil does have a diagram containing such a rectangular portion.

We mentioned earlier that the invariants of knot theory fall into two main classes: the ones we can actually calculate from diagrams (but which we must first prove to be well-defined) such as $\tau(K)$; and the ones such as $u(K)$, which are easy to define (often by minimisation over all diagrams) but not directly calculable from a single diagram. Any single diagram will give an *upper bound* for an invariant defined by minimisation over all diagrams, but finding *lower bounds* for this type of invariant constitutes an important problem in knot theory. Here is an theorem illustrating a beautiful way in which we may use $\tau(K)$ to give lower bounds for $u(K)$.

Theorem 3.3.20. *The unknotting number of a knot is bounded below by* $u(K) \geq \log_3(\tau(K)) - 1$,

Proof. Suppose K_+ and K_- are two knots having diagrams D_+ and D_- which are identical except at one crossing C, as shown below.



Suppose that the outside portion of the diagram (the part common to both) has arcs $A_1, A_2, \ldots, A_m$. We will consider colourings of this portion in the usual way, with colours $x_1, \ldots, x_m$ satisfying the usual equations at each crossing in that portion. This gives us a subspace of solutions $X \subseteq \mathbb{F}_3^m$.

Let A_i, A_j, A_k, A_l be the four arcs which enter the disc region we are looking at. (They might not all be distinct - this won't make a difference.) Let T_+, T_- be the spaces of 3-colourings of K_+, K_-: each is obtained by adding *two* new equations to those defining X:

$$\text{For } T_+: \qquad x_i = x_l \qquad x_i + x_j + x_k = 0$$

$$\text{For } T_-: \qquad x_j = x_k \qquad x_j + x_i + x_l = 0$$

This is a non-standard way of writing things: the first equation on each line tells us that two of the arcs are actually connected across C and so must have the same colour; the second is the usual colouring equation at C. But these equations may be rewritten:

$$\text{For } T_+: \qquad x_i = x_l \qquad x_i + x_l = x_j + x_k$$

$$\text{For } T_-: \qquad x_j = x_k \qquad x_i + x_l = x_j + x_k.$$

(If you substitute the first equation into the second and remember to work mod 3, you can easily recover the earlier version). Therefore we can write

$$T_+ = X \cap W \cap V_+ \qquad T_- = X \cap W \cap V_-$$

where W, V_+, V_- are the spaces of solutions of the equations

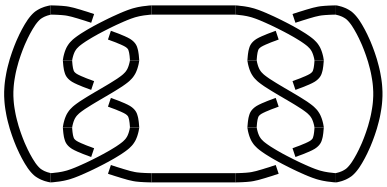
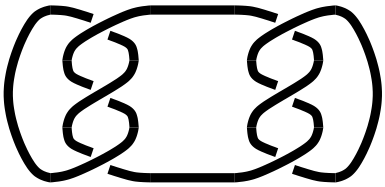
$$x_i + x_l = x_j + x_k \qquad x_i = x_l \qquad x_j = x_k$$

respectively. Now we can use the formula

$$\dim(P + Q) = \dim(P) + \dim(Q) - \dim(P \cap Q)$$

for subspaces P, Q of a vector space to show that either $\tau(K_+), \tau(K_-)$ are equal, or one is three times the other. (This last bit is left as an exercise! Recall that $P + Q$ is the subspace spanned by sums of vectors from P and Q.) □

Exercise 3.3.21. Show that both the *reef* and *granny* knots below have unknotting number 2.

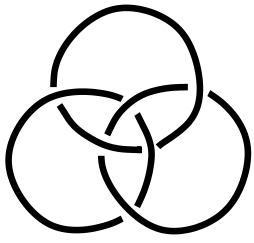
Square (reef) knot  Granny knot 

Exercise 3.3.22. (2003F) Let L be an oriented link, and let $\bar{L}$ be its mirror image.

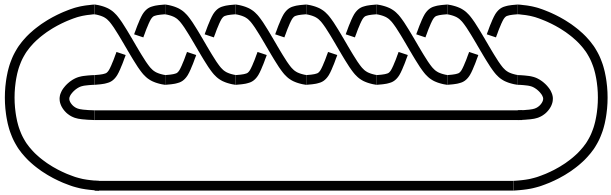
(i). Prove that the number of 3-colourings satisfies $\tau(\bar{L}) = \tau(L)$

(ii.) Prove that the linking number satisfies $\mathrm{Lk}(\bar{L}) = -\mathrm{Lk}(L)$.

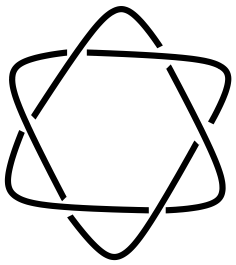
Exercise 3.3.23. (2005M) Calculate the number of 3-colourings of the link shown below. Is it equivalent to the unlink?



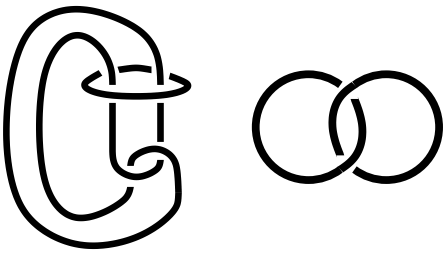
Exercise 3.3.24. (2003M) Calculate the number of 3-colourings of the link shown below. Is it equivalent to the unlink?



Exercise 3.3.25. (2007M) Calculate the number of 3-colourings of the link shown below. Is it equivalent to the unlink?



Exercise 3.3.26. (2001M) Calculate the numbers of 3-colourings of the two two-component links shown below. Are these links equivalent?



3.4. ***p*-colourings.**

There is no need to stick with just three colours. If we use p colours (p a positive integer) then it is still possible to set up conditions on the three colours incident at a crossing such that the resulting number of solutions is an invariant. At this stage, like the definition of a 3-colouring, there is no good motivation for the conditions we choose: the fact that they happen to work has to suffice! Eventually though we will be able to explain what these new invariants measure. In what follows we will actually assume that p is *prime*, so that the colours $\{0, 1, \ldots, p-1\}$ form a *field* $\mathbb{F}_p$. This means that we can work with vector spaces, just as before. (Otherwise our set of colours would be only a ring, not a field, and the set of solutions would merely be a module over this ring, instead of a vector space, which complicates matters.)

Definition 3.4.1. Let p be a prime. Let $T_p(D)$ be the set of colourings of the arcs of a diagram D with elements of $\mathbb{F}_p$, such that at each crossing, where arc A_i is the one going over and arcs A_j, A_k are the ones ending, the following equation is satisfied:

$$2x_i - x_j - x_k = 0 \mod p.$$

Theorem 3.4.2. *$T_p(D)$ is a vector space over $\mathbb{F}_p$, and if two diagrams D, D' differ by a Reidemeister move then there is a bijection between $T_p(D), T_p(D')$. Therefore the number $\tau_p(D)$ of elements of $T_p(D)$ is a power of p, and is an invariant of links $\tau_p(L)$, the* number of p-colourings of L.

Proof. $T_p(D)$ is a set of solutions of homogeneous linear equations over $\mathbb{F}_p$, so it is a vector space and has $p^{\dim T_p(D)}$ elements. The bijections are established just as before: one checks that any colouring of the left-hand diagram can be turned into one of the right-hand one, not changing any of the colours outside the region being altered. □

Remark 3.4.3. τ_2 is not interesting, as it equals 2 to the power of the number of components of a link.

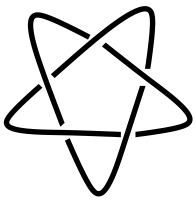
Remark 3.4.4. The invariant τ_3 is the same as our earlier number of 3-colourings τ, since the equation $2x_i - x_j - x_k = 0 \mod p$ is equivalent to the equation $x_i + x_j + x_k = 0 \mod 3$. But be careful: for general p, the equation used to define p-colourings does *not* correspond to the condition "all the colours are the same, or all different"; that only happens when $p = 3$.

Remark 3.4.5. The complete set of invariants $\{\tau_p\}$ is quite strong. It is certainly possible that one invariant τ_p fails to distinguish a pair of knots, while some other one τ_q does distinguish them. The more invariants one uses, the better "separation" of knots occurs. However, there are still pairs of inequivalent knots K, K' which have equal p-colouring invariants for all p. So we haven't succeeded in "classifying knots".

Exercise 3.4.6. Calculate the number of 5-colourings of the knot 6_1.

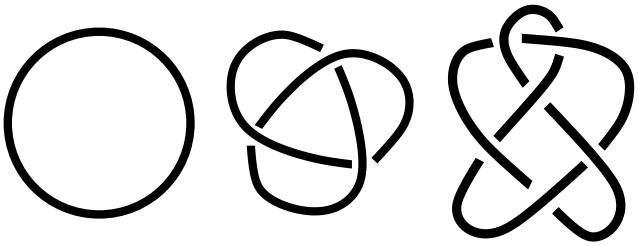
Exercise 3.4.7. Calculate the numbers of 5-colourings and 7-colourings of the knot 7_1.

Exercise 3.4.8. (2003F) Compute the number of 5-colourings of the knot shown below. Is it equivalent to the unknot?

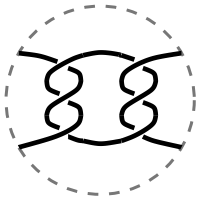


Exercise 3.4.9. (2001F) (i). Define the *number of* 3*-colourings* $\tau(D)$ and the *number of* p*-colourings* $\tau_p(D)$ (for p a prime greater than 3) of a knot diagram D.

(ii). Compute the numbers of 3-colourings for the unknot, trefoil and the knot 6_2 shown below. (You can use any method you want.)



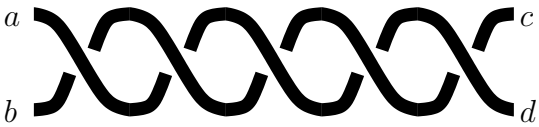
(iii). Suppose a knot diagram D contains a portion like the one shown below. By colouring this portion so that the four outgoing strings have the same colour, show that the knot represented by D cannot be the unknot or 6_2.



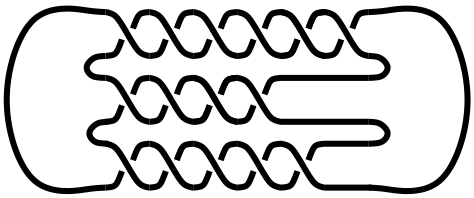
(iv). Show, in contrast, that there is a (non-standard) diagram of the trefoil knot which does contain such a rectangular portion.

Exercise 3.4.10. (2005F) (i). Define the *number of admissible* 5*-colourings* of a link diagram. Explain, in particular, what rule the colours must satisfy at a crossing.

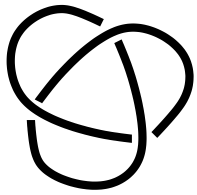
(ii). Suppose the diagram below is given an admissible 5-colouring, with colours a, b, c, d emerging at the ends as shown. Show that $a = c$ and $b = d$.



(iii). Use this result to compute the the number of 5-colourings of the pretzel link $P(6, 4, 5)$ shown below.

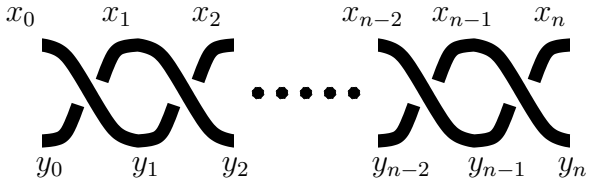


Exercise 3.4.11. (2000F) (i). Let D be a knot diagram. Define the *number of p-colourings* $\tau_p(D)$ of D. If D is the trefoil diagram shown below, write down an explicit set of linear equations (over the field $\mathbb{F}_p$ of p elements) whose solutions corresponds bijectively to p-colourings of D.

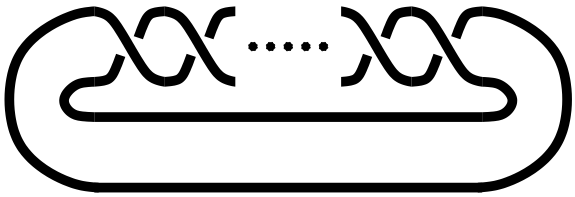


(ii). Compute, for each prime number $p \geq 3$, the number τ_p of p-colourings of the trefoil.

(iii). Suppose a p-coloured knot diagram D contains a portion like the one shown below, with colours x_i, y_i appearing at the places indicated. What are the equations relating x_{i+1}, y_{i+1} to x_i, y_i? Let $z_i = x_i - y_i$. What are the equations relating z_{i+1}, x_{i+1} to z_i, x_i? Use this to express x_n, z_n in terms of x_0, z_0.



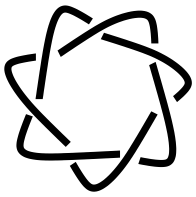
(iv). Let m be a positive integer and $p \geq 5$ be prime. The $(2, 2mp + 3)$ torus knot can be described by a diagram D of the form shown below, with $2mp + 3$ crossings. Show that $\tau_p(D) = p$.



Exercise 3.4.12. (2007F) (a). Let p be a prime number. Give the definition of the *number of p-colourings* $\tau_p(D)$ of a link diagram D.

(b). Show that the number of p-colourings is invariant under Reidemeister moves, and hence that τ_p is an invariant of unoriented links.

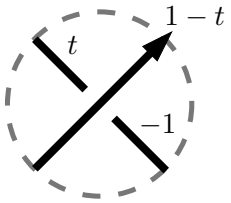
(c). Calculate the invariant τ_5 of the following knot:



3.5. **Optional: colourings and the Alexander polynomial. (Unfinished section!)**

Here is a final variation on the idea of "solving linear equations coming from knot diagrams", which leads to one of the many possible definitions of the *Alexander polynomial* invariant $\Delta_K(t)$ of knots. From this point of view it appears quite unmotivated; a more meaningful definition of the Alexander polynomial can be given using standard tools of algebraic topology (namely, homology theory and covering spaces) which are unfortunately beyond the scope of this course. Nevertheless, there is a lot of merit in the simplicity of the following approach.

Consider *oriented* diagrams, and fix a parameter $t \in \mathbb{C}^*$ – that is, a *non-zero* complex number. Consider colourings of the arcs A_i by *complex numbers* x_i such that at each crossing, the incident colours (numbers) x, y and z on the over, left and right arcs (respectively) satisfy the linear equation



$$(1-t)x + ty - z = 0.$$

(In an oriented diagram, each crossing may always be viewed so that its over-arc points upwards; then it makes sense to distinguish the other two arcs from one another by saying that one is "to the left" and one "to the right". These concepts do of course depend on our standard concept of handedness ("orientation") of the plane itself.)

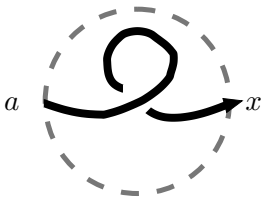
Let $T_t(D)$ be the set of solutions of these equations. Because the equations are linear, $T_t(D)$ is a complex vector space. It will always be at least one-dimensional, because putting all the colours x_i equal to one another (a "monochromatic" colouring) gives a solution. There's no point *counting* solutions because there are infinitely many; what we should do is count the *dimension* of the space. Therefore define $n_t(D) = \dim T_t(D)$; of course n_t is the *nullity* of the square matrix encoding the equations (for this chosen value of t).

Just as we did in the world of p-colourings, we can compute n_t by doing Gaussian elimination (working now over the complex numbers instead of the field with p elements) and counting the number of zero rows we end up with. Note that since the entries of the matrix depend on the parameter t, so does the nullity: if we change the value of t we may get quite different values. (A somewhat similar observation is that the *matrix* used for computing p-colourings (the one whose entries are all 2, -1 or 0) did not depend on p, though its *nullity* did.)

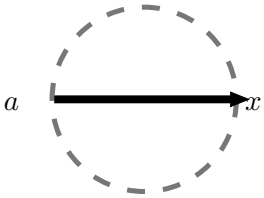
Theorem 3.5.1. *For each fixed non-zero t, the nullity $n_t(D)$ is invariant under Reidemeister moves, and hence n_t is an invariant of oriented links.*

Proof. The proof is very similar to the proof of invariance of p-colourings. We need to set up a bijection between $T_t(D)$ and $T_t(D')$ whenever D and D' are related by a Reidemeister move. All this comes down to is checking that a colouring of the local region under consideration in D corresponds to a unique colouring of the corresponding local region in D' *with the same colours appearing on the boundary*, for this ensures that we can copy all the colours on the rest of the diagram over without change. Because of the orientations there are now several variant versions of each Reidemeister move that need to be checked (four for R1, two for R2, eight for R3) and only one of each will be shown below. If you are careful you can check all the others yourself; much better would be to come up with an argument for why reversing the orientation on one of the strings will not affect whether the process works, since this would resolve almost all the possibilities in one go. (There would still technically be two cases for R1, if you think about it – but we saw in chapter 2 that using R2 and R3, each of them implies the other, so we can ignore that.)

For R1, the equations shown below obviously have the same sets of solutions.

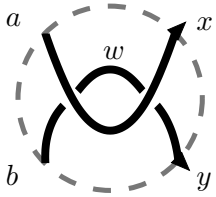


$(1-t)a + ta - x = 0$

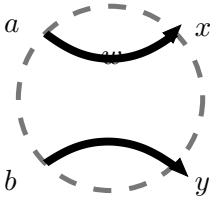


$a = x$

For R2, if colours a and b enter on the left, with x and y on the right and w in the middle, then



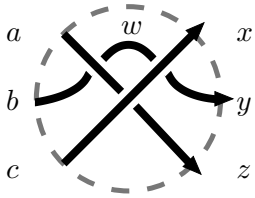
$$\begin{aligned} a &= x \\ (1-t)a + tw - b &= 0 \\ (1-t)a + tw - y &= 0 \end{aligned}$$



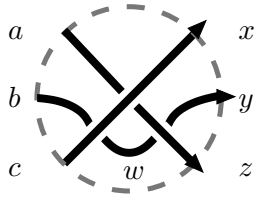
$$\begin{aligned} a &= x \\ b &= y \end{aligned}$$

The left-hand equations are solved by $a = x, b = y$ and w is determined uniquely by $w = t^{-1}(b - (1-t)a)$. Here it is important that t is not allowed to be zero: if it were, the choice of w would be arbitrary, and there would be more colourings of the left diagram than the right!

For R3, if a, b, c enter at the left, then we can solve for the x, y, z which emerge on the right hand-side and the w in the middle short arc:



$$\begin{aligned} c &= x \\ (1-t)a + tw - b &= 0 \\ (1-t)c + ta - z &= 0 \\ (1-t)c + tw - y &= 0 \end{aligned}$$

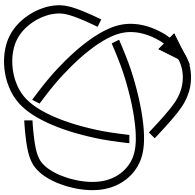


$$\begin{aligned} c &= x \\ (1-t)c + tb - w &= 0 \\ (1-t)c + ta - z &= 0 \\ (1-t)z + ty - w &= 0 \end{aligned}$$

.

Each set of equations has the same solution $x = c, y = (1-t)(c-a) + b, z = (1-t)c + ta$, together with a unique value for w, given in terms of a, b, c, t. The two solutions for w are actually *different*, but that doesn't matter: we just need to know that there is a *unique* choice on each side. □

Example 3.5.2. The (left, and with indicated orientation) trefoil is symmetrical so it's easy to write down the matrix as



$$\mapsto \begin{pmatrix} -1 & 1-t & t \\ t & -1 & 1-t \\ 1-t & t & -1 \end{pmatrix}$$

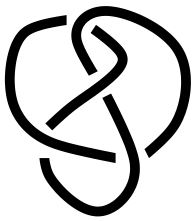
If we add the first two columns to the third, we kill it (this will always happen since the row sums are zero). Then we can use the first and second rows to eliminate the third. We have reduced to a matrix

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

which clearly has rank 2 (and thus nullity 1) unless its two first columns are multiples of one another, in which case it has rank 1 (and nullity 2). This happens if and only if the obvious 2×2 determinant $(-1)(-1) - t(1-t)$ is zero. Thus

$$n_t(\text{left trefoil}) = \begin{cases} 2 & \text{if } t^2 - t + 1 = 0 \text{ (that is } t = \frac{1 \pm \sqrt{3}i}{2}) \\ 1 & \text{otherwise.} \end{cases}$$

Example 3.5.3. The figure-eight gives rise to the following matrix:



$$\mapsto \begin{pmatrix} -1 & 1-t & 0 & t \\ 0 & t & -1 & 1-t \\ 1-t & 0 & -1 & t \\ -1 & t & 1-t & 0 \end{pmatrix}$$

By adding the columns we can zero the last column. By adding the second row to the first, then the second column to the first and third, we get

$$\begin{pmatrix} -1 & 1-t & 0 & 0 \\ 0 & t & -1 & 0 \\ 1-t & 0 & -1 & 0 \\ -1 & t & 1-t & 0 \end{pmatrix} \mapsto \begin{pmatrix} -1 & 1 & -1 & 0 \\ 0 & t & -1 & 0 \\ 1-t & 0 & -1 & 0 \\ -1 & t & 1-t & 0 \end{pmatrix} \mapsto \begin{pmatrix} 0 & 1 & 0 & 0 \\ t & t & t-1 & 0 \\ 1-t & 0 & -1 & 0 \\ t-1 & t & 1 & 0 \end{pmatrix}.$$

Now using the top row to clear the second column, observing that the last row is minus the third, and rearranging we get

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

from which it's clear by the same sort of determinant calculation as above that

$$n_t(\text{figure-eight}) = \begin{cases} 2 & \text{if } t^2 - 3t + 1 = 0 \text{ (that is } t = \frac{3 \pm \sqrt{5}}{2}) \\ 1 & \text{otherwise.} \end{cases}$$

Remark 3.5.4. Notice that the columns always sum to zero, so we can always immediately clear the rightmost one, as happened above. Therefore $n_t(K) \geq 1$, for any t and any knot. (The rows do *not* necessarily sum to zero in general, but there must always be a linear dependence between them, since the fact that the columns are dependent means the determinant of the whole matrix is always zero and hence its rank cannot be maximal. This is the principle that "row rank equals column rank" for any matrix.)

We were thinking above of n_t, *for each fixed value of* t, as an integer-valued knot invariant, and the whole collection $\{n_t\}_{t \in \mathbb{C}^*}$ as a *family* of integer-valued knot invariants. For example, we can't distinguish the trefoil from the figure-eight using n_{23+47i} (as it's 1 for each knot), but we can using each of the four invariants $n_{\frac{3+\sqrt{5}}{2}}, n_{\frac{3-\sqrt{5}}{2}}, n_{\frac{1+\sqrt{3}i}{2}}, n_{\frac{1-\sqrt{3}}{2}}$.

An alternative viewpoint is to think of whole family of invariants $\{n_t\}_{t \in \mathbb{C}^*}$ as a *single* invariant n of knots. The most obvious way to do this would be to define an invariant n whose value is an *integer-valued function of* t by setting $n(K)(t) := n_t(K)$. More interestingly, we could combine all the $\{n_t\}$ into a single invariant R whose value $R(K)$ is a *subset of* $\mathbb{C}^*$ consisting of the values of t for which $n_t > 1$. We should really count these elements *with multiplicity*, just as we do for the set of roots of a polynomial, so as not to forget the actual values of n_t at these special points; it makes sense to define the multiplicity of the special value t to be $n_t - 1$, so that all the boring (non-special) points t with $n_t = 1$ have multiplicity 0. In the examples above we can see that the set $R(K)$ actually *is* the set of roots of a polynomial: $t^2 - t + 1$ for the trefoil, and $t^2 - 3t + 1$ for the figure-eight. Perhaps this is always the case? If so, we should expect there to exist an invariant of knots – let's call it Δ – whose values are *polynomials* $\Delta(K) \in \mathbb{Z}[t]$.

How did these polynomials emerge in the example calculations above? Although we thought we were trying to calculate the nullity of a matrix whose entries were simply complex numbers – because we had in mind a *previously fixed value* for $t \in \mathbb{C}^*$ – it now appears that we were doing some kind of Gaussian elimination for matrices whose entries are *polynomials in* $\mathbb{Z}[t]$ – that is, with t viewed as an indeterminate – at least until the end, where we used a rather ad hoc determinant argument in which t was once again viewed as a fixed complex number.

Is there a general process for computing the nullities $\{n_t\}$ simultaneously so that we can see they correspond to the roots of a polynomial $\Delta(K)$?

To be continued...!

3.6. **Some additional problems.**

These multiple-choice questions are intended to help you check that you have "internalised" the ideas properly. It's very important in learning maths to have a good feeling for what is right, what is wrong, what is plausible or implausible, and to be able to mentally try to validate or invalidate (by looking for simple counterexamples) statements you may be presented with (even your own - it's always good to try to "knock down" things which you think are true, but are uncertain about). You may or may not be able to *prove* or *disprove* the statements, but you should at least have a feeling about whether you should be trying to prove or to disprove them!

Exercise 3.6.1. (2001M) Say whether each of the following statements is true or false. (You don't have to give any explanation, or write anything other than "true" or "false".)

(1). If K_1, K_2 are knots with $\tau_3(K_1) = \tau_3(K_2)$ then K_1 is equivalent to K_2.

(2). If K_1, K_2 are equivalent knots then $\tau_3(K_1) = \tau_3(K_2)$.

(3). If K_1, K_2 are knots with $\tau_3(K_1) = \tau_3(K_2)$, then $\tau_5(K_1) \neq \tau_5(K_2)$.

(4). If an oriented link L has $Lk(L) = 0$, then L is equivalent to an unlink.

(5). A knot K for which $\tau_7(K) = 7$ must be an unknot.

(6). There is no knot K with $\tau_3(K) = 6$.

(7). For any oriented knot K, $\tau_5(rK) = \tau_5(K)$. (Here, rK denotes the orientation-reversed version of K.)

(8). For any knot K, $\tau_5(\bar{K}) = -\tau_5(K)$. (Here, $\bar{K}$ denotes the mirror-image of K.)

(9). For any oriented link L, $Lk(rL) = -\,Lk(L)$. (Here, rL denotes the link obtained by reversing the orientations of all the components of L.)

(10). The function f defined on knot diagrams by setting $f(D) =$ (number of crossings of D) defines an invariant of knots.

(11). For any n, the number of knots with crossing number at most n is finite.

(12). For any n, the number of knots with unknotting number at most n is finite.

(13). If L is an oriented link with n components, then $|\,Lk(L)| \leq n$.

(14). If L is an oriented link which has a diagram with c crossings, then $|\,Lk(L)| \leq \frac{1}{2}c$.

(15). If L is a link with n components then $\tau_2(L) = 2^n$.

Exercise 3.6.2. (2007M) Say whether each of the following statements is true or false. (You don't have to give any explanation, or write anything other than "T" or "F".)

(1). There exists a knot K with $\tau_3(K) = 18$.

(2). If K_1, K_2 are knots with $\tau_3(K_1) = \tau_3(K_2)$ then K_1 is equivalent to K_2.

(3). If K_1, K_2 are knots with $\tau_5(K_1) \neq \tau_5(K_2)$, then $\tau_3(K_1) \neq \tau_3(K_2)$.

(4). If K is a knot with $\tau_3(K) = 27$ then its unknotting number cannot be 1.

(5). For any knot K, $\tau_2(K) = 2$.

(6). For any non-trivial knot, $\tau_7(K) > \tau_5(K)$..

(7). For any knot K, the unknotting number $u(K)$ is less than or equal to the crossing number $c(K)$.

(8). If L is an oriented link with n components, then $|\,Lk(L)| \leq n$.

(9). Any oriented Brunnian link with more than two components has linking number zero.

(10). For any oriented link L, $Lk(rL) = Lk(L)$. (Here, rL denotes the link obtained by reversing the orientations of all the components of L.)

(11). For any oriented link L, $Lk(\bar{L}) = Lk(L)$. (Here, $\bar{L}$ denotes the mirror image link obtained by switching all the crossings in some diagram of L.)

(12). If an oriented link L has $Lk(L) = 0$, then L is equivalent to an unlink.

(13). For each n, the number of knot diagrams with n crossings is finite.

(14). The number of knots K with $\tau_3(K) = 9$ is finite.

(15). There are only finitely many knots with unknotting number 1.

Exercise 3.6.3. (2005M) Say whether each of the following statements is true or false. (You don't have to give any explanation, or write anything other than "T" or "F".)

(1). For each n, the number of knot diagrams with n crossings is finite.

(2). For any knot K, $\tau_5(\bar{K}) = \tau_5(K)$. (Here, $\bar{K}$ denotes the mirror-image of K.)

(3). For any oriented knot K, $\tau_5(rK) = \tau_5(K)$. (Here, rK denotes the orientation-reversed version of K.)

(4). There exists a knot K with $\tau_3(K) = 18$.

(5). If L is an oriented link with n components, then $|\,Lk(L)| \leq n$.

(6). If K_1, K_2 are knots with $\tau_3(K_1) = \tau_3(K_2)$ then K_1 is equivalent to K_2.

(7). If K_1, K_2 are knots with $\tau_5(K_1) \neq \tau_5(K_2)$, then $\tau_3(K_1) \neq \tau_3(K_2)$.

(8). If L is an oriented link which has a diagram with c crossings, then $|\,Lk(L)| \leq \frac{1}{2}c$.

(9). If K_1, K_2 are equivalent knots then $\tau_3(K_1) = \tau_3(K_2)$.

(10). For any knot K, $\tau_2(K) = 2$.

(11). For any oriented link L, $Lk(\bar{L}) = Lk(L)$.

(12). If an oriented link L has $Lk(L) = 0$, then L is equivalent to an unlink.

(13). The number of knots K with $\tau_3(K) = 9$ is finite.

(14). The unknotting number of the trefoil is 3.

(15). There are only finitely many knots with unknotting number 1.

Exercise 3.6.4. (2003M) Say whether each of the following statements is true or false. (You don't have to give any explanation, or write anything other than "true" or "false".)

(1). If K_1, K_2 are knots with $\tau_5(K_1) = \tau_5(K_2)$ then K_1 is equivalent to K_2.

(2). If K_1, K_2 are equivalent knots then $\tau_7(K_1) = \tau_7(K_2)$.

(3). If K_1, K_2 are knots with $\tau_5(K_1) \neq \tau_5(K_2)$, then $\tau_3(K_1) \neq \tau_3(K_2)$.

(4). If an oriented link L has $Lk(L) = 0$, then L is equivalent to an unlink.

(5). For any non-trivial knot, $\tau_7(K) > \tau_5(K)$..

(6). There is no knot K with $\tau_3(K) = 18$.

(7). For any oriented knot K, $\tau_5(rK) = \tau_5(K)$. (Here, rK denotes the orientation-reversed version of K.)

(8). For any knot K, $\tau_5(\bar{K}) = \tau_5(K)$. (Here, $\bar{K}$ denotes the mirror-image of K.)

(9). For any oriented link L, $Lk(\bar{L}) = Lk(L)$.

(10). For any knot K, $\tau_2(L) = 2$.

(11). The function f defined on knot diagrams by setting $f(D) =$ (number of arcs of D) defines an invariant of knots.

(12). For any n, the number of knot diagrams with n crossings is finite.

(13). The number of knots K with $\tau_3(K) = 9$ is finite.

(14). If L is an oriented link with n components, then $|\, Lk(L)| \leq n$.

(15). If L is an oriented link which has a diagram with c crossings, then $|\, Lk(L)| \leq \frac{1}{2}c$.

4. The Jones polynomial

The Jones polynomial is another invariant of links which can be defined combinatorially, using diagrams. It was invented in 1984 by Vaughan Jones (hence* the symbol V), who was working in a completely different area of mathematics – operator algebras – but gradually came to realise that some of the things he had discovered could be used to define a link invariant. This essentially accidental and miraculous discovery came as a complete surprise to knot theorists, and attracted lots of attention because of its novel formulation, which did not appear to be within the standard framework of 3-dimensional topology at all. Over the next five years, lots of work was done leading to vast generalisations of Jones' construction, and many connections to the field of "quantum algebra", a subject more or less initiated by Vladimir Drinfeld in 1986. Because of this, the mathematics dealing with "things not unrelated to the Jones polynomial" is now known as *quantum topology*.

Despite these efforts, the "raison d'etre" of the Jones polynomial – a really convincing and direct way to define it and illuminate its "topological meaning" – was still lacking. Finally in 1989 the physicist Edward Witten gave an equally miraculous explanation using the methods of quantum field theory (applied not to the real 4-dimensional physical world, but to a simplified 3-dimensional of the kind generally referred to by physicists as a "toy model"). Because Witten's explanation relies on the notoriously non-rigorous *Feynman path integral* method from QFT, it cannot be viewed (at present) as a rigorous mathematical *theorem*; but it is unquestionably the most profound interpretation we have (at present) of what the Jones polynomial *really is*. Jones, Drinfeld and Witten were all awarded Fields Medals in 1990, in large part for these works.

(Witten's introduction of QFT as a method of making amazing conjectures in geometry and topology is probably the most important overall influence on these topics in the last 20 years. People often ask whether knot theory is related to the physics of string theory. The answer is "not in the sense you might imagine". The basic objects of string theory are indeed 1-dimensional loops, rather than the traditional 0-dimensional particles, but they are moving in a space of dimension much bigger than 3 (in fact typically 10 or 26), so there is no interesting knotting possible. The sense in which the subjects are both string theory and the theory of the Jones polynomial are examples of quantum field theories.)

From a knot-theoretic point of view, the Jones polynomial is a wonderful thing. It is extremely good at distinguishing knots – it seems to be much more powerful than the previously-known computable knot invariants. It can distinguish knots from their mirror images, which relatively few of the previously-known invariants could do. It can be used to prove the 100-year-old "Tait conjectures" about alternating knots. And it is so easy to work with that it can be fitted into two weeks of an undergraduate course on knot theory!

The approach we will take is not Jones' original one, which is quite different and a bit harder. We will construct it using the *Kauffman bracket polynomial*, a method invented a year or so after Jones' discovery by Lou Kauffman.

* The way new mathematical inventions are named can be quite amusing. Suppose you invent a new knot invariant: it would be considered crassly immodest to say something like "I call this the Roberts polynomial", though it's perfectly acceptable to give it the symbol R (or even better, ρ!) But by far the best technique is to give the polynomial an unusably convoluted ("binary recursive regular isotopy polynomial") or insufficiently specific name ("the diagram polynomial") or just not to name it at all (there are lots of papers called things like "A new polynomial invariant of links"). All of these tricks will (fingers crossed) force everyone else to refer to the thing using your name, but preserve your own modesty!

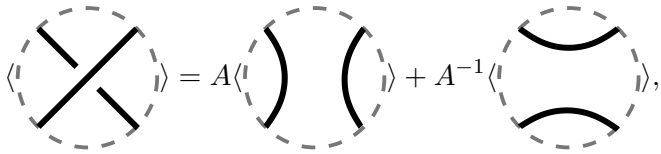
4.1. **The Kauffman bracket.**

The invariants in this section will mostly be *polynomial-valued* rather than *integer-valued* like the ones we've seen so far. In fact they will mostly be *Laurent polynomials* rather than common-or-garden polynomials: they are allowed to have *negative powers* of the variable occurring. Just as $\mathbb{Z}[x]$ denotes the set of ordinary polynomials in x with integer coefficients, $\mathbb{Z}[x^{\pm 1}]$ is the symbol used to denote Laurent polynomials in x with integer coefficients.

Definition 4.1.1. The *Kauffman bracket polynomial* of an *unoriented link diagram* D is a Laurent polynomial $\langle D\rangle \in \mathbb{Z}[A^{\pm 1}]$, defined by the following recursive rules:

(0). It is invariant under planar isotopy of diagrams. (I have numbered this rule "zero" because it's the sort of rule that almost goes without saying, and in fact will barely be mentioned from now on.)

(1). It satisfies the *skein relation*



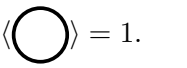
which is a linear relation amongst the bracket polynomials of diagrams differing only locally inside a small disc as shown. That is, the three pictures really represent large diagrams which are identical outside the small dotted circle.

(2). It satisfies $\langle D \amalg U\rangle = (-A^2 - A^{-2})\langle D\rangle$, where U is any closed crossingless loop in the diagram. In other words, disjoint unknot diagrams may be removed at the cost of multiplication by $(-A^2 - A^{-2})$.

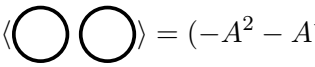
(3). It satisfies the normalisation $\langle U\rangle = 1$; the bracket of a crossingless unknot diagram is the constant polynomial 1.

These axioms in fact suffice to calculate the bracket of any diagram. One can use the skein relation (1) to express the bracket of an n-crossing diagram in terms of those of a pair of $(n-1)$-crossing diagrams, and repeat until one has only crossingless diagrams. These are evaluated using rules (2) and (3).

Example 4.1.2. Let's compute the bracket for the simplest possible diagrams. Firstly we have the 0-crossing unknot diagram, whose evaluation is by definition (using rule (3)):

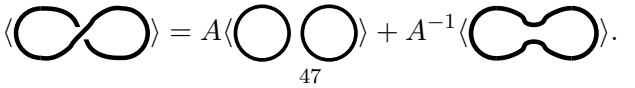


Next, there is the 0-crossing diagram of an unlink with two components. By rule (2), this is equal to $(-A^2 - A^{-2})$ times the bracket of the unknot diagram. Hence



Similarly, a three-component unlink diagram could be evaluated by using rule (2) twice:

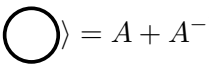
and so on. Now let's consider a 1-crossing unknot diagram, and apply rule (1) to calculate it:



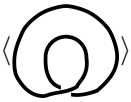
We calculated the first diagram above, and the second one – the "peanut" – is planar isotopic to a plain circular unknot, so is worth 1. Therefore we get $A(-A^2 - A^{-2}) + A^{-1}$, so

$\langle$  $\rangle = -A^3.$

Similarly for the other 1-crossing diagrams we get

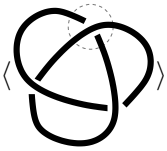
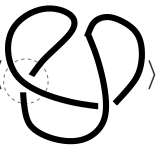
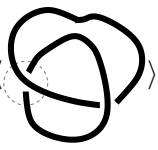
$\langle$  $\rangle = A\langle$  $\rangle + A^{-1}\langle$  $\rangle = A + A^{-1}(-A^2 - A^{-2}) = -A^{-3},$

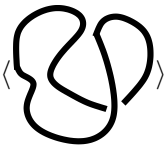
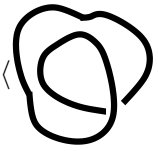
and to complete the list (omitting the analogous calculations)

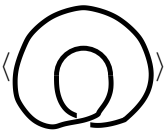
$\langle$  $\rangle = -A^{-3} \qquad \langle$  $\rangle = -A^3.$

Remark 4.1.3. Note that one immediately sees that the Kauffman bracket is *not* an invariant of links: above there are three different diagrams of the unknot with different brackets!

Example 4.1.4. Consider the standard left-handed trefoil diagram, and apply the skein relation at one of its three crossings. This gives two new diagrams, which we hit once more with the skein relation at the indicated crossings, then simplify by removing circles and using planar isotopy:

$\langle$  $\rangle = A\langle$  $\rangle + A^{-1}\langle$  $\rangle.$

$= A\left\{A\langle$  $\rangle + A^{-1}\langle$  $\rangle\right\} + A^{-1}\left\{A\langle$  $\rangle + A^{-1}\langle$  $\rangle\right\}.$

$= A^2(-A^2 - A^{-2})\langle$  $\rangle + 1.\langle$  $\rangle + 1.\langle$  $\rangle + A^{-2}\langle$  $\rangle.$

$= (-A^4 + 1)(-A^3) + A^{-2}(-A^{-3}) = A^7 - A^3 - A^{-5}.$

Remark 4.1.5. (1). When applying the skein relation at a crossing one must be careful to look at the crossing "the right way up", so that the overpass goes from bottom left to top right. Then the term getting the A is the "vertical" reconnection, and the term getting A^{-1} is the "horizontal" one. (Alternatively, one can think of "turning left from the overpass" down onto the underpass to get the A term, and turning right to get the A^{-1} one.)

(2). Notice that "the right way up" is not really well-defined: rotating a crossing by 180° makes it look the same. Fortunately, the diagrams on the right-hand-side of the skein relation are also the same when rotated by 180°, so this ambiguity does not matter.

(3). We don't have orientations on these diagrams, so there is only one type of crossing. You should never use the words *positive* and *negative* to describe crossings unless they have orientation arrows!

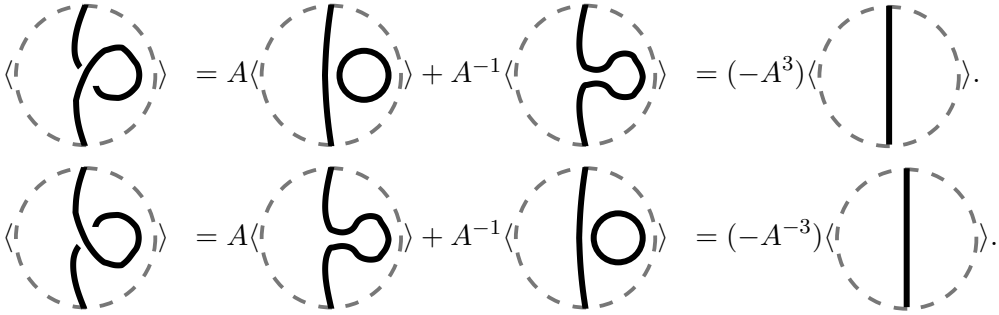
(4). The fact that a 0-crossing unlink diagram with n components evaluates to $(-A^2 - A^{-2})^{n-1}$ is annoying, and it can be corrected easily by changing the normalisation rule (3). Suppose we allow ourselves to remove the *last* unknot circle while multiplying by $(-A^2 - A^{-2})$, leaving the "empty diagram" whose bracket we define to be $\langle\emptyset\rangle = 1$. If we do this then things work more nicely in some ways, though we now have a strange value for the bracket of the unknot! In most

recent work, this other normalisation is more standard, but I am sticking to the original definition of Kauffman in these notes.

(5). You might feel slightly uneasy (and if you don't, you should!) about whether the process of evaluating the bracket of a diagram by recursively applying the rules above is actually well-defined. Is it really obvious (or even true) that applying the rules in some different order (say by choosing different sequences of crossings to break) will always give the same result? In fact it *is* true, and in section 4.3 we'll see a way of thinking which shows this very clearly, but for now I leave you to ponder the question, and we'll proceed assuming it is all right. (One reason to think about this is that later we will see a slightly different skein relation where this is a real problem!)

(6). The word *skein* ("a loosely wound ball of wool") was introduced into knot theory by John Conway in 1970. As you might have noticed, most words having to do with entanglement have already been used in knot theory, so I assume he took out his thesaurus...

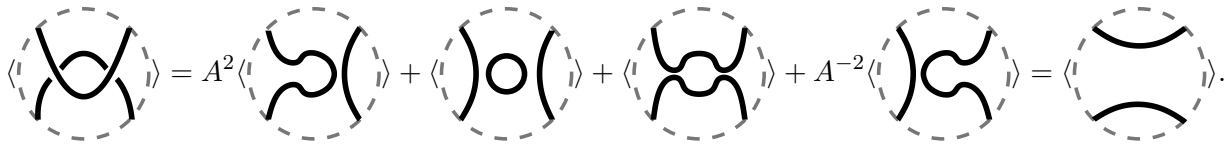
A very important idea in "skein theory" is that because the relations are *local*, we can use them to make partial evaluations of small pieces of link diagrams. The following identities are between brackets of diagrams differing only in the portions shown, like the skein relation (1) itself. (The calculations amount to the same ones we did earlier for 1-crossing diagrams.)



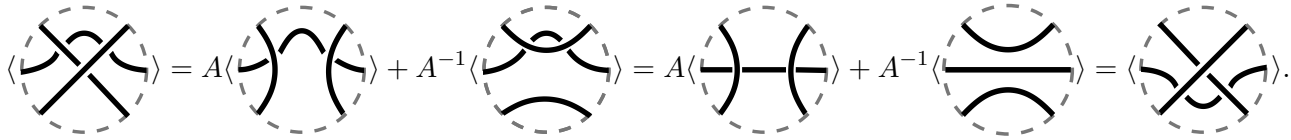
These two formulae describe precisely the *non-invariance* of the bracket under the first Reidemeister move R1. It might seem that failing to be invariant under R1 makes further consideration of the bracket worthless, but this is not so!

Lemma 4.1.6. *The Kauffman bracket is invariant under R2 and R3.*

Proof. For R2, applying the skein relation twice, then removing the little circle, "a miracle occurs": the $(-A^2 - A^{-2})$ from the little circle cancels two of the other terms, leaving just the one we want!

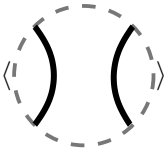


For R3, we apply the skein relation just once, use the invariance under R2 just established, and then the vertical symmetry of the picture:



□

Exercise 4.1.7. Suppose we defined a Kauffman-bracket-like polynomial (of planar isotopy classes of diagrams) in three variables A, B, d by the following modification of the skein relations:

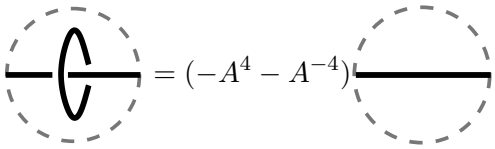
(1). $\langle$  $\rangle = A\langle$  $\rangle + B\langle$  $\rangle$

(2). $\langle D \amalg U \rangle = d\langle D \rangle$

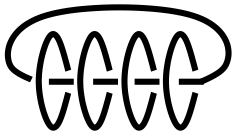
(3). $\langle U \rangle = 1$.

Examine how this "bracket" changes when we perform a Reidemeister move II or III on a diagram. Deduce that we *must* set $B = A^{-1}$ and $d = -A^2 - A^{-2}$ in order to get invariance under these moves: this justifies these previously unexplained (and rather bizarre) choices!

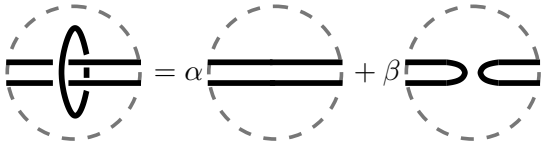
Exercise 4.1.8. Prove that the following relation, between Kauffman brackets of diagrams which differ locally as shown, holds:



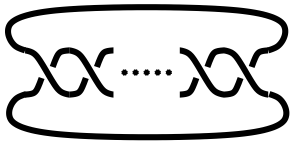
and use this to compute the bracket of the diagram below.



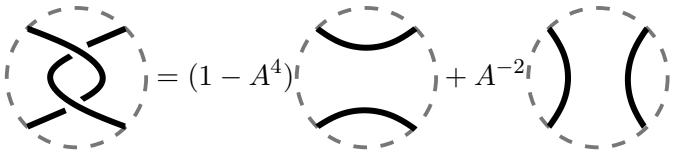
Exercise 4.1.9. Find the constants $\alpha, \beta \in \mathbb{Z}[A^{\pm 1}]$ such that



Exercise 4.1.10. (2007F) Consider the family of link diagrams D_n (for $n \geq 0$, n being the number of crossings in the diagram), as shown below. By applying the skein relation and Reidemeister move 1 appropriately, express the Kauffman bracket of D_n in terms of that of D_{n-1}. Use this "reduction" formula to evaluate completely the Kauffman bracket $\langle D_n \rangle \in \mathbb{Z}[A^{\pm 1}]$.

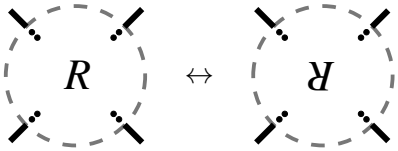


Exercise 4.1.11. Prove that the following relation, between Kauffman brackets of diagrams which differ locally as shown, holds.



Use this to calculate the bracket of the "bracelet" link diagram D_n with n components:

Exercise 4.1.12. Suppose a diagram D contains a subdiagram R which has exactly four strings emerging from it, as shown below. (For example, R might be a single crossing, or one of the diagrams used in Reidemeister II, or much more complicated.) Let D' be the new diagram obtained by rotating the part R by 180°. This operation is called *mutation* of the diagram. Show that the Kauffman brackets of D and its mutant D' are equal.



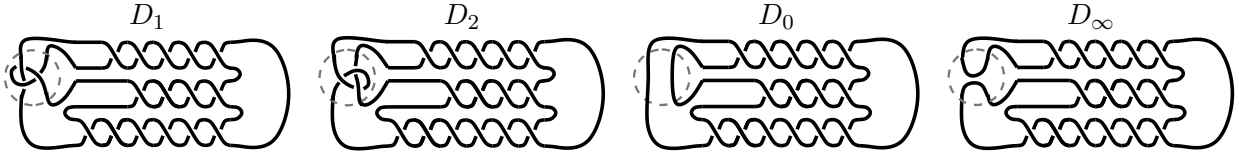
Exercise 4.1.13. (1999F) (i). Define the *Kauffman bracket* polynomial $\langle D \rangle$ of an unoriented knot diagram D. State Reidemeister's theorem about diagrams of equivalent knots.

(ii). Prove that the bracket is invariant under the second and third Reidemeister moves.

(iii). Consider the four knot diagrams D_1, D_2, D_0, D_∞ which are identical except inside a small disc region as shown below. Show that there exist polynomials $f(A), g(A) \in \mathbb{Z}[A^{\pm 1}]$ such that

$$\langle D_1 \rangle = f(A)\langle D_0 \rangle + g(A)\langle D_\infty \rangle.$$

(iv). By expressing $\langle D_2 \rangle$ similarly, show that $\langle D_1 \rangle = \langle D_2 \rangle$.



Exercise 4.1.14. (1998F) (i). Explain what is meant by a *knot*, an *equivalence of knots*, a *regular projection* of a knot, and a *knot diagram*.

(ii). Define the *Kauffman bracket* $\langle D \rangle \in \mathbb{Z}[A^{\pm 1}]$ of a knot diagram D. State Reidemeister's theorem characterising diagrams of equivalent unoriented knots.

(iii). Prove that the Kauffman bracket is invariant under Reidemeister moves R2 and R3.

(iv). Let $p(D) = \langle D \rangle(-1)$ be the evaluation of the Kauffman bracket polynomial of D at $A = -1$. By considering how the operations of changing a crossing of D and performing a Reidemeister move R1 on D affect $p(D)$, show that $p(D) = (-2)^{\mu - 1}$ for any diagram D of a link with μ components.

4.2. **Correcting via the writhe.**

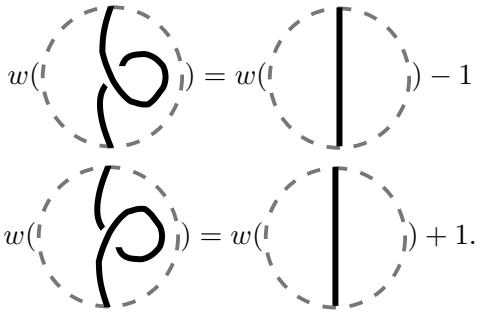
Now the Kauffman bracket is very close to being a link invariant, as it fails only R1, and even then just multiplies in a simple way by $-A^{\pm 3}$, depending on the "handedness" of the kink. What do we really mean by "handedness" here?

If orientations are chosen everywhere then each crossing has a sign, in particular the sign at a kink will measure this handedness, and we can introduce a correction to the bracket to make it a genuine invariant.

Definition 4.2.1. If D is an *oriented* link diagram, then the *writhe* $w(D)$ is just the sum of the signs of *all crossings* of D. (It differs from the total linking number in the fact that the self-crossings *do* contribute here, and there is no overall factor of $\frac{1}{2}$.)

Lemma 4.2.2. *The writhe of an oriented link diagram is invariant under R2, R3 but changes by* ± 1 *under R1.*

Proof. This is just another variation of the proof of theorem 3.1.3 on invariance of the linking number. For R2 and R3, it is even easier, as there is no reason to consider whether the strands involved belong to the same component or not. For R1, there is an obvious change. The slightly surprising thing is that the following identities hold *regardless of the orientation on the string* (easy check):



□

Remark 4.2.3. The orientation is necessary in order to define the writhe, as otherwise one cannot distinguish a "positive" or "negative" crossing. However, the notion of a "positive" or "negative" *kink* is defined independently, as one sees from the above.

Theorem 4.2.4. *If D is an oriented link diagram, then the polynomial $f_D(A) = (-A^3)^{-w(D)}\langle D\rangle$ is invariant under all three Reidemeister moves, and hence defines an invariant of oriented links.*

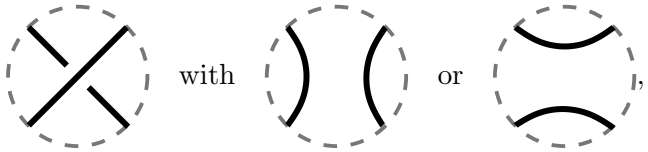
Proof. Certainly it is invariant under R2, R3 since both the writhe and bracket are. All that remains is R1. If a diagram D is altered by the addition of a positive kink somewhere, then its Kauffman bracket multiplies by $(-A^3)$ and its writhe increases by 1; therefore $f_D(A)$ is unchanged. Similarly for the negative kink case. □

This polynomial $f_D(A)$ is (once we make a certain change of variable) the Jones polynomial. Let us put off further examination of its properties just for a moment.

4.3. **A state-sum model for the Kauffman bracket.**

It may not be completely clear that the Kauffman bracket is really well-defined by the axioms we gave earlier. It is worth thinking a little more about how to compute it, as this will be important later and also gives a better idea of the computational difficulties involved.

Definition 4.3.1. A *state* s of a diagram D is an assignment of either $+1$ or -1 to each crossing. Clearly a c-crossing diagram has 2^c states. Given a state s on D, we may form a new diagram sD by *resolving* or *splitting* the crossings of D: this means replacing



according as the state takes the value $+1$ or -1 at the crossing. Thus, sD is a crossingless diagram, consisting simply of a certain number of disjoint loops: denote this number by $|sD|$. For a state s, let $\sum s$ denote the sum of its values.

Remark 4.3.2. The value of $\sum s$ is between $-c$ and $+c$, but it always has the same parity as c.

Remark 4.3.3. If s, t are two states on a diagram D differing only at one crossing, then $|sD| = |tD| \pm 1$, because changing which way that crossing is resolved either joins two previously disconnected loops, or splits a previously connected loop in two.

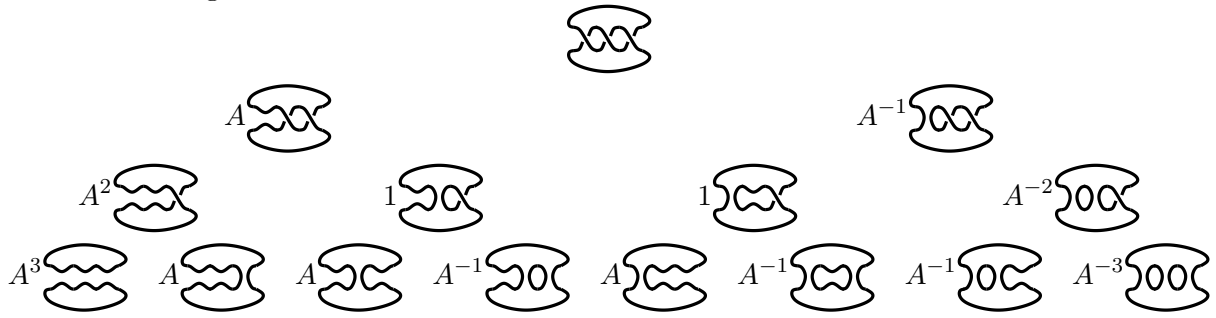
Proposition 4.3.4. *The Kauffman bracket can be expressed by the explicit "state-sum" formula*

$$\langle D \rangle = \sum_s \langle D|s\rangle,$$

where s runs over all states of D, and $\langle D|s\rangle$ denotes the contribution of the state s, namely

$$\langle D|s\rangle = A^{\sum s}(-A^2 - A^{-2})^{|s(D)|-1}.$$

Proof. Number the crossings of D from 1 to c. Apply the skein relation at crossing 1, reducing $\langle D\rangle$ to a linear combination of the brackets of two other diagrams, each with crossings numbered from 2 to c. Apply the skein relation to each of these diagrams at the crossing numbered 2: we now have a linear combination of four diagrams, each with crossings numbered from 3 to c. Repeating, we boil down to a linear combination of 2^c brackets of crossingless diagrams, indexed by states s in the obvious way, and each with a prefactor $A^{\sum s}$. Finally an n-component crossingless diagrams has bracket $(-A^2 - A^{-2})^{n-1}$. For example, a trefoil diagram with crossings numbered left to right gives the following *skein tree*:



The eight terms at the bottom have $2, 1, 1, 2, 1, 2, 2, 3$ loops, so we end up with the sum

$$A^3(-A^2-A^{-2})+A+A+A^{-1}(-A^2-A^{-2})+A+A^{-1}(-A^2-A^{-2})+A^{-1}(-A^2-A^{-2})+A^{-3}(-A^2-A^{-2})^2$$

which equals $-A^5 + A^{-3} + A^{-7}$. Eh? This is not the same as the value we got earlier for the trefoil, in fact its powers of A are exactly the negatives of what we had before! But this is not a mistake:

the trefoil diagram above is of the *right trefoil*, whereas earlier we evaluated a diagram of the left trefoil. This is a very interesting hint that the Kauffman bracket does "see" mirror images, and we will examine this more carefully after we construct the Jones polynomial.

But to conclude the proof: really all we have done is fix some ordering for applying the skein relations and introduced notation so as to write an explicit formula for the result; having done that, we see that the formula does not involve the chosen ordering of crossings at all. It's also clear that if we were to remove circles at earlier stages, we'd get the same result. So we do indeed have a well-defined Kauffman bracket, not depending on the order of application of skein relations. □

Remark 4.3.5. An alternative treatment of the bracket would be to define it straight-off via the state-sum formula, and then prove that it satisfies the skein relations. But whatever we do, we'll need to use both the state-sum and the skein relations in developing the theory further.

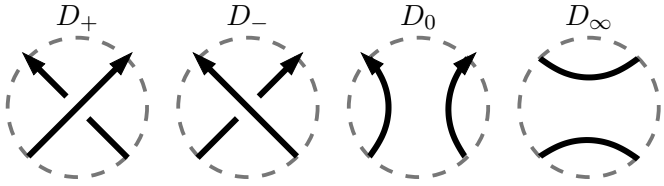
Remark 4.3.6. The computation of the Kauffman bracket is something which can easily be *programmed* on a computer, but is less easily actually *carried out*. For a c-crossing diagram it involves 2^c terms being added, hence 2^c operations, and therefore is an "exponential time" computation. If c is 100, for example, it looks as if it might take longer than the age of the universe to run! Contrast this with the algorithm for computing the number of p-colourings τ_p of the diagram, which is just Gaussian elimination on a $c \times c$ matrix: this takes of the order of c^2 operations, which is very fast even when c is enormous. (A human can compute τ_p of a 10-crossing diagram without too much difficulty, but will go insane trying to compute its bracket using the state-sum!)

Exercise 4.3.7. (2000F) (i). Define the *Kauffman bracket* polynomial $\langle D \rangle \in \mathbb{Z}[A^{\pm 1}]$ of an unoriented link diagram D. State Reidemeister's theorem about diagrams of equivalent links.

(ii). Prove that the Kauffman bracket is invariant under the Reidemeister moves of types II and III.

(iii). Define the *writhe* $w(D)$ of an oriented link diagram D. Show that the function $f(D) = (-A^3)^{-w(D)}\langle D \rangle$ defined on oriented link diagrams is an invariant of oriented links.

(iv). If D_+, D_-, D_0, D_∞ are four link diagrams differing only locally as shown below, and the first three are oriented as shown, find a linear relation between the polynomials $f(D_+), f(D_-), f(D_0)$.



4.4. **The Jones polynomial and its skein relation.**

Definition 4.4.1. The *Jones polynomial* $V_L(t)$ of an *oriented link* L is the polynomial obtained by computing $f_D(A) = (-A^3)^{-w(D)}\langle D\rangle$ for any diagram D of L, and then substituting $A = t^{-1/4}$.

The Jones polynomial therefore takes values in the ring $\mathbb{Z}[t^{\pm 1/4}]$ of integer-coefficient polynomials in a variable called $t^{1/4}$ and its inverse $t^{-1/4}$; in this ring, the symbol t is really just a shorthand for the element $(t^{1/4})^4$, and so on. The change of variables may seem bizarre and ugly, but we will see below why it is useful.

Because the basic variable in the ring is $t^{1/4}$, writing $V_L(t)$ for the polynomial is a sloppy piece of notation: the polynomial depends on $t^{1/4}$ rather than just t. You can't, for example, evaluate a polynomial like $3t^{1/4} + 2$ at $t = 1$ without making a choice of fourth root of 1. You may say that obviously we should choose $t^{1/4} = 1$ and avoid bringing complex numbers into things – but how would you evaluate the same polynomial at $t = -1$?

I'm afraid I will oscillate between the notations $V_L(t), V_L, V(L), V(t)$ according to whether I'm trying to emphasise the role of the link, the variable, or both. Sorry if this is confusing.

The following theorem will be key to understanding the Jones polynomial in its own right, separate from the construction via the Kauffman bracket.

Theorem 4.4.2. *The Jones polynomial satisfies*

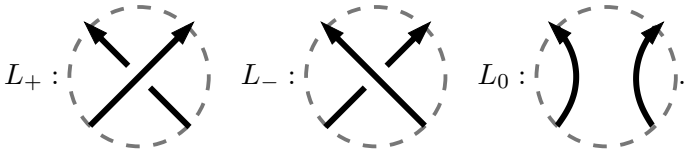
(1). *It is an invariant of oriented links lying in* $\mathbb{Z}[t^{\pm 1/2}]$.

(2). *The Jones polynomial of the unknot is* 1.

(3). *There is a skein relation*

$$t^{-1}V(L_+) - tV(L_-) = (t^{1/2} - t^{-1/2})V(L_0),$$

whenever L_+, L_-, L_0 *are three oriented links differing only locally according to the diagrams*



Proof. (1) Of course we know that the polynomial is an invariant of oriented links, but we are also asserting here that it takes only *half-integral* powers of t, and that the *quarter-integral* powers are not in fact needed. This is proved by looking at the definition via the Kauffman bracket state-sum. Each state contributes a term which is a power of $(-A^2 - A^{-2})$, giving all even powers of A, times $A^{\sum s}$, to the bracket $\langle D\rangle$. Since any sum over the crossings of $+1$ and -1's has the same parity as the number of crossings c in the diagram, the parity of Σs doesn't depend on the state, and we see that all the powers of A occurring in the state-sum are even or odd according as the number of crossings of the diagram is even or odd. For the same reason, the writhe $w(D)$ also has the same parity as $c(D)$, so after including the correction factor $(-A^3)^{-w(D)}$ we end up with only even powers of A. Therefore the Jones polynomial involves only powers of $t^{1/2}$ after all.

(2) This is trivial!

(3) We must compare the Kauffman brackets of the three diagrams D_+, D_-, D_0 in the Jones skein relation. It is convenient to define an additional diagram D_∞, the "horizontal" smoothing of the crossing (if D_0 is considered as the "vertical" one). Unlike D_0, this diagram does not have a natural orientation, because those induced from the rest of the link conflict with each other (it would require us to make arrows collide head-on!). So it does not make sense to speak of the oriented link

L_∞, let alone its Jones polynomial. However, unoriented diagrams do have a Kauffman bracket, which we can use as follows. By the Kauffman skein relation:

$$\langle D_+\rangle = A\langle D_0\rangle + A^{-1}\langle D_\infty\rangle,$$

$$\langle D_-\rangle = A\langle D_\infty\rangle + A^{-1}\langle D_0\rangle.$$

Multiply the first equation by A and the second by A^{-1} and subtract, to eliminate D_∞:

$$A\langle D_+\rangle - A^{-1}\langle D_-\rangle = (A^2 - A^{-2})\langle D_0\rangle.$$

Now substitute $\langle D\rangle = (-A^3)^{+w(D)}f(D)$ for each bracket, and note that the writhes of D_+, D_- are one more and one less than that of D_0. The result is

$$A(-A^3)^{w(D_0)+1}f(D_+) + A^{-1}(-A^3)^{w(D_0)-1}f(D_-) = (A^2 - A^{-2})(-A^3)^{w(D)}f(D_0)$$

$$\text{or} \quad -A^4 f(L_+) + A^{-4}f(L_-) = (A^2 - A^{-2})f(L_0),$$

which gives the Jones skein relation after the substitution $A = t^{-1/4}$ and a change of sign. □

What's nice about this theorem is that its three properties can be used *directly* to calculate the Jones polynomial, bypassing the Kauffman bracket/writhe construction completely. Let's work our way upwards with some calculations of increasing difficulty, by analogy with what we did when we first introduced the bracket.

Example 4.4.3. Our fundamental piece of information is that the Jones polynomial $V(U)$ of the unknot U is 1. When we looked at the bracket, the next things we looked at were the 1-crossing unknot diagrams, but there is no point looking at these in the world of the Jones polynomial; because it is an invariant of oriented links, and we know the 1-crossing diagrams represent knots equivalent to the unknot, they have Jones polynomial 1. The first non-trivial calculation we need to do is that of the 2-component unlink U_2.

(1). To compute $V(U_2)$, use the following trick: apply the skein relation to the three diagrams

L_+ :  L_- :  L_0 :  .

The first two are unknots, so again the isotopy invariance tells us that $V(L_+) = V(L_-) = 1$; the third is U_2. Therefore

$$t^{-1}.1 - t.1 = (t^{1/2} - t^{-1/2})V(U_2)$$

and so (dividing both sides by $t^{1/2} - t^{-1/2}$) we have find that $V(U_2) = -t^{1/2} - t^{-1/2}$.

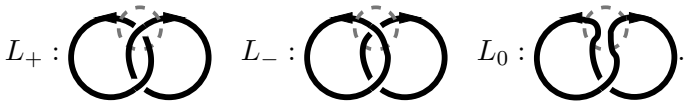
(2). There is a local version of the same trick. Consider the "skein triple" obtained by inserting positive and negative kinks into an arc of a given oriented link L, as shown below.

L_+ :  L_- :  L_0 :  .

Since the first two links are equivalent to L, and the third is the disjoint union $L \amalg U$ of L with an unknot, a calculation analogous to the previous one gives a formula for how the Jones polynomial of L changes under disjoint union with an unknot:

$$V(L \amalg U) = (-t^{1/2} - t^{-1/2})V(L).$$

(3). We can arrange for the positive Hopf link H_+ (the one with linking number $+1$) to be L_+, with L_- a 2-component unlink and L_0 an unknot.

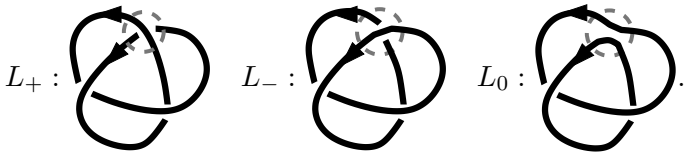


This gives

$$t^{-1}V(H_+) - t(-t^{1/2} - t^{-1/2}) = (t^{1/2} - t^{-1/2}).1$$

from which $V(H_+) = -t^{5/2} - t^{1/2}$. By an analogous calculation we find that the polynomial for the negative Hopf link is $V(H_-) = -t^{-5/2} - t^{-1/2}$. (The two oriented links are therefore inequivalent, although the linking number had already told us that).

(4). The right-handed trefoil T_R (the one whose standard diagram has positive writhe) can be arranged as L_+, such that L_- is an unknot and L_0 is the positive Hopf link whose Jones polynomial we just worked out.



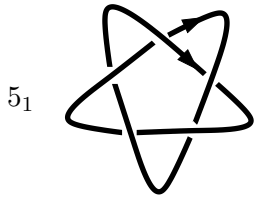
Thus

$$t^{-1}V(T_R) - t.1 = (t^{1/2} - t^{-1/2})(-t^{5/2} - t^{1/2}),$$

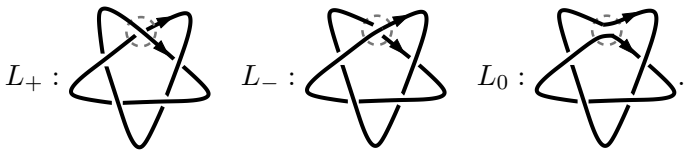
from which $V(T_R) = -t^4 + t^3 + t$.

(5) If we do the corresponding calculation for the left trefoil (do it!), we get $V(T_L) = -t^{-4} + t^{-3} + t^{-1}$; so *the left and right trefoils are inequivalent knots.*

Example 4.4.4. Let's calculate the Jones polynomial of the (right-handed version of the) knot 5_1, with choice of orientation shown below.

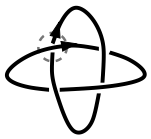
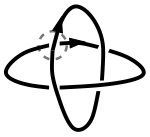


First pick a crossing to attack: we may as well choose the top one (the diagram is symmetric anyway). This is a positive crossing, so we view our 5_1 as the "L_+" in the skein relation. By switching and smoothing we obtain two new links: L_- is a right trefoil, whose polynomial we know; L_0 is a sort of double-strength Hopf link with linking number $+2$ – let's call it H_2 – whose polynomial we don't yet know.



$$t^{-1}V(5_1) - t(-t^4 + t^3 + t) = (t^{1/2} - t^{-1/2})V(H_2).$$

So we must now work out $V(H_2$. Choose a crossing and see what we get: H_2 will be the "L_+", while L_- is a standard positive Hopf link, and L_0 is a right-handed trefoil.

L_+ :  L_- :  L_0 :  .

Therefore

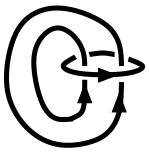
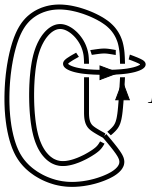
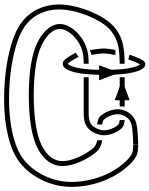
$$t^{-1}V(H_2) - t(-t^{1/2} - t^{5/2}) = (t^{1/2} - t^{-1/2})(-t^4 + t^3 + t).$$

After a bit of (easily messed-up!) manipulation, we end up with

$$V(H_2) = -t^{11/2} + t^{9/2} - t^{7/2} - t^{3/2},$$

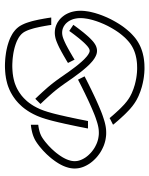
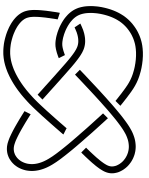
$$V(5_1) = -t^7 + t^6 - t^5 + t^4 + t^2.$$

Exercise 4.4.5. Calculate the Jones polynomial of the following links.

L_1 :  L_2 :  L_3 :  L_4 : 

Do the results suggest any interesting phenomenon?

Exercise 4.4.6. Calculate the Jones polynomials of the knots 4_1, 5_2 and 6_1.

4_1 :  5_2 :  6_1 : 

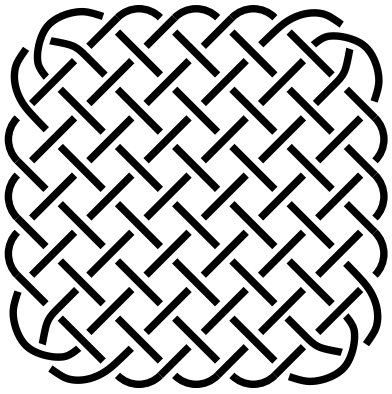
Remark 4.4.7. (1). You can check almost any calculation of a knot invariant on Chuck Livingston's invaluable knotinfo website `http://www.indiana.edu/~knotinfo/`

(2). While calculating Jones polynomials, it's a good idea to give a name to each link you need to work with, and *not* just refer to unknown new links as L_+, L_-, L_0, because after you've used the skein relation several times, you'll end up with multiple L_-'s, and so on, leading to appalling confusion! (Above, wherever I wrote a polynomial skein formula, I avoided using symbols like $V(L_0)$: I either substituted the known value, or gave the new link a new symbol.)

(3). Notice that although we started off computing Jones polynomials by building up a "library" of evaluations, going from simpler to more complicated ones, this is not the way we generally need to work. We usually begin with an unknown link, use the skein relation to break it down into simpler things, and recursively evaluate those. Nevertheless, the saved results of earlier calculations can save us a lot of time when we do this: it's annoying to have to keep working out the Hopf link polynomials over and over again! If you do build up a reasonable "library" you will find that using the Jones skein relation rather than the Kauffman state sum is actually *much quicker* in practice as soon as the knots get larger.

(4). The idea of using the skein relation to replace a "complicated" link by a pair of "simpler" ones seems clear, but if you think about it for a moment, it's *not* really clear what these notions

mean. Consider this diagram:



Where would we start applying the skein relation, and why? How would we continue? Is it clear that our "simplifying" process might not just go round in circles somehow? We'll sort this out in the next section but one.

4.5. **Some properties of the Jones polynomial.**

The following theorem is suggested by earlier calculations (for example, of the two trefoils).

Theorem 4.5.1. *The Jones polynomial of the mirror-image $\bar{L}$ of an oriented link L is the* conjugate *under $t \leftrightarrow t^{-1}$ of the polynomial of L. In other words,*

$$V_{\bar{L}}(t) = V_L(t^{-1}).$$

Proof. It is easy to see by thinking about the Kauffman skein tree that the bracket $\langle \bar{D} \rangle$ of the mirror of a diagram D is just $\langle D \rangle$ with A and A^{-1} exchanged. Additionally, mirror-imaging negates the writhe of any oriented diagram, since positive and negative crossings are exchanged. Therefore $f_{\bar{D}}(A) = f_D(A^{-1})$, and we are done. □

Corollary 4.5.2. *Any knot K whose Jones polynomial $V_K(t)$ is not* palindromic (*i.e. symmetrical under exchanging t and t^{-1}*) *is* chiral*, i.e. distinct from its mirror-image.*

Exercise 4.5.3. Compute the Jones polynomial of the figure-eight knot 4_1 in two ways: first via the Kauffman bracket, then using the Jones skein relation. Check that the results agree and are consistent with the figure-eight being amphichiral (equivalent to its mirror image)

Exercise 4.5.4. Give a formula for the Kauffman bracket of the connected sum of two diagrams $\langle D_1 \# D_2 \rangle$ in terms of $\langle D_1 \rangle$ and $\langle D_2 \rangle$. Derive a formula for the Jones polynomial of the connect-sum of two oriented knots. Work out the corresponding formula for the *disjoint* union of two knots.

Exercise 4.5.5. Do the Kauffman bracket and writhe depend on the orientation of a diagram? Show that the Jones polynomial of a *knot* doesn't depend on its orientation, but give an example demonstrating that this independence of orientation is *not* generally true for links with more than one component.

Exercise 4.5.6. Show that the Jones polynomial of any knot, evaluated when $t = 1$, is equal to 1.

Exercise 4.5.7. You probably solved some of the exercises in the last few sections using only the Kauffman bracket approach, and some of them using only the Jones skein relation approach. Contemplate them again from the *other* point of view each time! You may find some of them feasible but some impossibly hard! This will show you that one tool is often much more suitable than the other, and it's important to be able to work with either!

4.6. **A characterisation of the Jones polynomial.**

We can view the three "skein" properties of theorem 4.4.2 as *axioms* for the Jones polynomial which suffice to *characterise* it as the unique invariant with these properties:

Theorem 4.6.1. *Suppose I is a $\mathbb{Z}[t^{\pm 1/2}]$-valued function of oriented links which satisfies*

(1). *isotopy invariance – it is an invariant of oriented links*

(2). *the Jones skein relation $t^{-1}I(L_+) - tI(L_-) = (t^{1/2} - t^{-1/2})I(L_0)$*

(3). *the normalisation $I(U) = 1$.*

Then $I(L) = V(L)$ for all oriented links L.

As stated above, this is only a *uniqueness* theorem; it does not assert that any such function I actually exists at all, just that the number of such functions is at most 1 (i.e. 1 or 0)! But as we have already established – by means of the Kauffman bracket and theorem 4.4.2 – that there *does* exist (at least) one function – the Jones polynomial V – satisfying the above relations, it would be reasonable to fold this into the theorem and restate it to say "there exists a unique function satisfying the following three properties".

But quite often in mathematics, proofs of existence and uniqueness of some interesting structure are established by completely different methods; in this case the proof of the uniqueness part (that is, the theorem as stated above) has nothing at all to do with the Kauffman bracket, and that's why it seems more natural to state it independently.

The theorem can also be interpreted as saying that recursive application of the axioms in the manner of example 4.4.4 *does* actually suffice to calculate the Jones polynomial for any link; in fact this is what we really prove, but it's easier to state the theorem cleanly as above.

Let's warm up for the proof by considering formally the analogous (but much easier) situation for the Kauffman bracket.

Theorem 4.6.2. *Suppose that I is a $\mathbb{Z}[A^{\pm 1}]$-valued function on unoriented diagrams which satisfies*

(0). *planar isotopy invariance*

(1). *the Kauffman skein relation $I(D) = AI(D_0) + A^{-1}I(D_\infty)$*

(2). *the circle removal formula $I(D \amalg U) = (-A^2 - A^{-2})I(D)$*

(3). *the normalisation for the unknot diagram $I(U) = 1$.*

Then $I(D) = \langle D \rangle$ for all diagrams D.

Proof. Proceed by induction on the number of crossings of D. Any 0-crossing diagram D is simply a collection of n disjoint loops, so by using (0), (2) and (3) we find that $I(D) = (-A^2 - A^{-2})^{n-1}$, which agrees with the bracket $\langle D \rangle$. Now suppose that $I(D) = \langle D \rangle$ for all diagrams with fewer than c crossings, and that D has c crossings. Applying the skein relation at some crossing gives

$$I(D) = AI(D_0) + A^{-1}I(D_\infty);$$

but since the bracket satisfies the same skein relation

$$\langle D \rangle = A\langle D_0 \rangle + A^{-1}\langle D_\infty \rangle$$

and the inductive hypothesis is that $I(D_0) = \langle D_0 \rangle$ and $I(D_\infty) = \langle D_\infty \rangle$, we get $I(D) = \langle D \rangle$. □

I said it was easy! Unfortunately a similar induction on number of crossings won't work for the Jones polynomial because the skein relation does not strictly reduce the number of crossings: the L_- diagram has the same number of crossings as L_+. We need some more sophisticated notion of

complexity which measures not just the number of crossings of a diagram, but how far away it is from being an unlink.

Definition 4.6.3. (1). Define the *complexity* $\kappa(D)$ of a *diagram* D to be the ordered pair of integers (c, m), where c is the number of crossings of D, and m is the minimal number of crossing switches needed to change D into a diagram of an unlink.

(2). *Order* these complexities "lexicographically" by the rule

$$(a, b) < (c, d) \iff a < c \quad \text{or} \quad a = c \text{ and } b < d,$$

so that the m part acts as a tie-breaker if the more important c parts are equal.

(3). Define the *complexity* $\kappa(L)$ of a *link* L to be the minimal (with respect to this ordering) complexity $\kappa(D)$ of any diagram D representing it.

Lemma 4.6.4. *Given any oriented link L with complexity (c, m), there exists a diagram D and a choice of crossing C so that the three links L_+, L_-, L_0 associated to D and C are L and two new links of complexity strictly lower than (c, m).*

Proof. Take a minimal complexity diagram D for L: it will have c crossings, and we know that changing some set of m of these will give a diagram of an unlink. Let C be one of these m crossings; suppose (WLOG) that it is positive. Then D_+ is a diagram of L; switching and resolving C gives us two new diagrams D_-, D_0 representing new links L_-, L_0.

Consider the complexity of these new diagrams. Firstly, D_0 has $c - 1$ crossings, and although we don't know the minimal number of crossing changes m_0 it might take to turn it into an unlink diagram, this doesn't matter, since $(c - 1, m_0) < (c, m)$ whatever m_0 is, and hence $\kappa(D_0) < (c, m)$. Secondly, although D_- still has c crossings, it will take only $m - 1$ further crossing changes to turn it into an unlink diagram. Thus $\kappa(D_-) = (c, m - 1) < (c, m)$.

The complexities of the *links* $L-$ and L_0 are bounded above by those of (any of their diagrams, and in particular) D_- and D_0, so each has complexity strictly less than (c, m). □

Now we can prove the theorem.

Proof. Suppose that L is a link with complexity (c, m), and that we know $I(L') = V(L')$ for all links L' with complexity less than (c, m). By the lemma we can find a skein relation involving L (WLOG as L_+) where L_- and L_0 are of complexity less than (c, m). The skein relations for I and for V state

$$t^{-1}I(L) - tI(L_-) = (t^{1/2} - t^{-1/2})I(L_0)$$
$$t^{-1}V(L) - tV(L_-) = (t^{1/2} - t^{-1/2})V(L_0)$$

but by the inductive hypothesis we know $I(L_-) = V(L_-)$ and $I(L_0) = V(L_0)$. Hence $I(L) = V(L)$.

It remains to discuss the base of the induction. Links of complexity $(0, 0)$ are those which can be drawn with no crossings, that is unlinks U_n or arbitrary numbers of components $n \geq 1$. But it's easy to prove that $I = V$ for these by evaluating each using the trick from example 4.4.3. □

Exercise 4.6.5. Suppose L_+, L_-, L_0 are links differing just at one crossing, as in the skein relation, and that L_+ has μ components. What are the possibilities for the number of components of L_- and L_0? Show that for links with an odd number of components (including knots) the Jones polynomial contains only integral powers of t and t^{-1}, and for links with an even number of components it contains only half-integral powers (i.e. $\ldots, t^{-\frac{1}{2}}, t^{\frac{1}{2}}, t^{\frac{3}{2}}, \ldots$). (Hint: use induction again. Do you think it is possible to prove this result by using only about the Kauffman bracket state-sum, not the Jones skein relation?)

Exercise 4.6.6. (2007F) Give the (usual) definitions of the *complexity* $\kappa(D)$ of an unoriented link diagram D and the complexity $\kappa(L)$ of an oriented link L.

(b). What are the links with complexity $(0, 0)$?

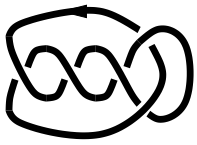
(c). The Jones polynomial is a function on oriented links satisfying the skein relation $t^{-1}V(L_+) - tV(L_-) = (t^{1/2} - t^{-1/2})V(L_0)$, where L_+, L_-, L_0 is a skein triple. Prove by induction on complexity that the Jones polynomial of a link with an odd number of components has no half-integral powers of t, while that of a link with an even number has only half-integral powers.

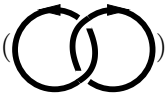
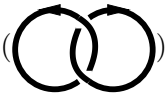
Exercise 4.6.7. (2003F) (i). Write down the axioms which define the Jones polynomial of oriented links. (Don't use the Kauffman bracket.)

(ii). Use these axioms to calculate the Jones polynomial of the knot 5_2 shown below. To speed up your calculation, you may use without proof the evaluations I've given you.

(iii). Is this knot amphichiral?

5_2



$V($  $) = -t^{\frac{1}{2}} - t^{-\frac{1}{2}}$ $V($  $) = -t^{\frac{5}{2}} - t^{\frac{1}{2}}$ $V($  $) = -t^{-\frac{5}{2}} - t^{-\frac{1}{2}}$

4.7. **How powerful is the Jones polynomial?**

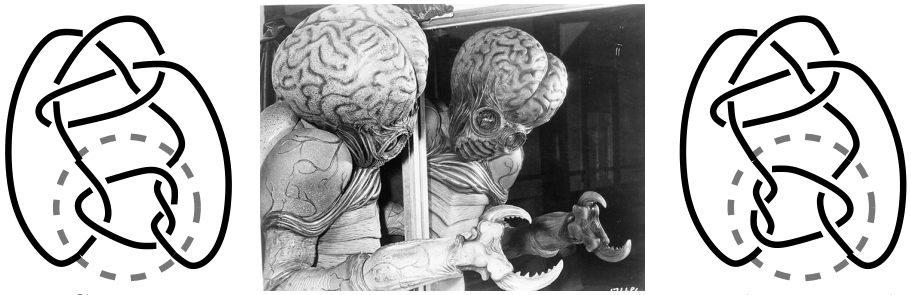
Recall that an invariant i might satisfy $i(K_1) = i(K_2)$ even when K_1 and K_2 are known (using some *other* invariant) to be inequivalent: we say "i fails to distinguish K_1 and K_2." For example, τ_3 can't distinguish the figure-eight from the unknot (since $\tau_3(4_1) = \tau_3(U) = 3$), whereas τ_5 can (since $\tau_5(4_1) = 25$ and $\tau_5(U) = 5$.) If this happens often, we think of an invariant as *weak*; if it is rare, the invariant is *strong*. Ideally, of course, we would like to find an invariant which *never* fails to distinguish different knots.

If we look at the standard table of 249 knots with 10 crossings or fewer and evaluate the invariant τ_3, we get *only 5 different values* (3,9,27,81,243). On the other hand, the same tabulation for $V(t)$ gives *248 different values*, with only one pair of distinct knots, 5_1 and 10_{132}, having the same polynomial. The Jones polynomial, therefore, is as strong as the 3-colouring number is weak!

(Since the standard table does not list knots and their mirrors (if distinct) separately, this is actually an underestimate of the difference in power. Most of the knots – 229 out of 249 – have distinct mirror images; but τ_3 is completely oblivious to this, while the Jones polynomial distinguishes each knot from its mirror with just two exceptions: 9_{42} and 10_{125}.)

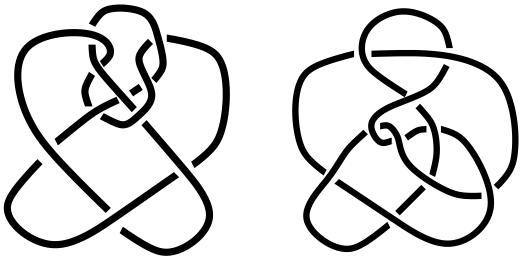
Exercise 4.7.1. Recall from exercise 4.1.12 that the Kauffman bracket does not change under mutation. What about the writhe, if the diagram is oriented? Prove that if an oriented link L is changed by mutation (of one of its diagrams D) into a new oriented link L', then the Jones polynomials of L and its mutant L' are equal: *the Jones polynomial cannot distinguish different links which are related by mutation.*

Many other knot invariants we know share this defect, so mutant pairs are relatively hard to distinguish. The most famous example of a pair of (distinct) mutants is the following: the Conway and Kinoshita-Terasaka knots (also known as $11n34$ and $11n42$).



Conway Kinoshita-Terasaka

(On the subject of knots which are difficult to tell apart, here is another famous example. In Rolfsen's classic 1976 book *Knots and links* there is a table of all the distinct knots up to 10 crossings which compiles the information worked out (by hand, in the pre-computer era!) by many mathematicians over the years. It contains the following knots numbered 10_{161} and 10_{162}.



These are in fact *the same knot*, as was discovered by Kenneth Perko, and they are therefore known as *the Perko pair*, even though there's only one of them!)

Although the Jones polynomial cannot distinguish all knots, the following problem is still open:

Conjecture 4.7.2. *The unknot is the unique knot K with $V(K) = 1$.*

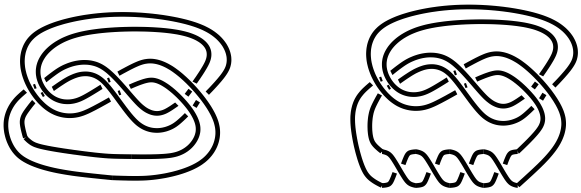
If this is true, then the Jones polynomial can *detect the unknot*: in other words, if you are given an unknown, complicated-looking knot, then all you need to do to decide whether it is or isn't unknottable is to compute its Jones polynomial (this can be done from any old diagram, of course) and see whether or not it is 1!

We don't have any particular reason for believing that the conjecture is true (Vaughan Jones originally posed it simply as a question), but we've had plenty of time since 1984 to look for counterexamples (a non-trivial knot K with $V(K) = 1$) and never found any. This makes people feel that maybe there really *aren't* any! There have been attempts to *disprove* the conjecture by constructing counterexamples, but without success. For example, suppose that we constructed a complicated diagram D of the unknot which could be mutated into a new diagram D'. Since mutation doesn't change the Jones polynomial, the link L represented by D' would have $V(L) = V(U) = 1$. However, it can be proved using geometric topology that L is always *itself* an unknot, so this method could never disprove the conjecture.

Mathematicians often have hard-to-justify "gut feelings" about whether open problems are true or false. For example, many years before Perelman finally proved it in 2003, almost all topologists believed that the Poincaré conjecture was true. I think that as of 2012, most people also think that the above Jones conjecture is most likely to be true, perhaps because we seem to be getting closer.

(Reduced) Khovanov homology, invented by Mikhail Khovanov in 1998 and closely related to the Jones polynomial, has very recently been proved to detect the unknot. Khovanov's invariant is quite sophisticated and beautiful, but the simplest way to express it is as a *two-variable* polynomial invariant of knots $\mathrm{Kh}_K(q,t) \in \mathbb{Z}[q^{\pm 1}, t^{\pm 1}]$. The Khovanov polynomial *determines the Jones polynomial* via the substitution $V_K(q) = \mathrm{Kh}_K(-q^{1/2}, -1)$. The Khovanov polynomial of the unknot is 1; Kronheimer and Mrowka proved in 2011 that it is the unique knot with this property.

Another reason is a theorem of Jørgen Andersen that the collection of all *cabled Jones polynomials* detects the unknot. For any knot K, an *n-cabling* of K, written K^n, is an n-component link obtained by simply replacing the original single strand with n parallel ones. If you take any diagram and draw n parallel curves all around, as in the left-hand picture of a 2-cable of the trefoil, you get a diagram of a cabling.



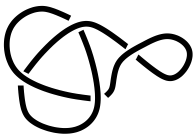
The right-hand picture is *also* a 2-cabling of the knot; the problem is that the number of twists of the strands around each other has not been specified. (You can see that the cabled link depends on the diagram: Reidemeister moves 2 and 3 do not change it, but R1 does.) The standard convention is to ensure that the pairwise linking numbers are zero by starting from a diagram with writhe zero.

Any invariant i of links may always be composed with the operation of cabling: we can define $i^n(K) = i(K^n)$. Even if i fails to distinguish K_1 and K_2, it's possible that i^n might for some n. Thus, each invariant gives rise to a *family* of *n-cabled invariants*. Andersen's theorem is that the family of all n-cabled Jones polynomials, for all $n \geq 1$, collectively detects the unknot.

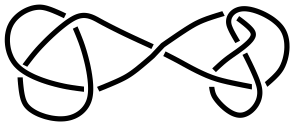
4.8. **Alternating knots and the Jones polynomial.**

Definition 4.8.1. Recall that a knot *diagram* is *alternating* if one passes alternately over and under crossings as one moves around the knot. A *knot* is alternating if it has *some* alternating diagram (it will always have non-alternating diagrams too).

A natural conjecture which energes when playing with knot diagrams is that *alternating diagrams are minimal*, i.e. that any knot represented by an alternating diagram cannot be represented by any other diagram with fewer crossings. (The up-and-down weaving pattern just seems for some reason to tie things together in a way that can't be simplified.) However, this overly naive conjecture is wrong, as a very simple counterexample shows:



Second attempt: *every alternating diagram which doesn't have a kink is minimal.* This is false too: in the picture below, half of the whole diagram could be turned over (a sort of "macro" version of R1) to effect a reduction in number of crossings.



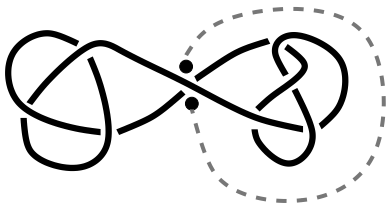
Third attempt: *every alternating diagram which doesn't have a macro-kink like this is minimal.* There are worrying signs here that we're entering into a runaway loop of finding counterexamples and then adding hypotheses to the conjecture, in an ad hoc way, so as to rule them out. But in this case we are lucky! It's not easy to find a counterexample to this third version, and in fact it is the most famous of the *Tait conjectures*, made about 120 years ago. No progress was made on any of these until the advent of the Jones polynomial, after which they were all proved by Murasugi, Kauffman, Thistlethwaite, and Menasco. In this section we will prove the (above) Tait conjecture.

Remark 4.8.2. The exploration above is a wonderful example of how research in mathematics is really done. We begin with some naive observation and/or intuition which makes us think that perhaps *every X is P*. We look at examples of Xs and see whether they are all P. If some aren't, we try to spot patterns in these, leading to a refined statement *every X which is also Q is P*. Perhaps we can almost prove this, modulo a gap which we could bridge if we assumed X is also R. If after further work we can't show that every X is R, we'd better add that X is R to our list of hypotheses for the theorem. But even now we've got a proof, it's possible that somebody will say "but what about C? It is an example of an X which is Q and R, but not P!" You realise that you were only ever thinking about Xs which are S: that your proof *implicitly* uses that X is S, but that the counterexample C to your theorem is not S. So, adding the previously implicit assumption that X is S to the list of hypotheses, honour is saved – until someone finds another problem, and around we go. (See "Proofs and Refutations" by Lakatos for more such philosophy.)

Back to the Tait conjecture. We need some proper terminology:

Definition 4.8.3. (1) A *region* of a link diagram is a connected component of the complement of the knot projection. We include the infinite "outside" of the diagram as one of the regions; this might seem a bit anomalous, but it becomes much more sensible if we imagine diagrams as drawn on a *sphere* rather than the plane; in that case no region is special. (Recall that removing a point from a sphere gives us back the plane.)

(2) An *isthmus* of a link diagram is a crossing at which there are fewer than four distinct regions incident. This implies that one can move, in the plane from a point in one of the quadrants incident at the crossing to a point in the opposite quadrant, without hitting the diagram again. Therefore every isthmus is, as its name suggests, a unique bridge between two separate pieces of diagram. A diagram is *reduced* if it has no isthmi. Any unreduced diagram can be made reduced by repeatedly flipping over one part of the diagram, destroying an isthmus (or "removing a macro-kink" as I temporarily described it earlier), until there are none left. (The property of being reduced is usually easy to observe in a diagram.)



(3) The *span* or *breadth* of a (Laurent) polynomial in a variable A is the difference between its highest (most positive) and lowest (most negative) powers of A appearing. For example, the span of $-A^3+2+A^{-3}-3A^{-4}$ is 7, as is that of $A^{12}-2A^9-A^6+3A^5$. The second polynomial here is just the first multiplied by $-A^9$, illustrating the fact that multiplying by a scalar (here -1) and/or a power of A doesn't change the span. The consequence of this is that although the Kauffman bracket is not invariant under R1, its span is, and hence *the span of the Kauffman bracket is a topological invariant of unoriented links*! All the results below will involve the A-span of the Kauffman bracket, but we could rewrite them in terms of the t-span of the Jones polynomial, which is exactly one quarter of the A-span of the bracket (because $A = t^{-1/4}$).

(4) A *minimal* diagram is one whose number of crossings equals the crossing number of the knot, so that the knot has no diagrams with fewer crossings. (There could be several other diagrams with the same number – we are not claiming that a minimal diagram is actually unique.)

Here are the theorems we will prove, and their corollary, Tait's conjecture.

Theorem 4.8.4. *For any c-crossing knot diagram D, the span of $\langle D\rangle$ is less than or equal to $4c$.*

Theorem 4.8.5. *For any c-crossing reduced alternating knot diagram D, the span of $\langle D\rangle$ is $4c$.*

Corollary 4.8.6. *Any reduced alternating knot diagram is minimal.*

Proof. (Of corollary) If our given reduced alternating diagram D has c crossings, then the span of the Kauffman bracket of the knot represented by D equals $4c$, by the second theorem. However, the *span* of the Kauffman bracket is a knot invariant and so if there were a diagram of the same knot with fewer than c crossings, the first theorem would be contradicted. □

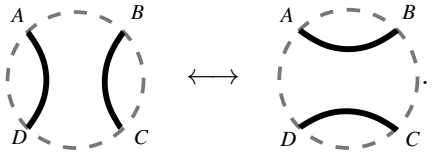
This corollary can be restated and specialised in more catchy ways:

Corollary 4.8.7. *Any non-trivial reduced alternating knot diagram represents a non-trivial knot.*

Corollary 4.8.8. *All reduced alternating diagrams of a knot K have the same number of crossings.*

To prove the theorems, we need to identify the highest and lowest powers occurring in the bracket of a c-crossing diagram D, and for this we think about the state-sum model. Let s_+ and s_- be the states consisting entirely of pluses and minuses respectively; it seems reasonable that s_+ might contribute the highest positive power, and s_- the highest negative power, because for these states $\sum s = \pm c$ is extremal. What we need to understand is how $|sD|$ changes as s ranges over all states.

Here is a very simple observation about the number of loops: if two states s and t differ at a single crossing, then $|sD| = |tD| \pm 1$. This is because the smoothed, crossingless diagrams differ by a single local "switching" alteration:



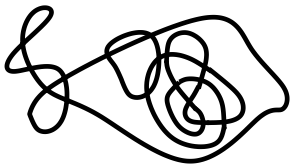
Now the unseen, outside part of the diagram, common to both sides, either connects $A \leftrightarrow B$ and $C \leftrightarrow D$, or $A \leftrightarrow D$ and $B \leftrightarrow C$; it cannot join $A \leftrightarrow C$ and $B \leftrightarrow D$ because this would require there to be a crossing. So the effect of switching is always to break one component into two, or to combine two components. The number of components must therefore alter by 1. We can iterate this observation: if two states s and t differ at k crossings then $|sD| \leq |tD| + k$.

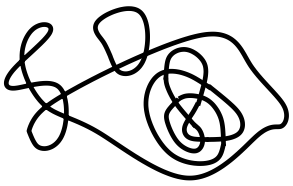
Proof. (of theorem 4.8.4) Recall the notation $\langle D|s\rangle = A^{\Sigma s}(-A^2 - A^{-2})^{|sD|-1}$ for the term in the state-sum corresponding to state s. The highest power of A appearing in $\langle D|s_+\rangle$ is $c+2(|s_+D|-1)$, which we will denote by H. Suppose we move from s_+ to a state s_1 which has one crossing labelled -1 and all the rest $+1$. The highest power of A in $\langle D|s_1\rangle$ is $(c-2)+2(|s_1D|-1)$, but because $|s_1D| = |s_+D| \pm 1$, this is either H or $H-4$. This argument also shows that whenever we go from a state to a state with one more minus label, the highest power in its term cannot increase. Therefore H is the highest power which could occur in *any* term of the state-sum. By symmetry, we can show that the lowest power L which could occur is $L = -c - 2(|s_-D| - 1)$.

Now the span of the bracket is bounded above by $H - L = 2c + 2(|s_+D| + |s_-D| - 2)$, so all we need to finish the proof is to show that $|s_+D| + |s_-D| \leq c + 2$. This is a purely combinatorial fact about planar diagrams, no longer having anything to do with the bracket, and it comes from applying the Dual State Lemma (below) to D and the state s_+.

(Why only *bounded above*, and not *equal* to $H-L$? Well, there could be *cancellation*: although we know for sure that there is a power of A^H occurring in $\langle D|s_+\rangle$, such powers could also occur (perhaps with negative coefficients) in other terms, and hence might cancel out in the overall sum, causing a possibly drastic drop in maximum degree of $\langle D\rangle$. A trivial example: $(x^3+2x+1)+(-x^3-x+3) = x+4$ is a sum of two polynomials of degree 3 which only has degree 1!) □

Definition 4.8.9. A diagram is *connected* if its underlying projection is a connected subset of the plane. (It is irrelevant whether the diagram is of a link or a knot, though clearly any knot diagram is connected.)

Connected: 

Disconnected: 

Lemma 4.8.10 (Dual state lemma)**.** *For any state s, let $\hat{s}$ denote its* dual *or* opposite*, given by exchanging all pluses and minuses.For any* connected *diagram D and any state s, we have*

$$|sD| + |\hat{s}D| \leq c + 2.$$

Proof. The key idea is that for any connected diagram D, we can always construct (in at least one way) a state t for which tD consists of a single loop. The construction is quite easy: simply resolve the crossings of the shadow one at a time, each time making a choice of vertical or horizontal resolution which ensures that the new shadow remains connected. This can always be done, because

if there were a crossing such that neither the horizontal nor vertical smoothing gave a connected shadow, then (look at the last figure but one) the string leaving at the top left A could not return at either of the bottom corners C, D (as this would make the horizontal smoothing connected) or at either of the right-hand two B, C (which would make the vertical smoothing connected), giving a contradiction. On finishing we get a connected crossingless shadow – that is, a single loop.

Now let k be the number of crossings at which s and t differ. We can transform sD to tD by making k switches, each of which alters the number of loops by ± 1. Hence $|sD| \leq |tD| + k$. Similarly, we can transform $\hat{s}$ to t by making switches at the *complementary* set of $c - k$ crossings, showing that $|\hat{s}D| \leq |tD| + c - k$. Finally add the two equations and recall that $|tD| = 1$. □

Proof. (of theorem 4.8.5) The proof of 4.8.4 we gave above, which applied to *all* diagrams, combined two inequalities:

(a) $\text{span}\langle D \rangle \leq H - L$

(b) $|s_+D| + |s_-D| \leq c + 2$

If we could show that each of these was an *equality* in the case of a reduced alternating knot diagram, then the same proof would give the desired equality $\text{span}\langle D \rangle = 4c$.

As noted above, the inequality (a) reflected possible cancellation of extremal terms in the state-sum. We knew that s_+ contributed an A^H term, but a one-minus state s_1 could also contribute this same power if $|s_1D| = |s_+D| + 1$. If we can show that this never happens (that is, $|s_1D| = |s_+D| - 1$ for all one-minus states), then A^{H-4} is the highest power of A coming from any state other than s_+; no cancellation can occur, and so the maximal power would be exactly H. (We'd also need to show the dual statement, that a state s_1 with one plus satisfies $|s_1D| = |s_-D| - 1$, so as to make the lowest power precisely L.) The lemma below will show that this is the case for a reduced alternating diagram, and that (b) is also an inequality, which will finish the proof. □

Lemma 4.8.11. *For a reduced alternating knot diagram ,*

(i) $|s_+D| + |s_-D| = c + 2$

(ii) $|s_1D| = |s_+D| - 1$*, where* s_1 *is any state which has exactly one minus*

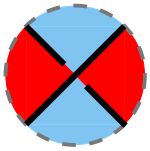
(iii) $|s_1D| = |s_-D| - 1$*, where* s_1 *is any state which has exactly one plus.*

Proof. First consider the nature of the regions of a general link diagram. For a *disconnected* diagram, there might be regions with more than one boundary component: topologically these are annuli, or discs with multiple interior holes. But for a *connected* diagram (in particular a knot diagram), the regions are all topologically discs, and each one has just one boundary component, a polygon (interpreted liberally: it can have 1 or 2 sides!) having crossings as vertices and (curved) arcs of the knot joining crossings as edges. (The outside region, as noted before, is only anomalous if we think on the plane instead of the sphere: on the sphere, it also becomes simply a disc with a polygon as boundary).

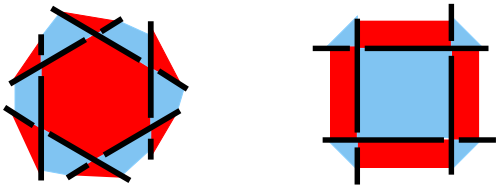
Recall also (from exercise 2.3.10) that the number of regions in a knot diagram is the number of crossings plus 2. (My favourite proof of this is as follows: replace the knot with its "lazy man" unknot, which has the same projection, crossings and regions. Now use Reidemeister moves to unknot it, and check that the numbers of regions and crossings both change in the same way under the three moves. Finally, the result holds for a 0-crossing circular diagram.)

(Both these observations can also be proved more slickly using Euler characteristic arguments from the chapter on classification of surfaces.)

Now consider alternating knot diagrams. Many of their nice properties come from the following simple construction. Colour the corners of the regions incident at each crossing either red (dark) or blue (light), according to the picture

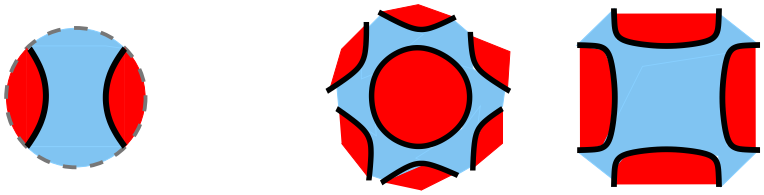


The property of *alternation* means that the patches of colour assigned to the corners of each region are the same: consequently, each region gets a well-defined colour, red or blue, depending on which way its boundary crossings "circulate", as in the following pictures.



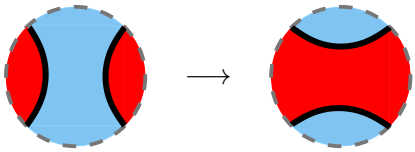
We usually call this kind of colouring a "chessboard shading", for obvious reasons. (I've called the colours red and blue rather than black and white because when I'm at the blackboard, shading a black region with white chalk, and you're at your desk, shading a white region with black ink, things get very confusing!)

On resolving the diagram according to s_+, we get a crossingless diagram with coloured regions. Near a crossing it looks like the first picture below, and around each region it looks like the second or third, according to the "circulation" of the boundary crossings.



Thus, there is one loop in s_+D for each red region of the original coloured diagram. Dually, there's one loop in s_-D for each blue region. Since the total number of regions in a connected diagram is equal to $c+2$, we have proved (i).

Moreover, if s_1 is any state with just one minus then s_1D differs from s_+D near just one crossing according to the picture



Since the original diagram was *reduced*, the two red areas on the left here belong to *distinct* regions (otherwise the crossing would have been an isthmus), and therefore on the right, they (and their two boundary loops) have been merged together. Therefore $|s_1D|$, which is the number of loops in the right-hand figure, is one less than $|s_+D|$, the number of loops on the left, proving (ii). Similarly, considering s_- and blue regions, we get (iii). □

Exercise 4.8.12. Show that the crossing number of an alternating knot is equal to the span (in the variable t) of its Jones polynomial. (Note: I said alternating *knot*, not alternating *diagram*!)

Exercise 4.8.13. Show that the crossing number of the connected sum of two (oriented) alternating knots is equal to the sum of the crossing numbers of the two knots.

4.9. **Other knot polynomials.**

Here are three questions about working with the Jones polynomial using only the Jones skein relation, avoiding going back to the Kauffman skein method.

(1). *Can we give an actual algorithm for calculating the Jones polynomial using the Jones skein relation?* Although the uniqueness theorem implies that the axioms suffice to calculate the polynomial for any link, and we gave an *idea* of how to do this in the proof of uniqueness, we didn't specify an actual *algorithm* which could be implemented on a computer. Our procedure involved lots and lots of choices: choose a minimal diagram, choose one of the "unknotting" crossings in that diagram; resolve into two new links. Now repeat the whole process for those new links! It would be nice to pin down a less arbitary and more direct procedure for doing the calculation.

(2). *Why are the funny coefficients* $t, t^{-1}, (t^{1/2} - t^{-1/2})$ *in the Jones skein relation there?* The uniqueness proof we gave didn't depend on any special properties of these coefficients at all; it would still have worked if put independent variables x, y and z in their place. Although Kauffman's method explains them, can we find an explanation inside the Jones skein world?

(3). *Is it possible to prove directly using the Jones skein method, without using the Kauffman bracket, that the polynomial exists?*

(4). *Might there be other sets of skein-type axioms which define a polynomial invariant of links?* Suppose we claimed that there existed a unique polynomial invariant of oriented links satisfying some different bunch of axioms; when will this be true?

This last one is a question about the *consistency* of the axioms we define. For example, suppose we added a *fourth* axiom saying that $V(L \amalg U) = (-t - t^{-1})V(L)$. There's nothing unreasonable about this, on the face of it – in fact you might have worried that there wasn't an axiom for dealing free unknot components, as there had been in the Kauffman world. But we know from example 4.4.3 that the usual three axioms do give us such a rule *implicitly*, and that it clashes with the fourth axiom we added: the only way both can hold for all links is if $-t^{1/2} - t^{-1/2} = -t - t^{-1}$. This is clearly not true in $\mathbb{Z}[t^{\pm 1/2}]$, so there does *not* exist a polynomial-valued invariant satisfying these four axioms. However, if we choose a solution of this equation – a cube root of unity ω, as it turns out – then the $\mathbb{C}$-valued invariant $V(L)|_{t=\omega}$ does satisfy all four axioms! If we added as a *fifth* axiom that $V(t)$ multiplies under disjoint union of links then we would really be in trouble as there would be *no* function of any kind satisfying all these axioms: they have become *inconsistent*.

These questions are all closely related and can all be resolved by solving (1). Here is a precise algorithm, based on a limited number of explicitly stated choices, for performing the calculation of the Jones polynomial of any oriented link L.

(a) Pick an ordering of its components, a basepoint on each one, and choose a diagram D of L.

(b) Given these choices, there is now a canonically-defined diagram D^* of an unlink L^* obtained by converting each component of D into a standard descending projection (as we did when we defined unknotting number) starting from its given basepoint, and stacking the different components vertically over one another in order from lowest to highest.

(c) We can convert D into D^* by switching some of its crossings, and we can make these switches in a natural order, starting by moving around component 1 and switching the ones we find, then doing those around component 2 that are not already done, and so forth. Therefore we have a sequence of diagrams $D = D_0, D_1, D_2, \ldots, D_n = D^*$ where each consecutive pair forms the "L_+, L_-" of a skein triple, and there is a new link M_i arising from the smoothed diagram. The skein relations applied to these triples therefore express $V(L)$ in terms of $V(L^*)$ (which we know, as it is an unlink) and the polynomials $V(M_i)$ for these new links, each described by a diagram with one less crossing.

(d) The new links do come equipped with diagrams, but we need to tweak their basepoints and component ordering. If there is a new component with no basepoint, place one at the crossing, and place the component last in the ordering. If two components have merged, let the new single loop keep the basepoint and position in the ordering of the earlier-ordered of the two.

(e) Now we can recursively repeat for these simpler links M_i. Clearly we are constructing some sort of skein tree leading to a complete evaluation of $V(L)$.

What is required now is to answer question (3) by proving invariance of the choices of diagram (using Reidemeister moves), basepoint (by thinking about what happens when the basepoint is moved past a crossing), and component ordering (by considering what happens when some consecutive pair in the ordering is interchanged). This is quite hard to do, but it is worthwhile because it reveals the rather amazing fact that *the coefficients in the Jones polynomial skein relation may be arbitrary* and we still get an invariant; no consistency relation is imposed on them during the course of the above proof! Summing up:

Theorem 4.9.1 (Homogeneous HOMFLY polynomial)**.** *There is a unique invariant $P(L)$ of oriented links L, taking values in Laurent polynomials in three variables $\mathbb{Z}[x^{\pm 1}, y^{\pm 1}, z^{\pm 1}]$, which satisfies the skein relation $xP(L_+) + yP(L_-) + zP(L_0) = 0$ and the normalisation $P(U) = 1$.*

Proof. The existence proof (outlined above) is hard: see Lickorish's book. The uniqueness proof, however, is all-but-identical to that for the Jones polynomial. □

If you compute this polynomial you'll quickly notice that it is homogeneous in x, y, z, meaning all its terms have the same total degree – zero – of variables. (For the 2-component unlink, for example, we get $-(x + y)z^{-1}$.) This is easy to prove using induction and the fact that the skein relation itself is homogeneous (exercise!). But homogeneous polynomials in three variables are equivalent to inhomogeneous polynomials in two variables: if we set $z = 1$, for example, we get a polynomial in x and y but have lost no information: re-inserting powers of z to make the terms all have degree zero again gets us back where we started. It's actually nicer to set $xy = 1$ rather than $z = 1$, which leads us to the following inhomogeneous version of the polynomial:

Theorem 4.9.2 (Lickorish-Millett version of HOMFLY)**.** *There is a unique invariant $P(L)$ of oriented links L, taking values in Laurent polynomials in two variables $\mathbb{Z}[l^{\pm 1}, m^{\pm 1}]$ which satisfies the skein relation $lP(L_+) + l^{-1}P(L_-) + mP(L_0) = 0$ and the normalisation $P(U) = 1$.*

The name of the HOMFLY polynomial comes from its discoverers' initials: Freyd and Yetter, Lickorish and Millett, Hoste, and Ocneanu (working in two groups of two, and two solo) all discovered it more or less simultaneously in 1985. After finding out about each others' work, they wrote a collective paper about it; unfortunately they did not know at that point that Przytycki and Traczyk, working on the other side of the Iron Curtain, had *also* discovered it. To give them their proper credit, sometimes the polynomial is referred to as HOMFLY-PT.

Remark 4.9.3. There are many alternative ways to define the HOMFLY polynomial, some amazingly different from the skein method outline above, but none of them is as simple as the bracket method for computing the Jones polynomial.

Exercise 4.9.4. Calculate the HOMFLY polynomial of the two-component unlink and of the left- and right-handed trefoils.

Exercise 4.9.5. Show that the HOMFLY polynomial determines the Jones polynomial of a link.

Exercise 4.9.6. Setting $x = 1$ in the HOMFLY polynomial gives a polynomial $\nabla_L(z)$ of oriented links which is called the *Conway potential function* of a link. (Setting $z = t^{-1/2} - t^{1/2}$ in the Conway

polynomial gives the *Alexander polynomial* of the link, which is much older: Alexander defined it by a very different method in 1928.)

(i). Show by induction that for any link, the Conway polynomial lies in $\mathbb{Z}[z]$ (i.e. that it has no negative powers of z),

(ii) Show that it has all even or all odd powers of z according to the parity of the number of components of the link.

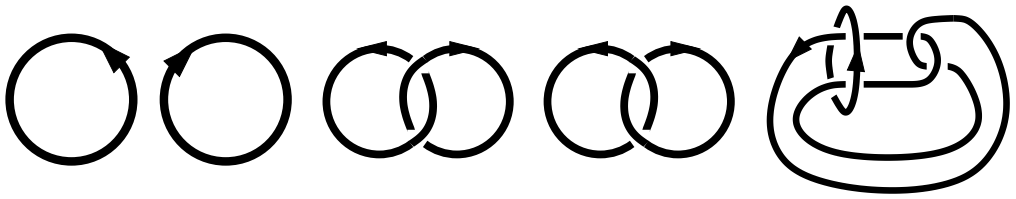
(iii) Show that for a *knot* K, $\nabla_K(z)$ always has constant term 1.

(iv) Show similarly that for a two-component link, there is never a constant term, but that the coefficient of z in the Conway polynomial equals the linking number of the link.

Exercise 4.9.7. (2005F?) The HOMFLY polynomial $P(L)$ may be defined by the three axioms: (a) it is a topological invariant of oriented links with values in $\mathbb{Z}[l^{\pm 1}, m^{\pm 1}]$; (b) the value of P of the unknot is 1; (c) it satisfies the skein relation

$$lP(L_+) + l^{-1}P(L_-) + mP(L_0) = 0,$$

where L_+, L_-, L_0 are three links differing locally in the same way as in the Jones polynomial skein relation. Using these axioms, compute the HOMFLY polynomials of the four links shown below:



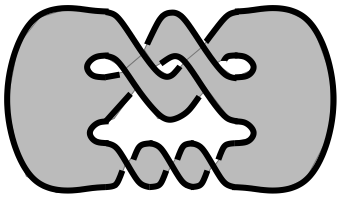
Exercise 4.9.8. How do the HOMFLY and Conway polynomials behave under reversing the orientations of all components of a link and under mirror-imaging a link?

Exercise 4.9.9. Prove that the values of the Jones polynomial and Alexander polynomial of a knot agree when $t = -1$. (This number is called the *determinant* of the knot.)

Exercise 4.9.10. Come up with three interesting questions about knots and links to which you don't know the answer. (The same question was asked in the introduction, but you know a lot more now! This is how mathematical research progresses - the more you know, the more you realise you don't know but want to find out about!)

5. A glimpse of manifolds

The knots and links we have studied so far are examples of compact 1-dimensional submanifolds of $\mathbb{R}^3$. We can also consider compact 2-dimensional submanifolds (perhaps with boundary) inside $\mathbb{R}^3$, in other words "knotted surfaces". There is a close connection between the two types of object, because the boundary of any knotted surface is a link, as shown below; and this relationship is very useful in studying knots.



The *intrinsic* topology of a link is not very interesting: from the perspective of an ant who lives on the link and can't see out, it is just a disjoint union of circles, and so is completely determined by its number of components (In fact a single ant can only explore *one* of the components, but let's assume there are ants living on all of the different loops!) Of course we are interested in the more subtle problem of understanding the ways in which these circles can sit inside $\mathbb{R}^3$.

The first thing we will do when studying surfaces is obtain a classification of their *intrinsic topology*, i.e. a list of possible homeomorphism types of compact 2-manifolds. Then we can begin to investigate the ways in which they can sit inside $\mathbb{R}^3$, and the relationship between knots and surfaces.

The classification of surfaces is a traditional part of a basic topology course, and it is usually done using the language of topological spaces, which comes earlier in the course. But typically, such descriptions resort at least at some point to dealing with surfaces built from triangles (in other words a combinatorial point of view) and do not prove the fact that topological surfaces can be triangulated.

In fact therefore none of the language of topological or metric spaces is necessary; the approach used below will be in fact be completely elementary and combinatorial, by analogy with how we handled knots earlier in the course.

However, in order to convince readers that it is not completely "handwavy", I feel it's important to at least begin by explaining briefly what the "true" definition of a surface should be. Readers who are already familiar with either the theory of metric spaces, or that of topological spaces, should be comfortable with what follows; everyone else should consider this an invitation to go away and learn some more about these topics (see the appendix) but need not feel obliged to really understand it fully. In a sense therefore, this section is completely optional; go directly to the next one (explaining the classification of surfaces) if you want, or even to the section after that where we use surfaces to study knots.

Here is the appalling official definition of a *manifold* (a *surface* is just a *2-dimensional manifold*):

Definition 5.0.11. An *n-dimensional manifold* is a second-countable Hausdorff topological space M, such that every point of M has a neighbourhood homeomorphic to $\mathbb{R}^n$.

The rest of this section contains brief explanations of the terms involved here, before we return and consider the definition in this light. In the next section we go combinatorial and redefine surfaces as being "polyhedral" in the sense of being made from triangles glued together.

5.1. **Metric spaces.**

The primary object of study in algebraic topology is the *topological space*. It is the most general kind of space in which one can do sensible analysis, by which I mean that the notions of continuity, limit etc. make sense. Let's begin working towards the definition by reciting the time-honoured definition of continuity for a real-valued function of a real variable:

Definition 5.1.1. A function $f : \mathbb{R} \to \mathbb{R}$ is *continuous at* $a \in \mathbb{R}$ if for every $\epsilon > 0$, there exists a $\delta > 0$ such that $|x - a| < \delta$ implies that $|f(x) - f(a)| < \epsilon$.

The intuition is, of course, that the function does not jump about locally: if one looks at a sufficiently small range of values about a, then the values of the function may be confined to be arbitrarily close to $f(a)$. Of course, a function is said without qualification to be *continuous* if it is continuous everywhere, i.e. for at all $a \in \mathbb{R}$.

Suppose now that we want to generalise to more complicated kinds of function, such as (let's not get carried away) a real-valued function of two real variables. Obviously the correct thing to do is repeat the same definition with the Euclidean distance $\|x - a\|$ replacing $|x - a|$ now that x, a are points of $\mathbb{R}^2$.

In fact the same principle will work to give a sensible definition of continuity of any function between subsets of a Euclidean space $\mathbb{R}^n$; all that is needed is the notion of distance between pairs of points. In this way, one can quite happily start talking about continuous functions between spheres of arbitary dimensions, because the n-sphere is usually thought of as simply the unit sphere inside the Euclidean space $\mathbb{R}^{n+1}$.

If we want to escape the confines of Euclidean space, it is necessary to abstract away the really essential aspects of Euclidean distance. It turns out that the most important thing is the *triangle inequality*: if you start writing out proofs of the simplest properties of continuous one-variable functions, you will need it pretty quickly. It is that for any three vectors $x, y, z \in \mathbb{R}^n$,

$$\|x - y\| \leq \|x - z\| + \|z - y\|.$$

Definition 5.1.2. A *metric space* is a set X equipped with a *metric* function $d : X \times X \to \mathbb{R}_{\geq 0}$ such that

(1). $d(x, y) = 0$ if and only if $x = y$

(2). $d(x, y) = d(y, x)$ for any x, y (symmetry)

(3). $d(x, y) \leq d(x, z) + d(z, x)$, for any x, y, z (triangle inequality).

The basic example is obviously $\mathbb{R}^n$ with the Euclidean distance function, as described above. With this notion, we can make an obvious definition of continuity:

Definition 5.1.3. A function $f : X \to Y$ between metric spaces is *continuous at* a if for every $\epsilon > 0$, there exists a $\delta > 0$ such that $d_X(x, a) < \delta$ implies that $d_Y(f(x), f(a)) < \epsilon$.

The axioms for the metric lead to the basic lemma that *the composition of two continuous functions is continuous.*

The next most obvious examples of metric spaces are subsets of $\mathbb{R}^n$, using the restriction of the distance function. Consider for example the unit sphere S^2 in $\mathbb{R}^3$; this distance function is the *chordal metric*, which measures the length of a straight-line chord ("as the mole burrows") joining the points inside the sphere.

On reflection, this is a bit tasteless, since to define it we used geometry external to the sphere! A nicer choice would be to use the great-circle distance on the sphere's surface, measuring "as the crow (or plane) flies". So we seem to have two reasonable choices of metric on S^2; which should we take as "the" metric?

The answer turns out to be: it doesn't much matter! Whichever metric we choose, the same functions into or out of the sphere turn out to be continuous. This is easy to prove, as a consequence of the fact that the two metrics d, d' on the fixed space X (in our case, S^2) are *Lipschitz-equivalent*, meaning there exist constants $K, k \geq 1$ such that for all points x, y,

$$\frac{1}{k} d(x,y) \leq d'(x,y) \leq K d(x,y).$$

In other words, the ratio between the distances, as measured using the two metrics, is bounded within some fixed band either side of 1. If this is the case then the identity map of X, considered as a map between the different metric spaces (X,d) and (X,d') or vice versa, is continuous (easy check). Consequently, if $f : (X,d) \to (Y,d_Y)$ is a map of metric spaces which is continuous, so is f thought of as a map starting from (X,d'), because this is the composite of the original f with the continuous identity map $(X,d') \to (X,d)$.

This example highlights a problem with the use of metric spaces as a foundation of the theory of continuity — it shows that the actual metric itself contains *far more information* than we need when simply thinking about continuity. The actual distances are not important to us; only the idea of continuity of functions! The notion of a *topological space* will be more economical: it will incorporate only what we actually need.

(Another reason for being unhappy with metric spaces is that not all the constructions we hope to perform with spaces work well. We can certainly take subspaces and products of metric spaces and get sensible induced metrics. But the notion of *quotient* is hopeless. Even *disjoint union* of two metric spaces is unpleasant: in $X \amalg Y$, we have perfectly sensible ways of measuring distance between pairs of points of X, and between pairs of points of Y. But what should be the distance between a point of X and a point of Y? Of course ad hoc definitions are available, but there is no canonical, choice-free method.)

5.2. **Topological spaces.**

To work further towards the definition of a topological space, it helps to rephrase the metric space definition of continuity, avoiding explicit dependence on the metric (which we are trying to get rid of).

Definition 5.2.1. Given a point x of a metric space X and a real number $\epsilon > 0$, let us define the *ball of radius ϵ at x* as

$$B_\epsilon(x) = \{y \in X : d(y,x) < \epsilon\}.$$

Definition 5.2.2. A set N is called a *neighbourhood* of a point $x \in X$ if it contains some ball $B_\epsilon(x)$ of positive radius about x.

Definition 5.2.3. A set U is set to be *open* if it is a neighbourhood of each of its points. Such a set can then be written in the form (check each direction of containment if this seems puzzling!)

$$U = \bigcup_{x \in U} B_{\epsilon(x)}(x).$$

This is a bit of a silly expression, from one point of view: we are writing a set as a union of small balls about all of its points in a very redundant way. However, the intuition that every open set can be expressed as some huge union of special kinds of standard small open sets is a valuable one.

Lemma 5.2.4. (*Local form*) *A function $f : X \to Y$ is continuous at $a \in X$ if and only if, for each neighbourhood N of $f(a)$, the inverse image $f^{-1}(N)$ is a neighbourhood of a.*

(*Global form*) *A function $f : X \to Y$ is continuous if and only if, for each open set U in Y, the inverse image $f^{-1}(U)$ is open in X.*

Proof. It's straightforward to check necessity and sufficiency in each case. □

This lemma then, removes the explicit dependence on the metric, as we desired — the *open sets* of a metric space provide enough information for us to talk about continuity. The conceptual leap to a topological space is then simply the realisation that we may as well only specify these open sets, rather than a metric. Remarkably, a few simple axioms suffice to make the structure behave (for the most part) in the way we have come to expect.

Definition 5.2.5. A *topological space* is a set X together with a set τ_X (its *topology*) of subsets of X, whose elements are referred to as the *open sets*, satisfying the axioms:

(1). the whole space X, and the empty set $\emptyset$ are open

(2). the union of an *arbitrary* family of open sets is again open

(3). the intersection of *finitely many* open sets is again open.

Definition 5.2.6. A function $f : X \to Y$ between topological spaces is *continuous* if for every open U in Y (i.e. $U \in \tau_Y$), the inverse image $f^{-1}(U)$ is open in X (i.e. $f^{-1}(U) \in \tau_X$).

I have chosen to write the "slogan" versions here, instead of emphasising the set-theoretic notation, which require one to be very careful not to confuse the symbols $\in$ and $\subseteq$.

It is convenient to define a *neighbourhood* of a point x in a topological space X to be any set which contains an open set containing x. (When X is a metric space, this coincides with our original definition.) It is then possible to define continuity of a function locally (that is, at a point) in terms of neighbourhoods, just as we did for metric spaces.

These definitions are somewhat frightening initially, and not just because all the geometry appears to have gone out of the window. The structures involved (topologies) can be absolutely enormous, and the whole apparatus appears unmanageable. Fortunately, the intuition developed by thinking with metric spaces is surprisingly helpful for understanding topological spaces, and after working through analogues of the basic theorems (and playing with some of the standard counterexamples) they begin to seem quite visualisable. As for the amount of structure being carried around — well, the metric on a metric space actually carries more information than the topology it defines (see below); it's just that its definition makes it seem "smaller".

It's clear that *any metric space can be regarded as a topological space* by simply looking at its collection of open sets, and forgetting the actual metric. (It is in this sense that topological spaces are "more general" than metric spaces: there is a topological space associated to each metric space, but there are many others besides.)

What is the correct notion of *equivalence* of metric and topological spaces?

For metric spaces the natural notion is *isometry*; existence of a bijective, distance-preserving map between the two spaces. But this is a very fine equivalence relation, much finer than we want for topological purposes. (The two different metrics on S^2 described above are *not* isometric!) But the natural definition for topological spaces, *homeomorphism*, turns out to be precisely what we want:

Definition 5.2.7. Two topological spaces are *homeomorphic* if there exists a pair of mutually-inverse continuous maps between them.

The term *isomorphic* would be just as good: an isomorphism between mathematical structures (topological spaces, groups, vector spaces, ...) can always be defined as a pair of mutually-inverse "structure-preserving" maps, where structure-preserving is interpreted appropriately: that is, as "homomorphism" in the case of groups, "linear map" in the case of vector spaces, and "continuous

map" in the case of topological spaces. (The language of *category theory* encapsulates this idea neatly.)

Example 5.2.8. (1). The open unit interval $(-1, 1)$ and the real line $\mathbb{R}$ are homeomorphic. Just use the map $x \mapsto x/(1 - x^2)$ and its inverse.

(2). Generalising this, we have that the open unit ball Int B^n and the space $\mathbb{R}^n$ are homeomorphic. The map $x \mapsto x/(1 - \|x\|^2)$ is perhaps the nicest choice of homeomorphism.

(3). The map $t \mapsto e^{2\pi it}$ is a continuous bijection between the interval $[0, 1)$ and the circle S^1. However, its inverse is *not* continuous, and therefore the circle is not homeomorphic to an interval (which is just as well, as topology would be boring if it were!)

Note that it makes sense (already implicitly used above) to talk about two metric spaces being homeomorphic: it means that their associated topological spaces are homeomorphic, or that there is a pair of mutually-inverse continuous maps going between them.

5.3. **Hausdorff and second countable.**

The final parts of the definition of a manifold that need explanation are the terms *Hausdorff* and *second countable*. (We're getting there, don't worry!)

Definition 5.3.1. A topological space is said to be *Hausdorff* if, for any pair of distinct points x, y, one can find disjoint open sets U, V containing x, y respectively.

This definition is part of a family of "separation axioms" dealing with whether points and/or open sets can always be "insulated" from one another by means of larger open sets. Hausdorffness (Hausdorffitude?) is the only one worth bothering with here (at all?), for the following simple reason. Any metric space is automatically Hausdorff: if x, y are distinct then $d(x, y) > 0$, and balls of radius $d(x, y)/3$ at x, y are disjoint, by the triangle inequality.

Second countability has to do with the analogue, in the world of topological spaces, of the concept of *generators* for a group (namely, a set of elements which when multiplied together in all possible ways, yields all the elements of the group). This notion is sometimes a convenient time-saving device in proofs.

Take any collection ρ of subsets of a set X whose union is all of X. It won't in general be a topology, but it is easy to construct a "smallest" topology (one with the fewest open sets) containing all those subsets, as follows. If we close ρ under finite intersections (by adjoining all sets which are intersections of finitely-many elements of ρ), we obtain a larger collection σ which obviously satisfies the third axiom for a topological space, and also contains the empty set. If we now close σ under arbitrary unions (by — what else? — adjoining all unions of elements of σ) we get a collection τ which satisfies the second axiom, contains X as well as $\emptyset$, and (check) still satisfies the third axiom; it is a genuine topology.

Any collection of open sets such as σ which, when closed under unions, generates τ, is called a *base* for τ; any collection such as ρ, which requires closure under both finite intersections and arbitrary unions to generate τ, is called a *sub-base*. In a metric space, for example, the collection of all open balls $B_\epsilon(x)$ is a base for the topology. It's easy to see that just the balls of radius $1/n$ (for positive integer n) will do. In $\mathbb{R}^n$, we may actually use balls of radius $1/n$ based at rational points (that is, points whose coordinates are all rational).

Definition 5.3.2. A topological space with a countable base is called *second countable*.

5.4. **Reconsidering the definition of n-manifold.**

The most important part of the definition is the idea of a space being *locally homeomorphic to* $\mathbb{R}^n$. This means that at any point $x \in M$, there exists a neighbourhood (we might as well assume it's actually an open set containing x) which is homeomorphic to $\mathbb{R}^n$, or equivalently (since this is also homeomorphic to $\mathbb{R}^n$) to the open unit ball B^n.

(It seems easier to visualise the neighbourhoods as being homeomorphic to n-balls. But if we use $\mathbb{R}^n$ as the target, we are asserting that there is at each point a way to choose a local coordinate systems $\{(x_1, x_2, \ldots, x_n)\}$, parametrising continuously and in a one-to-one fashion the points near x by arbitrary n-tuples of real numbers.)

Now any subspace of $\mathbb{R}^N$ is both Hausdorff and second countable. So any subspace which is locally homeomorphic to $\mathbb{R}^n$ ($N \geq n$) is an n-manifold. Conversely, it turns out that any n-manifold can be embedded in (mapped homeomorphically to a subspace of) $\mathbb{R}^N$, for some suitably large N. (Hausdorff and second countable are essential for this to work.)

So we could make an alternative definition: *an n-manifold is a subspace of some $\mathbb{R}^N$ which is locally homeomorphic to an open n-ball.* This is an equivalent definition to the earlier one, but it is a little misleading: thinking of manifolds as subspaces may help in visualising them, but one must remember that the way the manifold is embedded in $\mathbb{R}^N$ is *not* an intrinsic part of its definition; the equivalence relation we want to consider is *homeomorphism*, which does not address the ambient space.

Exercise 5.4.1. Determine which of the following spaces is a manifold:

$$S^1, \quad S^1 \amalg S^1, \quad \mathbb{R}, \quad I, \quad (0,1), \quad (0,1], \quad \text{the "open letter } Y\text{"},$$
$$S^2, \quad S^1 \times S^1, \quad \text{the open unit disc}, \quad S^1 \times S^1 - (\text{point}),$$
$$S^1 \times S^1 - (\text{open disc}), \quad \text{an open subset of } \mathbb{R}^2.$$

Are any of these spaces homeomorphic to one another?

Since we want to consider knots that are the boundaries of surfaces, we need to extend the definition as follows.

Definition 5.4.2. An *n-manifold-with-boundary M* is defined in the same way as a manifold, except that its points are allowed to have neighbourhoods homeomorphic *either* to $\mathbb{R}^n$ or to the upper half space $\mathbb{R}^n_{\geq 0} = \{(x_1, x_2, \ldots, x_n) : x_n \geq 0\}$.

Remark 5.4.3. Note that the second type of neighbourhood can be thought of as half of the open unit ball (the part with $x_n \geq 0$).

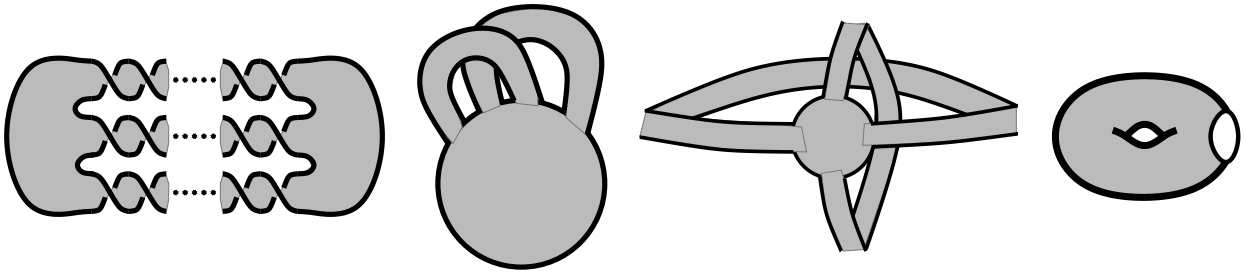
Exercise 5.4.4. Repeat exercise 5.4.1, identifying the manifolds-with-boundary.

Definition 5.4.5. The set of points which have no neighbourhood homeomorphic to $\mathbb{R}^n$ is called the *boundary* ∂M of M, and its complement $M - \partial M$ is called the *interior* of M. (Warning: these uses are different from the concepts of *boundary* (or *frontier*, and meaning closure minus interior) and *interior* of a subset of a topological space. In our case the manifold-with-boundary *is* the whole space, so its frontier is empty and its topological interior is itself!)

Exercise 5.4.6. Show that the boundary of an n-manifold-with-boundary is itself an $(n-1)$-manifold without boundary: "the boundary of a boundary is zero".

Remark 5.4.7. Note that a manifold is a special type of a manifold-with-boundary — it's just one where the boundary is the empty set! The converse is not true, however. The unit interval is a manifold-with-boundary, but not a manifold.

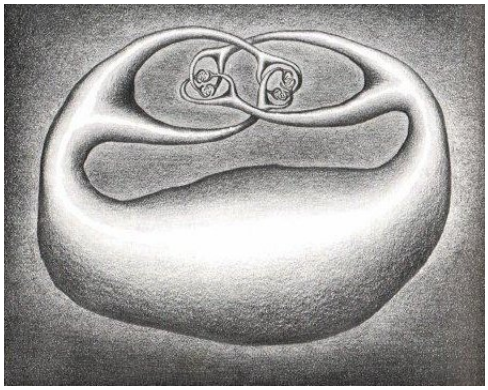
Exercise 5.4.8. These four surfaces are homeomorphic!



6. Classification of surfaces

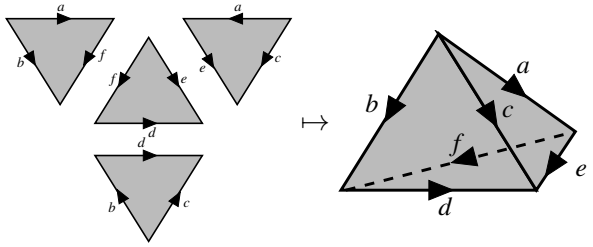
6.1. **Combinatorial models for surfaces.**

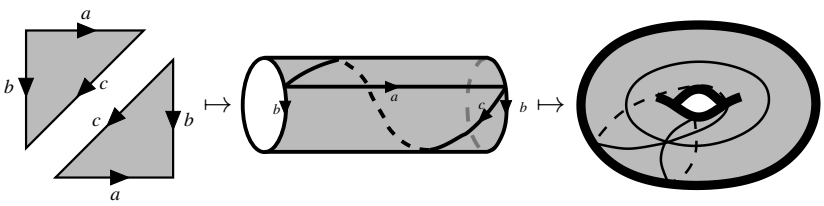
Just as when working with knots, we will find it helpful to work with a *combinatorial* (meaning, roughly, *discrete and finite*) version of the concept. The reasons are much the same: firstly, we'd like to be able to do combinatorial proofs, including using induction on the "number of constituent elements". Secondly, when we come to think about surfaces inside $\mathbb{R}^3$, it's necessary to have some way to rule out *wild* surfaces such as the (frankly awesome!) *Alexander Horned Sphere* shown below.



The appropriate concept is to build surfaces by gluing triangles together, edge-to-edge in pairs. Consider a collection of (closed, solid) Euclidean triangles $T_1, T_2, \ldots, T_f$ of arbitrary shapes and sizes.

Definition 6.1.1. A *gluing pattern* P on $\amalg T_i$ is an *oriented pairing* of the edges of the triangles. That is, we match each edge to another edge of the same length, in a specified direction. (Clearly the triangles will need to have even numbers of edges of each length appearing, and f itself will have to be even.) We usually indicate the pairing by labelling the edges by symbols, each appearing twice, and orienting each edge using an arrow, as shown below.

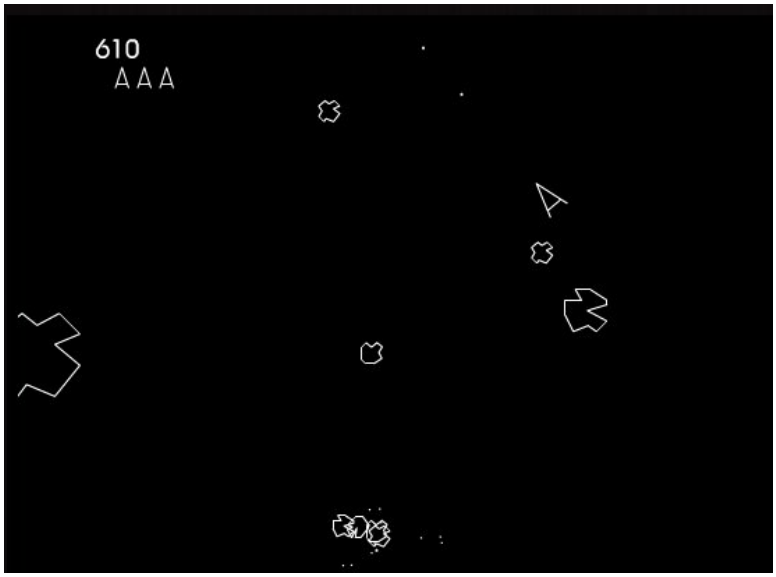


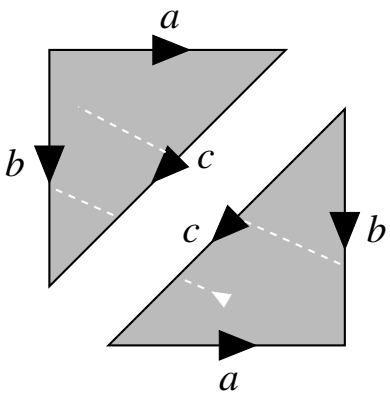


As the picture suggests, the gluing pattern generates an equivalence relation on $\amalg T_i$. Each point of an edge is identified with the corresponding point on the other like-labelled edge, using the unique length-preserving correspondence between the edges determined by the directions of their arrows. The resulting equivalence classes are by definition the *points* of the *closed combinatorial surface* Σ associated to P.

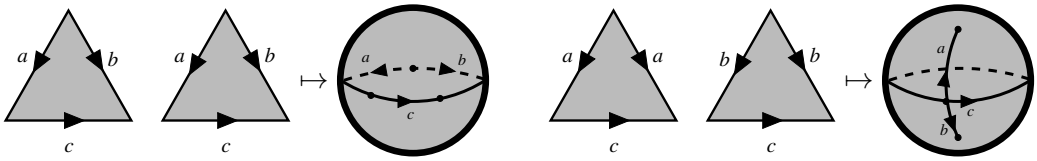
The surjective map $q : \amalg T_i \to \Sigma$, sending points to their equivalence classes, is injective on the interiors of triangles in T, two-to-one on points in the interiors of their edges, and at least two-to-one on vertices. We can define the *faces, edges* and *vertices* of Σ as the images under q of the original faces, edges and vertices of $\amalg T_i$. Clearly Σ has f faces, $3f/2$ edges, and a number of vertices somewhere between 1 and $3f/2$; as the above examples show, this will actually depend on the gluing pattern.

In the pictures above, the glued-up surfaces are drawn using a 3-dimensional perspective, but this is not a good way to think of them. In general we probably can't assemble the triangles sensibly inside $\mathbb{R}^3$ at all, but in any case we want to study them *intrinsically* as 2-dimensional objects and not in this artifical way. I think the best way to think about them is as universes in a game of "Multi-screen Asteroids". Imagine flying around in the triangular "screens"; if you fly off (or shoot off) the edge of one screen, you come out on another one. This internal navigation is the only way you have to explore your world.

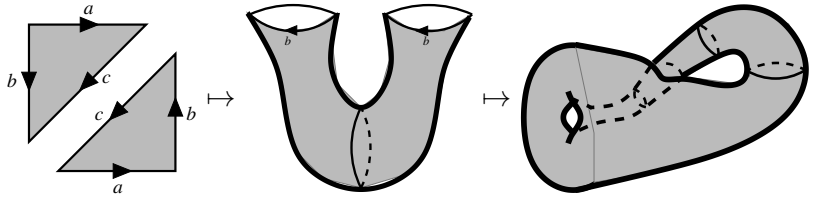




Here are some further examples: two different ways to glue two triangles so as to make a sphere

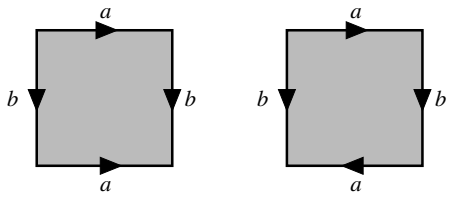


and two triangles glued to make a *Klein bottle*. The third picture here depicts the Klein bottle inside $\mathbb{R}^3$ but *intersecting itself*: the idea is that the tube curls around, passes inside itself, and connected to the other end. The Klein bottle can sit in $\mathbb{R}^4$ without self-intersections, but not in $\mathbb{R}^3$!



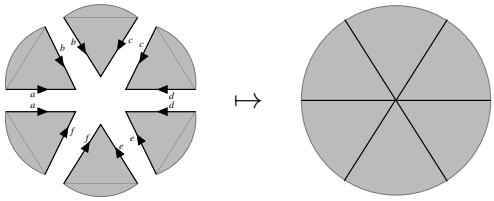
To cut down on the complexity of the gluing codes, we will often assemble some of the triangles in the plane, if this is possible. For example, here are simplified pictures for the torus and Klein

bottle.

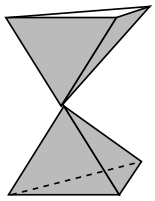


(I call this a "net", like the planar cut-out shapes you begin with when constructing cardboard models of polyhedra.)

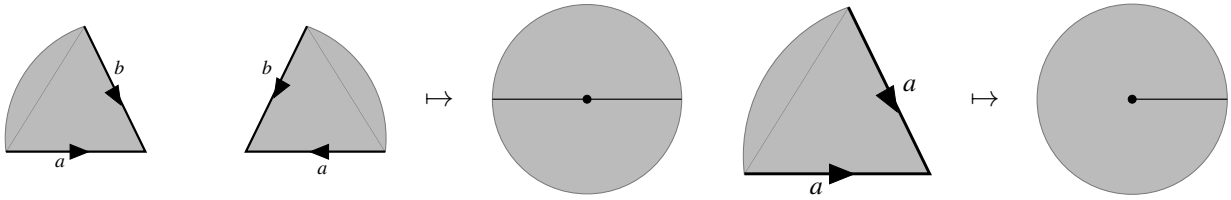
The local structure of Σ is very nice. Every edge of Σ has precisely two faces incident to it. Now consider the part of Σ formed from points of $\amalg T_i$ within some small distance ϵ of the vertices (take ϵ to be one third of the shortest length of an edge, for example). It consists of a union of $3f$ sectors of discs or radius ϵ; each sector has two straight edges, and these are glued in pairs. Therefore the sectors become glued into a bunch of discs of radius ϵ. There is one such disc for each vertex of Σ (we think of it as a neighbourhood of the vertex), and it consists of one or more sectors glued around in a circular fashion, as in the following picture. (Because the gluing is done along edges, we can never have "just the vertices touching" as in the second picture.)



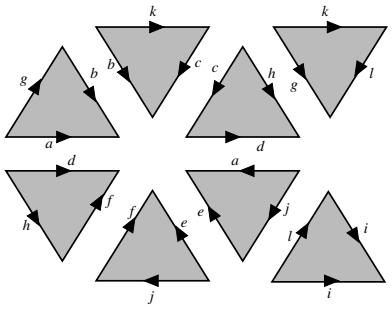
Doesn't occur:



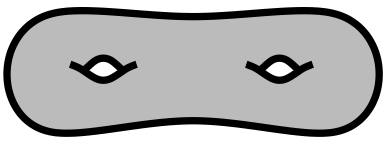
Warning: what we see here are only *corners* of triangles, not the whole triangles. The six sectors in the picture above may not all belong to different triangles – in the earlier torus example, all six corners of the two triangles come together at the single vertex! Moreover, the picture may not really look like a polygon, in that it might have only 1 or 2 sides!



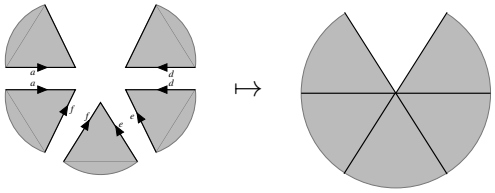
Exercise 6.1.2. Count the vertices of the surface coming from this gluing pattern.



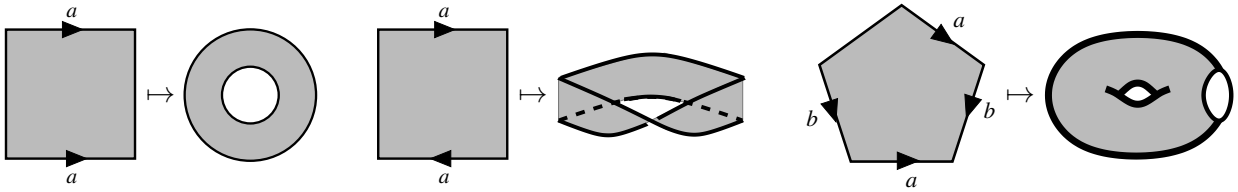
Exercise 6.1.3. Find a way to glue up the edges of a regular octagon in pairs so as to make a genus two surface like this one.



An extremely useful variation on the above to allow some edges to remain *unpaired*: redefine a gluing pattern as a pairing up of *some* of the edges of $\amalg T_i$ (there can now be an odd number f of triangles). After gluing, the unpaired (unlabelled) edges will remain as "free edges" of the surface Σ, with only one face incident to them, instead of two. A space Σ constructed from such a generalised gluing pattern is called a *combinatorial surface with boundary*. For such a surface, another kind of vertex neighbourhood is possible, formed by gluing a *chain* of sectors, beginning and ending with sectors possessing free edges, rather than a complete loop. Therefore the free edges of a surface with boundary must be glued end-to-end, forming a union of closed polygons, the *boundary* $\partial\Sigma$ of Σ.



Basic examples include the *annulus* (*cylinder*), the *Möbius strip*, and a torus with a hole.



Remark 6.1.4. (1) With the language of topological spaces, it is easy to describe the topology on Σ: it is the *quotient* or *identification space* topology induced by the equivalence relation. This just means that continuous functions on Σ are precisely the same thing as bunches of continuous functions on the individual triangles which agree at correspondig points of paired edges. The observations about the local nature of the surface above (the way corners of faces come together at vertices) show that indeed a closed surface is locally homeomorphic to $\mathbb{R}^2$, as required to be a 2-dimensional manifold. (Moreover, a *2-dimensional manifold with boundary* is a space which is locally homeomorphic to either $\mathbb{R}^2$ or the upper half-space, and this is what we get when we construct combinatorial surfaces with free edges.)

(2) We could try to describe the surface as a metric space instead. Define a path on the surface to be a sequence of continuous paths across triangles of the kind along which the spaceship in "Asteroids" might fly. Then define distance between points on Σ by taking the shortest length of such a path. The main defect of this definition is that it will only work for a connected surface, where all points actually can be joined by such paths. There are ways to rectify this defect, but they are not very natural. This is one good argument for preferring the language of topological spaces over that of metric spaces.

Finally, the following fact justifies the definitions we have just made: it allows us to consider only combinatorial surfaces, rather than having to work with arbitrary 2-manifolds.

Fact 6.1.5. Any compact 2-manifold (perhaps with boundary) is homeomorphic to a combinatorial surface (with boundary if appropriate).

Remark 6.1.6. (1). The same sort of technique can be used to talk about 1-dimensional manifolds: we pair the vertices of a disjoint union of closed unit intervals. Classification of 1-manifolds is an easy exercise: closed ones are just disjoint unions of polygonal circles.

(2). The obvious generalisation is to build n-dimensional objects by gluing together *n-simplexes* (the n-dimensional analogues of tetrahedra) in pairs along their faces. But one must be careful!

When $n \geq 3$, the result of such a gluing might *not* be an n-manifold. For example, in three dimensions, the boundary of the ϵ-neighbourhood of a vertex of the glued-up space could be any closed combinatorial surface, and only if all such boundaries are 2-spheres will the space be a 3-manifold. So the definition of an n-dimensional combinatorial manifold requires that the boundary of each vertex neighbourhood is a combinatorial $(n-1)$-manifold which is equivalent (see next subsection) to an $(n-1)$-sphere.

But even though this gives us a satisfactory world of combinatorial manifolds, we still need to be careful because fact 6.1.5 also starts to fail in higher dimensions! Mike Freedman proved in 1982 (at UCSD) that there exist 4-dimensional topological manifolds which are not homeomorphic to a combinatorial 4-manifold (this is part of the work for which he won a Fields Medal). Once we reach dimension 7, we find a reverse sort of problem: there exist topological manifolds which are homeomorphic to several different combinatorial manifolds which are combinatorially inequivalent (see below) from one another!

6.2. **Equivalence of surfaces.**

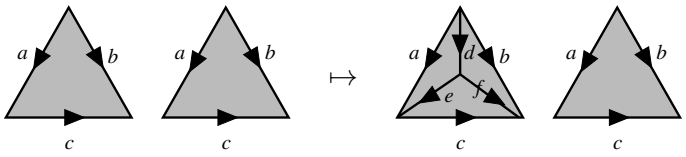
We need to introduce a suitable notion of equivalence for combinatorial surfaces which corresponds to the notion of homeomorphism in the world of topological 2-manifolds. In working with polygonal knots we used Δ-moves; what is a sensible analogue for surfaces?

Definition 6.2.1. We define two surfaces Σ_1, Σ_2 coming from gluing patterns P_1 and P_2 to be *isomorphic* if there is a bijection between the sets of vertices of the triangles of P_1 and P_2, taking triangles to triangles, and preserving identifications.

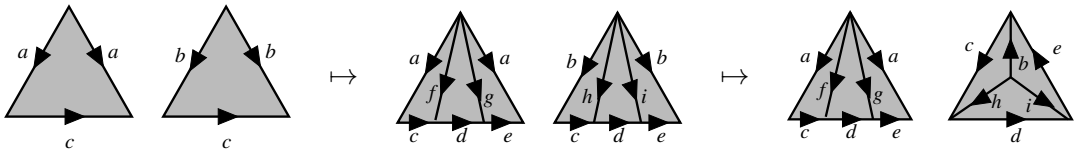
Example 6.2.2. (1) Any gluing pattern is isomorphic to one all of whose triangles are equilateral.

(2) The two different gluing patterns building the sphere out of two triangles are *not* isomorphic, even though they build the same kind of surface. Since isomorphism can't change the number of triangles making up the surface either, we see that it is inadequate as a notion of equivalence. So:

Definition 6.2.3. Define a *subdivision* of a gluing pattern P_1 to be another one P_2 obtained by decomposing each of the triangles of P_1 as a net of smaller triangles, such that the gluing of points on the boundaries is preserved. (The decomposition of the net into triangles still requires that triangles meet along full edges.) For example:



Now we can define *equivalence* as the equivalence relation generated by isomorphism and subdivision: two surfaces are equivalent if there is a finite sequence of isomorphisms and subdivisions (or reverses of subdivisions) relating them. The pictures above and below show that both of our simple gluing patterns for the sphere, as well as the tetrahedral pattern, are equivalent. We'll refer to any surface equivalent to our standard model of the sphere as "a sphere", anything equivalent to our standard torus as "a torus", and so on.

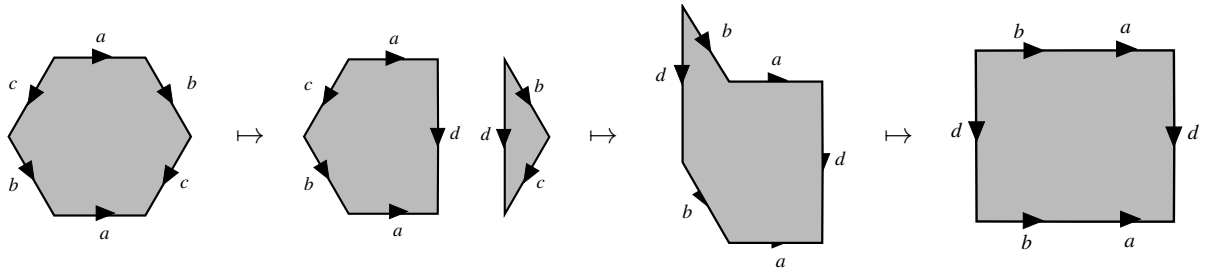


Ultimately the justification of this notion of equivalence is the following difficult theorem:

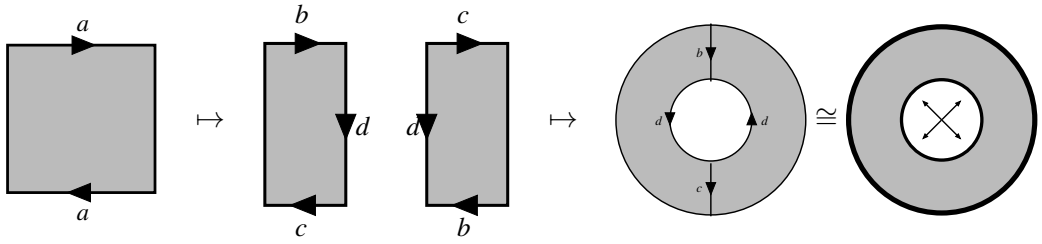
Fact 6.2.4. Two surfaces Σ_1 and Σ_2 produced from gluing patterns P_1 and P_2 are homeomorphic if and only if the patterns are equivalent in the sense defined above.

Here are some examples of manipulating gluing patterns to understand surfaces in different ways.

Example 6.2.5. A hexagon glued up as follows gives a torus. By splitting the pattern, reattaching one piece and then deforming it, we get an isomorphism to a standard "square torus" pattern.

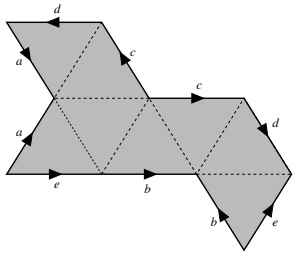


Example 6.2.6. A similar trick gives an interestingly different description of a Möbius strip. Splitting the gluing pattern down the middle into two rectangles, then stretching them around to form an annulus looks like this:



We're not really allowed to use curved edges, but if you imagine making a very fine subdivision you can always approximate such things, and it saves time to be able to draw this sort of picture. The right-hand picture is a *crosscap*: an annulus whose inner boundary circle is glued to itself in a 2:1 fashion by identifying antipodal points. Its outer boundary circle corresponds to the actual boundary circle of the Möbius strip. I often draw such a thing with a sort of "X" (cross) symbol as on the right, indicating the antipodal gluing.

Exercise 6.2.7. Prove that this net gives a surface equivalent to a regular tetrahedron.



Exercise 6.2.8. Let P be a convex polygon in the plane, and v a point inside P. The *cone* P_v is the surface got by dividing P into triangles, using arc-segments drawn from v to its vertices. (*Convex* means that whenever two points lie inside P, so does the arc-segment joining them.)

(a) Show that this surface is equivalent to the one consisting of a single equilateral triangle.

(b) Show that the intersection of two convex polygons is convex.

(c) Show that if w is another point inside P, not lying on a line through v and a vertex of P, that the surfaces P_w and P_v have a common subdivision.

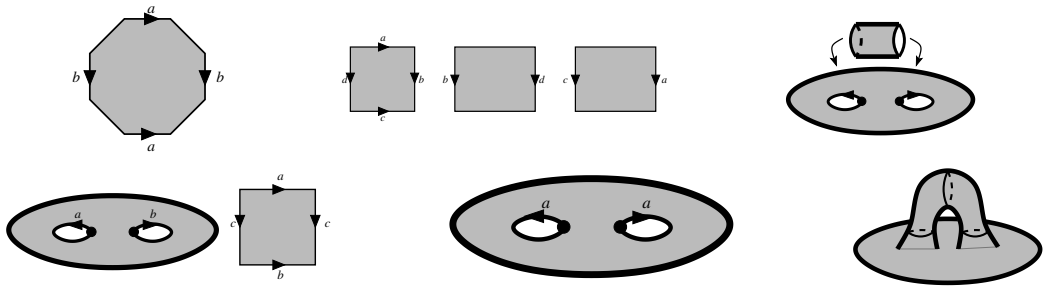
(d) Show by a similar argument that any two triangulations (subdivisions coming from a gluing pattern) of P into triangles are equivalent. (This justifies allowing ourselves to build surfaces by gluing convex polyhedra (not specifying how they are divided into triangles), and not just triangles.)

Remark 6.2.9. It will follow from the Jordan Curve Theorem, to be proved soon, that the inside of *any* closed polygonal curve in the plane can be divided into triangles, without introducing any more vertices, resulting in a gluing pattern which is equivalent to a disc (the surface coming from a single triangle without gluing). Moreover, any two triangulations of it are equivalent. Therefore we can forget about the precise decomposition, and work with larger cells, namely pieces of the plane bounded by polygons. We can simply attach gluing labels to the edges of these polygons and glue them to other planar bits. These large pieces simplify the descriptions of surfaces quite a lot. (We replace lots of small triangles by larger units).

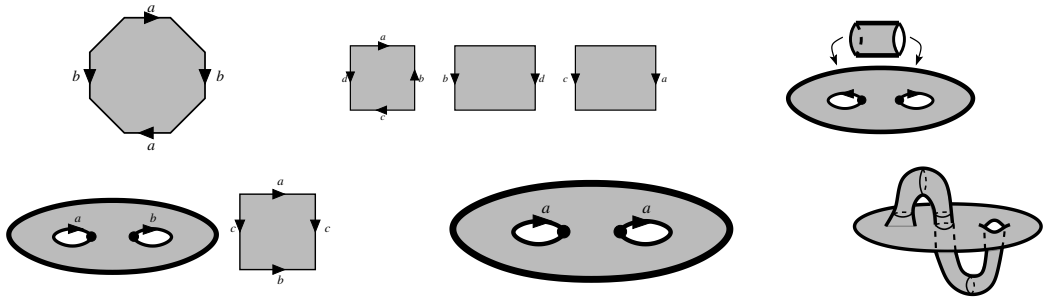
Exercise 6.2.10. We defined two surfaces to be topologically equivalent if they were related by a sequence of isomorphisms, subdivisions or inverse of subdivisions. Show that if instead we define $F_0 \sim F_1$ if and only if there exist subdivisions F_0', F_1' which are isomorphic, then $\sim$ is an equivalence relation, and is in fact the same relation as the first definition of topological equivalence.

Exercise 6.2.11. Classify all possible gluings of a square.

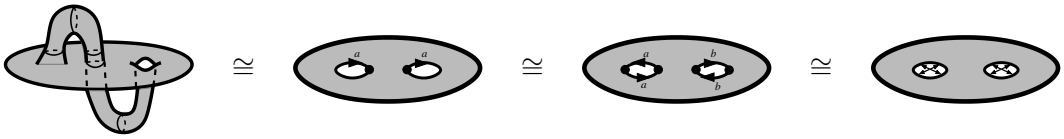
Exercise 6.2.12. Here are four surfaces described by gluings. Show using cutting and pasting arguments that they are all equivalent (this surface is really a torus minus a hole).



Exercise 6.2.13. Do the same for these four. (This surface is a Klein bottle minus a hole).

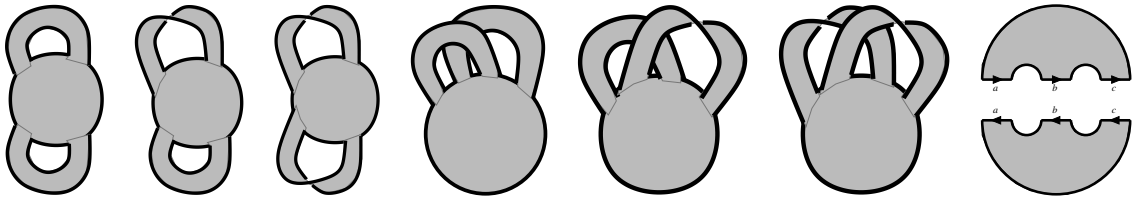


Exercise 6.2.14. Show that a twisted handle is equivalent to a disc with two crosscaps:



Exercise 6.2.15. Show that a disc with a handle and a crosscap attached is equivalent to a disc with three crosscaps:

Exercise 6.2.16. Identify the following surfaces in terms of more familiar things (sphere, torus, Möbius strip with a hole, and so on). (The pictures show things that are flat in the plane for the most part.)



6.3. **Basic properties of surfaces.**

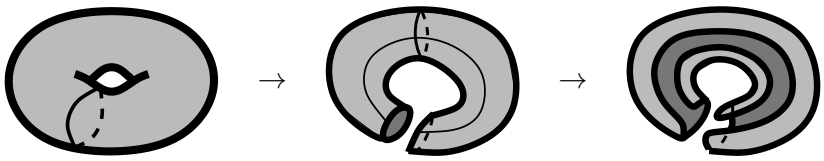
Definition 6.3.1. A surface is *connected* if it's possible to move from any face to any other face by stepping from face to face across shared edges. If it's not connected we'll refer to the equivalence classes (under this relation of faces being being joinable) as *components* of the surface.

It's important to note that the property of connectedness (and the number of components of a disconnected surface) is invariant under subdivision, and hence a well-defined number independent of gluing pattern. (The same goes for the number of boundary circles of surface.)

Definition 6.3.2. (1) A *simple closed curve* C in a surface Σ is a sequence of vertices and edges $v_0, e_1, v_1, e_2, v_2, \ldots, e_n, v_n = v_0$, such that each edge joins the vertices before and after it in the sequence, all the edges are distinct, and all the vertices are distinct (with the exception of $v_n = v_0$). The term "simple" refers to the fact that the curve does not intersect itself or repeat its tracks, and "closed" that it forms a complete loop, but I will just call the thing a *curve*. (We need to specify the sequence of vertices *and* the edges because multiple edges may connect a given vertex pair.) We will normally only consider curves whose edges are not boundary edges of the surface, so that we can sensibly cut along them (see below).

(2) An *arc* in a surface is just like a curve except that it is not closed: it starts and ends at distinct points. Usually we are interested in *boundary arcs* which run along the boundary edges of a surface, or *proper arcs* which start and end at points on the boundary but whose edges are internal edges of the surface.

Definition 6.3.3. If Σ is a surface and C a curve, then we can *cut along* C by simply removing (from the gluing pattern that builds Σ) the identification instructions on all edges that map to C. We get a new surface Σ' formed by identifying the same set of triangles used to build Σ, but without gluing across any of the edges of C. (This is *not* the same as the complement $\Sigma - C$ of C in Σ.) We can also cut surfaces along proper arcs, but this is less important to us.) For example, we can cut a torus along a curve then an arc.

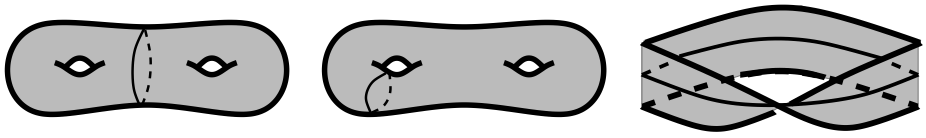


Definition 6.3.4. When we cut along a curve:

(1) The new surface Σ' will have either 2 or 1 more boundary components than Σ. The curve is called 2-*sided* or 1-*sided* respectively.

(2) The new surface Σ' will have 1 or 0 more components than Σ. We say C is a *separating* or *non-separating* curve respectively.

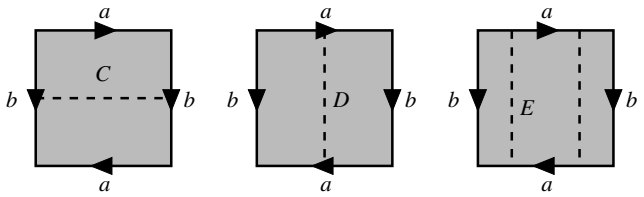
Here are examples of a 2-sided separating curve, a 2-sided non-separating curve and a 1-sided curve. (When you cut along the centreline of the Möbius strip, you get an annulus twice as long as the original strip, with two boundary components, compared with the Möbius strip's one.)



Exercise 6.3.5. Prove the assertions above, that cutting indeed creates 1 or 2 new boundary circles, and 0 or 1 new components.

Exercise 6.3.6. Show that cutting along a 1-sided curve cannot separate a surface: thus 1-sided curves are always non-separating.

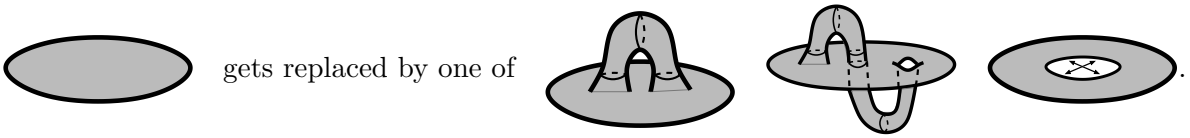
Example 6.3.7. Here are three curves on a Klein bottle. Cutting along them gives quite different results! The first is a non-separating 2-sided curve C which cuts the bottle into a cylinder. The second is a 1-sided (hence non-separating) curve D which cuts it into a Möbius strip. The third is a separating 2-sided curve E which cuts the bottle into *two* Möbius strips!



Exercise 6.3.8. Let Σ_1, Σ_2 be surfaces which are equivalent to the disc, and let A_1, A_2 be oriented arcs in their respective boundary circles, consisting of the same number of edges. After an isomorphism to resize the edges if necessary, we can glue the two surfaces together along these arcs. Show that the surface obtained is itself equivalent to a disc.

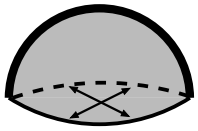
Exercise 6.3.9. Let Σ_1, Σ_2 be surfaces which are equivalent to the disc, and let their boundary curves be C_1, C_2 respectively, each with a chosen orientation and base vertex. After perhaps subdividing so that the curves have the same number of edges, and using an isomorphism to resize the edges as appropriate, we can glue the two surfaces along the curves (matching the base vertices). Show that the result is equivalent to a sphere.

Exercise 6.3.10. Consider the surfaces made by removing the interior of a closed disc from a torus or Klein bottle: they have one boundary circle, as does the crosscap. If D is a closed disc contained in a surface F then one may remove the interior of D, creating a boundary component, to which one can glue the boundary of any of the three surfaces just mentioned. These operations are called adding a *handle, twisted handle or crosscap* to F respectively:



The resulting surface is unique up to equivalence, i.e. it does not matter where the disc lies in the surface. Why?

Example 6.3.11. The surface obtained by gluing antipodal points of a disc's boundary is called the *projective plane*. If we draw the disc as a hemisphere, it looks like the hemisphere with points on the equator glued to their antipodes.



It can also be obtained by gluing the sides of a square according to the gluing pattern a, b, a, b. If we glue a disc onto the boundary circle of a crosscap we clearly get a disc with antipodal points of its boundary glued, so this shows that the projective plane can be thought of as the union of a Möbius strip and a disc.

Why is it called the *projective plane*? The answer is that it occurs naturally in the theory of 2-dimensional perspective drawings of 3-spaces, that is, projections of a 3d scene onto a 2d canvas by a (point) observer such as a painter. Digression to be written at some point...

6.4. **The Euler characteristic, graphs and trees in surfaces.**

You are probably familiar with the fact that for the five Platonic solids, the numbers of vertices, edges and faces satisfy *Euler's formula* $v - e + f = 2$. This formula is still true for irregular polyhedra, as long as they are convex: the number 2 reflects only the topology of the figure, in fact that its boundary is homeomorphic to S^2. We will extend this result during the course of the classification of surfaces.

Definition 6.4.1. For any combinatorial object A (something made of faces, edges and vertices,) the *Euler characteristic* of A is $\chi(A) = v - e + f$.

We will be concerned mainly with Euler characteristics of combinatorial surfaces and of combinatorial subsets of them. Here are some examples to illustrate how χ behaves.

Exercise 6.4.2. If $X = A \cup B$ is a combinatorial decomposition of a combinatorial object, then $\chi(X) = \chi(A) + \chi(B) - \chi(A \cap B)$.

Exercise 6.4.3. The Euler characteristic of any combinatorial circle is 0.

It will be useful in what follows to be able to use a bit of basic graph-theoretic terminology.

Definition 6.4.4. A *graph* G is defined by a set of vertices V, a set of edges E, and a choice, for each edge of E, of a pair of endpoints (elements of V, not necessarily distinct).

This kind of graph is therefore allowed to have isolated vertices with no incident edges; multiple edges between any given pair of vertices; and edges which loop from a vertex back to itself. (Sometimes people work with stricter definitions forbidding some of these occurrences; it's usually worth reading the definition carefully!)

Most of our graphs will arise from taking a collection of edges and gluing clumps of their endpoints together in some fashion: the set of vertices and edges of any surface (the way we defined it) is of this form, and can therefore not have isolated vertices. (In the world of topological spaces, such a graph is naturally viewed as a quotient of a disjoint union of closed unit intervals.)

Definition 6.4.5. (1) The *degree* or *valence* of a vertex is the number of incident *ends* of edges. (Thus, the "loop" graph with only one edge and one vertex has vertex of degree 2.)

(2) Any graph has an *Euler characteristic* $\chi(G) = V - E$ in the obvious way.

(3) A *path* in a graph is a sequence of vertices and edges joining them, in the obvious way. (We don't normally require these to be non-self-intersecting or non-backtracking.)

(4) A *cycle* is a simple closed curve, in the previous language: it *is* required to be non-self-intersecting and non-backtracking.

(4) A graph is *connected* if every pair of vertices can be joined by a path. (This coincides with the topological notion of connectedness.)

(5) A *tree* is a connected graph with no cycles. (A *forest* is a not necessarily connected graph with no cycles!)

Lemma 6.4.6. *Any connected graph G has $\chi(G) \leq 1$, with equality if and only if G is a tree.*

Proof. If G contains a vertex with degree 1 it may be *pruned* by removing that vertex and its incident edge (but not the vertex at the other end). The result is still connected (easy exercise), and has the *same* Euler characteristic. So apply pruning to G until one of two things happens: either all remaining vertices have degree 2 or more, or there is just one vertex left and no edges.

In the first case, the fact that the sum of degrees of all vertices equals twice the number of edges (counting up the number of *ends* of edges in two different ways) shows that $2e \geq 2v$ and hence that the Euler characteristic of the original graph was $\chi(G) = v - e \leq 0$. In the second case, since the single-edge graph has Euler characteristic 1, so too did the original G. Rebuilding G by reversing the pruning sequence (sprouting?) one can easily check that there can be no cycles (the point being that a univalent vertex could never be part of a non-backtracking cycle, and hence adding these does not alter the acyclicity of what we have already). □

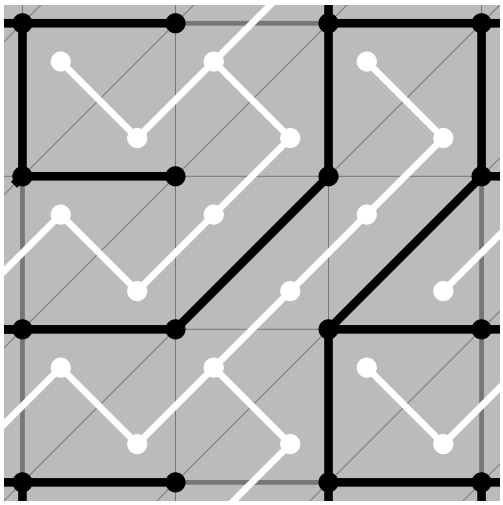
Lemma 6.4.7. *Every connected graph contains a* maximal tree*: a tree containing all its vertices.*

Proof. If the graph has a cycle, we can remove any one of the edges (choose arbitrarily) of that cycle without disconnecting it (because we can still move between the newly disconnected vertices by going the other way around the cycle.) Keep doing this until there are no cycles left. □

Exercise 6.4.8. Write down more careful, formal definitions and proofs of everything above!

Now let Σ be a connected surface. Associated to Σ there are *two* natural graphs. One is the obvious graph G formed by the vertices and edges. The other is the *dual graph* G' coming from the faces and edges: it is defined to have a vertex for every face of the surface, and an edge whenever two faces are adjacent. We could actually draw G' on the surface by placing vertices at the centres of the faces, and drawing edges between these. Each edge of G' would then cross transversely one of the edge of G.

A very useful construction is to pick a maximal tree T' inside this dual graph G'. Having done that, let H be the subgraph of G obtained by removing the edges which hit those of T. I claim that H is connected (proof below). Here is a picture illustrating this:



This is a model of the torus built from 18 triangles, having 9 vertices and 27 edges. It is drawn slightly enlarged at the edges to make the wrapping-around clearer: the four vertices along the top edge (for example) are the same as the four vertices along the bottom edge.

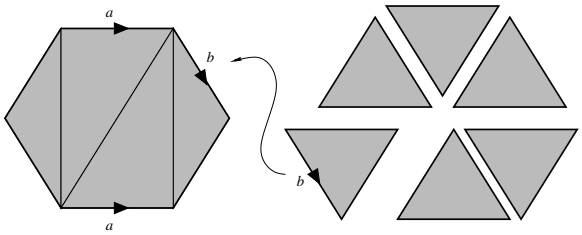
The white graph is a choice of maximal tree T': it has 18 vertices and 17 edges (count them!) The black graph is the graph H, consisting of all 9 of the vertices of the surface and the 10 edges of the surface which are not being crossed transversely by white tree edges. Notice that $10 + 17 = 27$; every edge of the original surface is associated with either T' or H.

Here is a different way to picture all this, which helps to explain why H is connected.

Lemma 6.4.9. *Σ may be obtained by gluing in pairs the sides of a regular $f+2$-gon in the plane.*

Proof. Imagine the f disjoint triangles $\amalg T_i$ out of which Σ is built all lying in a box on the floor. Since the surface is closed, their edges are labelled in pairs indicating how to assemble them to make Σ. We will "partially assemble" the surface as much as possible on the tabletop.

Pick up one triangle to begin with. It's possible that two of its edges are glued together; in that case, look at the third edge. Otherwise, pick an edge arbitrarily. The key idea is that there is a unique new triangle in the box which fits onto this edge. Pick this one up from the box, attach it to the first triangle on the table (resizing it if necessary), and deform the result to a square. Now we repeat. At each stage, look at the boundary of the regular polygon you have in your hand: its edges are all labelled, and some may in fact be paired with each other. But if there is a "free edge", one not paired with another edge of the polygon, then it must be paired with an edge of one of the triangles still in the box: pick this up, attach it along the edge you were considering, and deform the result to a regular polygon. As long as there are free edges remaining, there *must* be triangles still on the floor, and the process continues. Therefore it finishes when there are no free edges of the polygon left, and at this stage there cannot be triangles remaining in the box either, or we could start again and end up with a completely separate component of Σ, which was assumed to be connected. So all of them have been used, and the polygon (which gains a side for each triangle added after the first one) has $f+2$ sides. Notice that the polygon has no interior vertices – they are all on the boundary.



□

What is really happening here is a construction of the maximal tree T' in G': its "white edges" are precisely the duals of the internal edges of the polygon (the process adds one new white vertex and one new white edge at each stage, making it clear that it forms a tree). Therefore H, which consists of the remaining edges of the surface, is what you get by taking the boundary of the polygon and gluing up its edges in pairs appropriately. Since the boundary is connected, so is H. (In the reverse direction, if we started with the earlier picture of the graphs H and T' on the surface, we could obtain the polygon picture by cutting along all the black edges and simply opening out and reshaping the resulting surface until it is a regular polygon.)

This discussion leads to a vital ingredient in the classification theorem:

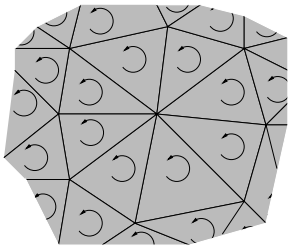
Lemma 6.4.10. *If Σ is a closed connected surface then $\chi(\Sigma) \leq 2$.*

Proof. By the above construction we see that every vertex of Σ belongs to H, every face corresponds to a vertex of T', and every edge of Σ corresponds either to an edge of T' or to an edge of H. Therefore the alternating sum gives $\chi(\Sigma) = \chi(T') + \chi(H)$. Since T' is a tree, $\chi(T') = 1$, and since H is a connected graph, $\chi(H) \leq 1$. □

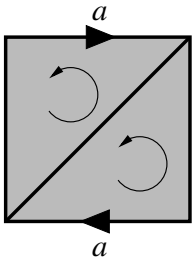
6.5. **Orientability.**

There are lots of equivalent definitions of orientability, and which one to use as *the* definition is a matter of taste.

Definition 6.5.1. An *orientation* of a closed combinatorial surface F is an assignment of a clockwise or anticlockwise "circulation" to each face (really an ordering of its vertices, considered up to cyclic permutation), such that at any edge, the circulations coming from the two incident faces are in opposition (if they were cogwheels, they would clash).



Not all surfaces have orientations. The torus and sphere, built from two triangles, clearly can be, but the Möbius strip below cannot: if we make the circulations oppose one another correctly at the diagonal edge, then they now agree rather than oppose along the edge a.



The concept would be useless if it depended on the gluing pattern, but fortunately it doesn't. If a surface has an orientation, then any subdivision of it inherits an orientation (make the small triangles circulate the same way their larger parent triangle does). Conversely, if the subdivision is oriented, then all small triangles in each large one circulate consistenly, so the large one inherits an orientation. Therefore the *orientability* or *non-orientability* of a surface is a property of its equivalence class, independent of the particular gluing pattern used. (Note: if a surface is orientable then it possesses 2^m different orientations, where m is its number of connected components.)

If we remove a collection of faces from a surface Σ and delete the gluing instructions from remaining faces which used to be glued to deleted faces, we get a new gluing pattern which we can think of as defining a *subsurface* of Σ. This allows us to state a different characterisation of orientability:

Lemma 6.5.2. *A surface is orientable iff it contains no subsurface equivalent to a Möbius strip.*

Proof. It's clear that any surface which possesses a Möbius strip subsurface cannot be oriented, because even by itself the strip cannot be oriented. Conversely, construct a maximal tree T' as in the last section. Choosing an orientation of one face, propagate it along edges of the tree to all the other faces: there is no choice about how to do this, since each face is reached by a unique chain of adjacent faces. Now look at each of the remaining "black edges" of the surface, the edges of the graph H. If the orientations of any pair of faces adjacent across one of these do not oppose one another, then we there cannot exist an orientation; moreover we have created a cycle of faces

(corresponding to the unique path in T' between the two faces, and the edge between them) forming a Möbius strip inside the surface. □

Exercise 6.5.3. Another option is to try to orient the *vertices* of the surface consistently. A *vertex orientation* is a choice of circulation direction for the collection of corners of faces incident at the vertex. This induces a transverse orientation of the incident edges. Two vertices connected by an edge must induce opposite transverse orientations of the edge.

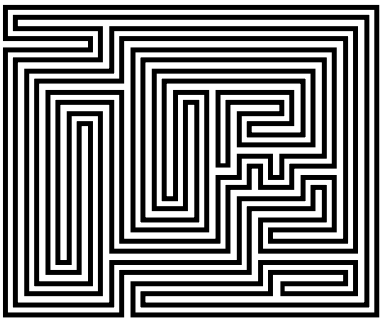
Prove that a surface is vertex-orientable if and only if it is face-orientable. (Hint: consistent circulations on the neighbourhoods of the vertices induce consistent ones on the triangles, and vice versa.)

Exercise 6.5.4. By a variation of the proof of the lemma above, show that a surface is vertex-orientable (and hence face-orientable, by the previous exercise) if and only if it does not contain a 1-sided curve.

Exercise 6.5.5. For a (connected) surface embedded in $\mathbb{R}^3$, yet another definition is available. Show that a surface is orientable if and only if it is possible to colour each of its triangles red on one side, blue on the other, such that adjacent faces have the same colour on the same side. Thus, the surface is orientable if and only if it is a *2-sided surface*. (Warning: the concept of being 2-sided has no meaning for a surface inside $\mathbb{R}^4$ or a higher dimensional space. But it is a helpful concept when working with surfaces in $\mathbb{R}^3$, as we will do in the next chapter.)

6.6. **The Jordan Curve Theorem.**

It is intuitively obvious that any non-self-intersecting polygon in $\mathbb{R}^2$ separates the plane into an "inside" and an "outside", but how do we prove this? Consider the following example!



Definition 6.6.1. A *path* in $\mathbb{R}^2$ is a union of some finite number n of arc segments: $P = [a_0, a_1] \cup [a_1, a_2] \cup \cdots \cup [a_{n-1}, a_n]$. We say P is *closed* or a *curve* if $a_n = a_0$. We say P is a *simple closed curve*, or for short simply a *polygon*, if the arc segments are all disjoint from one another except for their obvious intersections at endpoints.

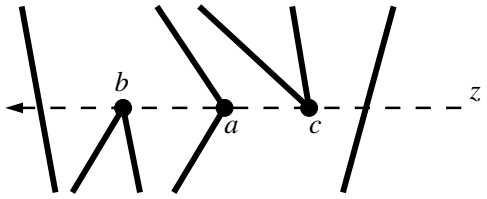
Definition 6.6.2. We say a subset A of $\mathbb{R}^2$ is *path-connected* if each pair of points $x, y \in A$ can be joined by a path lying in A. More generally, the points of a subset A of $\mathbb{R}^2$ can be partitioned into *path-components*, which are the equivalence classes obtained under the equivalence relation: $x \sim y$ if and only if x and y can be joined by a path in A.

Theorem 6.6.3 (Polygonal Jordan Curve Theorem)**.** *Let P be a polygon in $\mathbb{R}^2$. Then $\mathbb{R}^2 - P$ is disconnected.*

Proof. The idea is to decide whether a point z of $\mathbb{R}^2 - P$ is inside or outside by counting how many times, mod 2, we cross P as we move along a path from infinity to z. This is because each time we cross P, we must switch from the inside to the outside or vice versa. So we try to take this idea as the *definition* of "inside" and "outside" and prove that they are distinct.

It's easiest to assume that P has no vertices at the same height (y-coordinate), and in particular no horizontal edges. (We can rotate P a bit, so that none of the $\binom{n}{2}$ lines spanned by pairs of vertices of P is horizontal.)

Then we can assign a colour 0 or 1 to each point z not on P as follows. Let L_z be the left half of the horizontal line through x. If L_z does not hit any vertex of P, let $c(z)$ be the number of points of intersection of L_z and P. (This is a finite number, since there are no edges of P lying along L_z.) If L_z does hit a vertex of P (note that it can only hit one!), there are three configurations $(a, b, c$ below) according to whether the edges joining at the vertex go up or down: we count the intersection point only if it is of type a, and not if it is of type b or c.



Now consider how the colour $c(z)$ changes if we move z along a straight line in the plane from x to y, not crossing P. Unless L_z moves past a vertex of P at some moment, every edge of P that crosses L_x also crosses L_y , and so $c(x) = c(y)$. If we do pass a vertex, although the actual number of intersections might change by 2 (if the vertex is of type a or b), the number mod 2 does not

change. Therefore c is constant along all paths not crossing P. Since c obviously does take both values 0 and 1 (if P is not empty!), $\mathbb{R}^2 - P$ cannot be connected. □

The fact that every closed loop in the sphere separates it is in fact a way to *characterise* the sphere: among all closed 2-dimensional surfaces, *only* equivalent to S^2 have this property. This will be the basis for the classification of surfaces in the next section.

Remark 6.6.4. Although the above is all we actually need for classification of surfaces, it is important to consider various extensions and generalisations of the theorem.

(1). *P cuts $\mathbb{R}^2$ into exactly two connected pieces.* That's because the regluing of $\mathbb{R}^2 - P$ to obtain $\mathbb{R}^2$ is done along two connected curves (the 2 sides of P) and so if there were three or more components, there would be pieces not containing one of these curves which were unaffected by the regluing, and therefore still disjoint afterwards.

(2). *What goes in must come out.* Any closed path in $\mathbb{R}^2$ which is transverse to P (meets it in finitely many points, none a vertex of P) hits it an even number of times. That's because $c(z)$ changes from 0 to 1 or vice versa every time we cross P, and must return to its starting value as we complete the loop. (We could allow the path to hit vertices of P as long as we count them according to the configuration a, b, c, as above, but clearly the statement becomes silly if the path runs along P hitting it infinitely many times!)

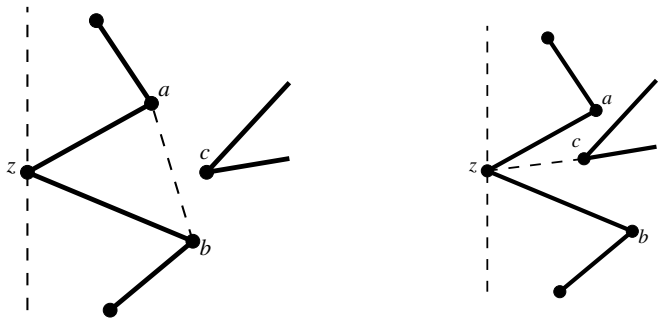
(3). The same theorem holds for the sphere as well as the plane

(4). There is a stronger form of the polygonal JCT which we used implicitly, right at the start of the course. The claim that a polygonal knot diagram with no crossings can be Δ-moved to a standard triangle, and therefore represents the unknot, actially makes use of the following theorem, which is sometimes referred to as the Schönflies theorem, but often just as the JCT.

Theorem 6.6.5. *Given any polygon P in $\mathbb{R}^2$, it's possible to divide the inside region into triangles by adding edges connecting vertices of P (without adding new interior vertices). Moreover, the dual graph of the triangulation we get (place vertices at barycentres of triangles and connect them by edges when their faces meet at an edge) is a tree. Hence the inside piece of $\mathbb{R}^2 - P$ is a surface with boundary and is equivalent to a disc.*

Proof. This time we assume there are no vertices of P with the same x-coordinate, and induct on the number of sides of the polygon. Let z be the left-most vertex of P, and a, b the vertices its neighbouring vertices in P.

If the triangle zab contains no other parts of P then we may simply cut it off, removing z and replacing it by an edge ab. The new polygon P' is divided into triangles with a dual tree, by inductive hypothesis, and therefore so is P.



If on the other hand zab does contain some other piece of P, let c be the leftmost vertex of this part of P inside zab. Divide P into two polygons using a new edge zc; each of these is shorter than

P, so by inductive hypothesis may be divided into triangles with a dual tree, and so P may be too, by taking the union of the divisions, and its dual tree is the obtained by gluing the two prior trees using one new connecting edge dual to zc, so it is also a tree.

Any collection of triangles glued edge-to-edge so that the dual graph is a tree is a disc, by the "disc union along arc with disc is disc" principle. □

As a corollary we get the theorem about Δ-moving P to a triangle: the tree of interior triangles shows how to trim off corners from P one at a time (find a univalent vertice of the tree, and cut off the triangle corresponding to this) until it has only three edges left.

(5). *The full topological form of the JCT.* The proper point-set topology statement of the Jordan Curve Theorem is the following, which is really quite difficult to prove.

Theorem 6.6.6. *If $C \subseteq \mathbb{R}^2$ is a subset which is homeomorphic to a circle, then it separates $\mathbb{R}^2$ into precisely two components, and the interior one is homeomorphic to an open disc.*

The corresponding fact on the sphere says that if $C \subseteq S^2$ is a subset which is homeomorphic to a circle, then it separates S^2 into precisely two components, and each one is homeomorphic to an open disc.

(6). *Analogues in higher dimensions.*

The *separation* part of the JCT is still true in higher dimensions. For example, a subset of $\mathbb{R}^3$ which is homeomorphic to a 2-sphere separates $\mathbb{R}^3$ into an inside and an outside. But the *Schönflies* part is more subtle! For a *piecewise-linear* (made out of triangles) surface equivalent to a sphere inside $\mathbb{R}^3$, the inside will indeed be the expected three-dimensional open ball. But if the sphere is *wild*, this might not be true: the inside of the Alexander horned sphere is *not* homeomorphic to an open 3-ball! In four dimensions (the hardest of all dimensions!) we don't even know whether the inside of a triangulated 3-sphere is equivalent to the standard 4-ball: this is the unsolved *4-dimensional Schönflies conjecture*!

6.7. **Recognising the 2-sphere.**

We want to prove what amounts to the *converse* of the Jordan Curve Theorem.

Theorem 6.7.1. *Let Σ be a connected closed surface. Then the following are equivalent:*

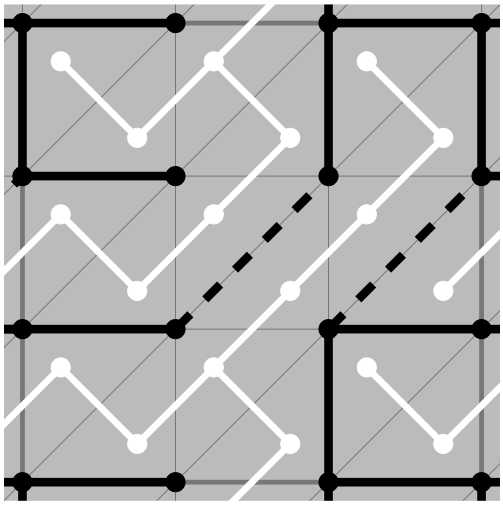
(1) *Σ is a sphere*

(2) *Every curve on Σ separates it*

(3) *$\chi(\Sigma) = 2$.*

Proof. The Jordan curve theorem tells us that (1) implies (2).

Now recall the earlier idea of picking a maximal tree T' in the dual graph, leaving a complementary connected graph H. Write the Euler characteristic as $\chi(H) = 1 - q$ for some $q \geq 0$, so that the Euler characteristic of the whole surface is $2 - q$. Pick a maximal tree T in H by deleting q edges suitably. In the earlier example, $q = 2$, and we can choose the two dashed edges shown below for deletion from H to obtain a black tree T.



If $q > 0$ then choose just one of the q deleted edges and reattach it to T. Since T was a tree, this creates a cycle consisting of a path in T and the new edge. This cycle clearly does not separate Σ because all the faces remain connected by the white tree T'. Therefore we've shown that $\chi(\Sigma) < 2$ implies there exists a non-separating curve, and hence (2) implies (3).

If $\chi(\Sigma) = 2$ then $q = 0$ and H is itself a tree. The whole surface can be expressed as a union of a small neighbourhood of T and a slightly-shrunk solid polygon, each of which is a disc, and therefore Σ is the union of two discs which is a sphere. □

Corollary 6.7.2. *The Euler characteristic $\chi(\Sigma)$ is a topological invariant: it depends only on the equivalence class of Σ, and is independent of the choice of gluing pattern.*

Proof. Subdiving a surface entails adding new vertices on its edges, splitting them into smaller edges, as well as dividing the interior of each face into smaller parts. Each vertex added to an edge splits one edge into two, so does not affect the Euler characteristic. Therefore all we need to do is show that whenever a single triangle is subdivided, the Euler characteristic of its internal parts equals 1 (the contribution for the single original face). But the Euler characteristic of the boundary

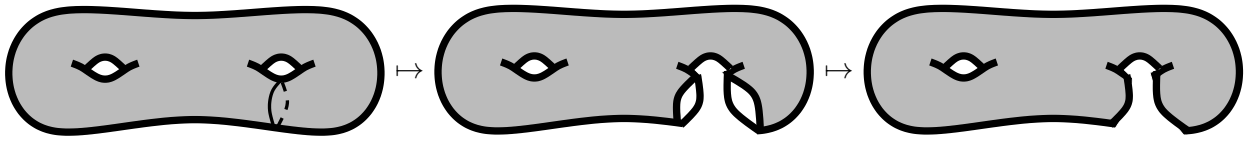
is zero (it's a polygon with as many vertices as edges) so all we need to do is show that the Euler characteristic of any subdivided triangle is 1.

To do this, double the triangle and glue the two copies along their boundaries. This makes a sphere, whose Euler characteristic 2 is twice that of the triangle, minus that of the boundary of the triangle (which is zero). QED. □

6.8. **The classification theorem.**

The classification of closed connected combinatorial surfaces is actually based on a very simple idea called *surgery*.

Definition 6.8.1. If C is a curve in Σ, then *surgery on* C is the operation of cutting Σ along C and then "capping off" each boundary component arising (there will be one or two) by gluing a cone of the appropriate number of sides onto it. (If C has d edges and is 2-sided then one will need two d-sided cones, but if C is 1-sided one needs one $2d$-sided cone.) Let us call the resulting surface Σ_C.

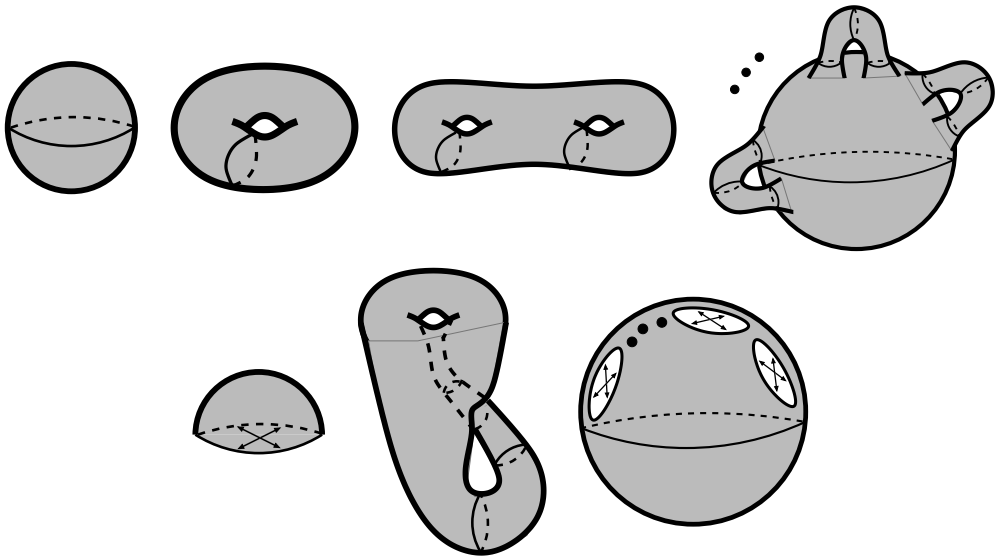


Exercise 6.8.2. If Σ' is obtained by cutting Σ along C, then $\chi(\Sigma') = \chi(\Sigma)$.

Exercise 6.8.3. If Σ_C is obtained by doing surgery along C, then $\chi(\Sigma_C)$ is either $\chi(\Sigma) + 1$ or $\chi(\Sigma) + 2$, depending on whether C is 1-sided or 2-sided.

In the last subsection we showed that a closed surface with no non-separating curves must be a sphere. Suppose we are given a closed surface Σ which does have at least one non-separating curve. We'll perform surgery on it, and repeat this process until there are none left, at which point we must have a sphere. Rebuilding the original surface by reversing the surgeries (rather like rebuilding a pruned tree via sprouting) makes it easily identifiable, and we'll get the following theorem.

Theorem 6.8.4 (Classification of closed surfaces)**.** *Any closed connected surface Σ is equivalent to exactly one of the surfaces M_g ($g = 0, 1, 2, \ldots$; a "sphere with g handles") or N_h ($h = 1, 2, 3, \ldots$; a "sphere with h crosscaps") shown below.*



The surfaces can be identified by their orientability and Euler characteristic: the M_g are orientable with $\chi(M_g) = 2 - 2g$, whereas the N_h are non-orientable with $\chi(N_h) = 2 - h$. Thus the Euler characteristic and the orientability form a "complete set of invariants" for closed connected surfaces.

Proof. The proof is best stated as an algorithm. We will first construct a finite sequence of closed connected surfaces $\Sigma = \Sigma_0, \Sigma_1, \ldots, \Sigma_k = S^2$, where each Σ_{i+1} is obtained from its predecessor Σ_i by surgery. Then, reversing direction, we'll rebuild Σ starting from the sphere.

To construct Σ_{i+1} from Σ_i, recall that $\chi(\Sigma_i) \leq 2$, by lemma 6.4.10. If $\chi(\Sigma_i) = 2$ then Σ_i is a sphere (and has no non-separating curves) by corollary **??**, so we are finished (with $k = i$). If instead $\chi(\Sigma_i) < 2$; then Σ_i is not a sphere, so it must contain a non-separating curve C_i. Doing surgery on C_i gives a closed surface Σ_{i+1} which is *still connected*, because C is non-separating, and with $\chi(\Sigma_{i+1})$ greater than $\chi(\Sigma)$ by 1 or 2, depending on whether C_i is 1- or 2-sided. Because of the overall bound 2 on Euler characteristic of closed connected surfaces, this process terminate in finitely-many steps.

To rebuild Σ we have to undo the effects of the surgeries, starting from S^2. A reversed surgery involves either removing a single (even-sided) cone and gluing the boundary up by identifying antipodal points (in other words, attaching a crosscap) or removing two cones and gluing the boundary circles together (attaching either a handle or twisted handle). Therefore any Σ is equivalent to a sphere with a handles, b twisted handles and c crosscaps attached, for some $a, b, c \geq 0$. (It doesn't matter where or in what order they are attached.) Since a twisted handle is worth two crosscaps, and a handle is worth two crosscaps *provided there is one there to start with* (see the earlier visualisation exercises), such a surface is equivalent either to M_a (if $b, c = 0$) or to $N_{2a+2b+c}$ (if $b + 2c \geq 1$).

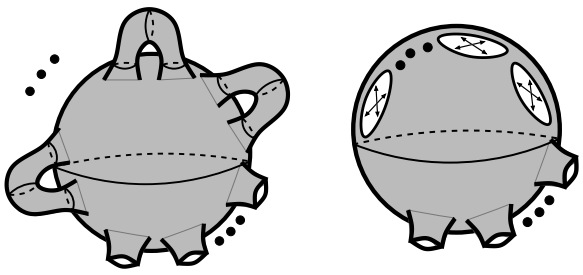
The rebuilding process computes the Euler characteristic for us. Starting from $\chi(S^2) = 2$, each attachment of a handle or twisted handle (reversal of a surgery on a 2-sided curve) decreases the Euler characteristic by 2 (remember that the surgery increased it by 2), and each attachment of a crosscap (reversal of a surgery on a 1-sided curve) decreases it by 1. Therefore, the Euler characteristic of a surface which gets reconstructed using a, b, c such things (as above) is $\chi(\Sigma) = 2 - 2a - 2b - c$. But if $\Sigma \cong M_g$ then $c = 0, a = g$ and hence $\chi(\Sigma) = 2 - 2g$, whilst if $\Sigma \cong N_h$ then $h = 2a + 2b + c$ so that $\chi(\Sigma) = 2 - h$.

Finally, we know that the Euler characteristic and orientability are topological invariants, so no two of the surfaces M_g and N_h could be equivalent. □

Remark 6.8.5. Surfaces with odd Euler characteristic must be non-orientable (since $2 - 2g$ is always even), but surfaces with even Euler characteristic need not be orientable! So surfaces with odd Euler characteristic are identified simply by that number, whereas this is not true for in the even case (both the torus and Klein bottle, for example, have Euler characteristic zero).

Remark 6.8.6. The *genus* g of a closed surface Σ is defined by $g(\Sigma) = 1 - \frac{1}{2}\chi(\Sigma)$ for an orientable surface and $g(\Sigma) = 2 - \chi(\Sigma)$ for a non-orientable one. Thus, $g(M_g) = g$ and $g(N_h) = h$. This is a more visualisable invariant than the Euler characteristic (it is the number of "holes" (handles) or crosscaps (Möbius strips) of the surface, depending on orientability), and the fact that it is a non-negative integer is also nice. However, it is less useful in calculations than the Euler characteristic, which has a nicer additive behaviour under cutting and pasting.

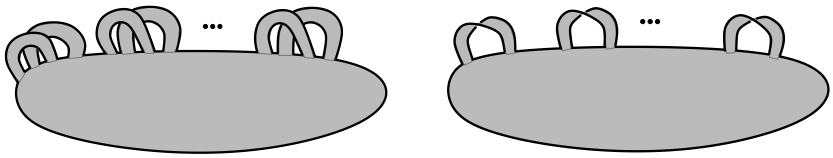
Theorem 6.8.7 (Classification of surfaces with boundary)**.** *Any connected surface Σ with $n \geq 0$ boundary components is equivalent to exactly one of the surfaces M_g^n ($g = 0, 1, 2, \ldots$; an "n-punctured sphere with g handles") or N_h^n ($h = 1, 2, 3, \ldots$; an "n-punctured sphere with h crosscaps") shown below.*



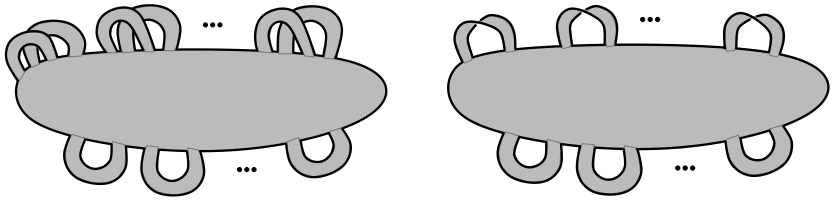
The surfaces can be identified by their number of boundary components, orientability and Euler characteristic: the M_g^n are orientable with $\chi(M_g^n) = 2-2g-n$, whereas the N_h^n are non-orientable with $\chi(N_h^n) = 2-h-n$. Thus the number of boundary components, Euler characteristic and the orientability form a complete set of invariants for connected surfaces.

Proof. We can just use the existing theorem. Given the surface with boundary Σ, cap off each of its n boundary circles with a cone to make a closed connected combinatorial surface $\hat{\Sigma}$ with $\chi(\hat{\Sigma}) = \chi(\Sigma) + n$. This $\hat{\Sigma}$ must be equivalent to one of the M_g or N_h, with $\chi(\hat{\Sigma}) = 2-2g$ or $2-h$ accordingly. Therefore Σ is one of these surfaces with n discs removed, and has the asserted Euler characteristic. Obviously these surfaces are pairwise non-equivalent, since the number of boundary components and the equivalence class of the capped-off surface are topological invariants. The final part is then obvious. □

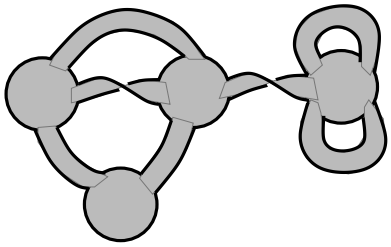
Exercise 6.8.8. Show that any compact connected surface with one boundary component is homeomorphic to one of the following surfaces.



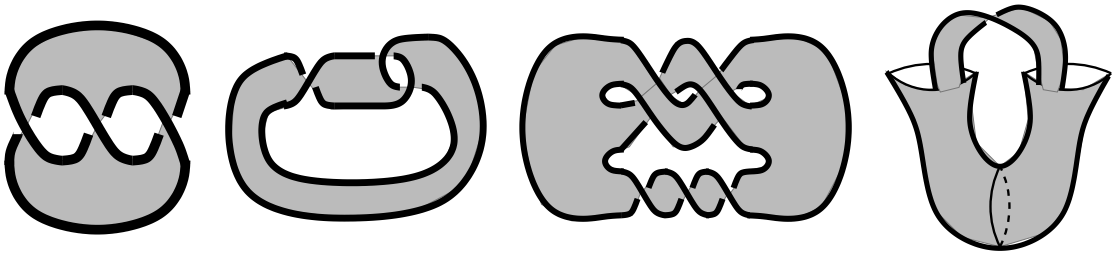
Exercise 6.8.9. Show that any compact connected surface with boundary is homeomorphic to one of the following surfaces.



Exercise 6.8.10. Suppose that a connected surface Σ is made by starting with v closed discs and attaching e bands to them, as in the example. Prove that $\chi(\Sigma) = v - e$. What does the formula suggest to you? (Think about the formula for the Euler characteristic of a graph.)

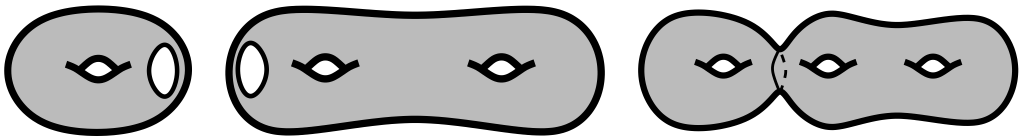


Exercise 6.8.11. Identify the following surfaces.



Exercise 6.8.12. Define the *connected sum* $\Sigma_1 \# \Sigma_2$ of connected combinatorial surfaces Σ_1, Σ_2 to be the surface made by removing a single triangular face from each and gluing the resulting

boundary triangles together. Show that $\chi(\Sigma_1 \# \Sigma_2) = \chi(\Sigma_1) + \chi(\Sigma_2) - 2$ and use this to prove that $M_g \# M_h \cong M_{g+h}$, $N_g \# N_h \cong N_{g+h}$ and $M_g \# N_h \cong N_{2g+h}$.



Exercise 6.8.13. In the last exercise there are actually two possible ways to do the gluing according to the orientations of the glued (triangular) boundary circles (as with the handle and twisted handle). Do these matter?

Exercise 6.8.14. Show (using Euler characteristic and the classification theorem) that cutting a sphere along a curve always results in two discs.

Exercise 6.8.15. Consider a knot projection P drawn on a sphere. By considering what happens when you cut along P, and using the Euler characteristic, show that the number of regions of P is the number of crossings plus 2.

Exercise 6.8.16. Suppose $C_1, C_2, \ldots C_n$ are disjoint curves on a sphere. By considering what happens when you cut along all the curves, and using the Euler characteristic, show that at least one of the complementary regions of the curves is a disc.

Remark 6.8.17. For closed connected 2-manifolds Σ we have shown that every closed curve separates Σ if and only if Σ is equivalent to the 2-sphere. It is natural to ask whether for closed connected 3-*manifolds*, every *closed surface* in M separates M if and only if M is equivalent to the 3-sphere.

This was conjectured by Poincaré around 1900, but he quickly found a rather amazing counterexample. If you glue together the opposite faces of a solid dodecahedron by translating each along a perpendicular axis and rotating by 36 degrees, you get a closed 3-manifold called the *Poincaré homology sphere* for which the conjecture fails.

Actually, the property "every surface in M separates M" is equivalent to the algebraic condition "the abelianisation of the fundamental group $\pi_1(M)$ is finite". The fundamental group of Poincaré's manifold is one with 120 elements called the *binary icosahedral group*, and its abelianisation is trivial. (These concepts will be explained in a later section.) Consequently Poincaré reformulated his conjecture with a stronger hypothesis by dropping the word "abelianisation":

A closed connected 3*-manifold with trivial fundamental group is homeomorphic to the* 3*-sphere.*

This *Poincaré conjecture* was one of the great motivating problems of 20th century mathematics, and drove the tremendous development of the theory of topology of manifolds. Most topologists always believed it to be true, and over the years, there have been quite a few premature announcements of proofs, which turned out not to stand up to careful scrutiny. Rather surprisingly, the Poincaré conjecture for *higher dimensions* (5 or more) was proved in the 60s by Smale and Stallings, and the version for 4-dimensional topological manifolds by Mike Freedman in 1982. But the original 3-dimensional versions withstood all attacks. In 2003 Grigory Perelman released an outline of a proof using differential geometry, and after several years of study in seminars all over the world, various people were able to write complete proofs of his theorem. Perelman was awarded the Fields Medal and the million-dollar Clay Millenium Prize, both of which he refused in a puzzling protest against "ethical standards in mathematics". (In my opinion, ethical standards in mathematics are higher than those in most other disciplines, because there is little room for subjectivity and little money!) There is now only one version of the Poincaré conjecture left-standing: that for smooth (or combinatorial) 4-dimensional manifolds. Nobody expects this to be resolved soon.

Exercise 6.8.18. (1997F) (i). Define a *closed orientable combinatorial surface*, a *non-separating simple closed curve* in a surface, and the *Euler characteristic* χ of a surface. State the classification theorem for orientable closed combinatorial surfaces.

(ii). Show that cutting such a surface Σ along a combinatorial (i.e. made of edges) non-separating simple closed curve yields a surface with 2 boundary components, and that doing surgery on the curve yields a surface Σ' with $\chi(\Sigma') = \chi(\Sigma) + 2$.

(iii). Let $\{C_1, C_2, \ldots, C_r\}$ be disjoint combinatorial simple closed curves in Σ such that $\Sigma - \cup C_i$ is connected. Show that $r \leq 1 - \chi(\Sigma)/2$.

Exercise 6.8.19. (1998F) *In this question, "surface" means "closed connected combinatorial surface".*

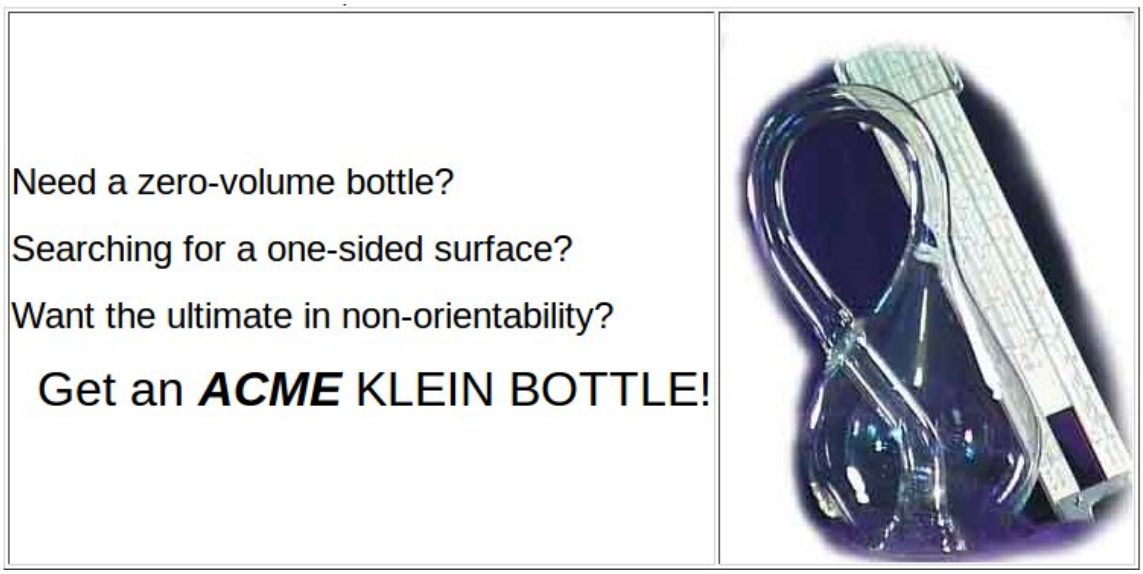
(i). Define the *Euler characteristic* of a surface, and state what it means for a surface to be *orientable*. List the possible homeomorphism types of surfaces, and explain how they may be identified by their Euler characteristic and orientability.

(ii). Define the *connected sum* $\Sigma_1 \# \Sigma_2$ of two surfaces Σ_1 and Σ_2. Show directly (not using the list from (i)) that $\chi(\Sigma_1 \# \Sigma_2) = \chi(\Sigma_1) + \chi(\Sigma_2) - 2$.

(iii). A surface Σ is called *prime* if (a) it is not a sphere and (b) in all splittings as a connected sum of surfaces $F = \Sigma_1 \# \Sigma_2$, at least one of Σ_1, Σ_2 is a sphere. Show that the torus and projective plane are prime, and (using the list from (i)) that they are the *only* prime surfaces.

(iv). A *prime decomposition* of a surface Σ is an expression of Σ as an iterated connected sum $\Sigma \cong \Sigma_1 \# \Sigma_2 \# \cdots \# \Sigma_k$, where the Σ_1 are prime and $k \geq 1$. Give a prime decomposition for each surface on the list from (i) other than the sphere. Show that such decompositions are not unique by finding a surface which has two different decompositions (merely reordering the factors does not count!)

Commercial break sponsored by ACME Klein Bottles (www.kleinbottle.com)



7. Surfaces and knots

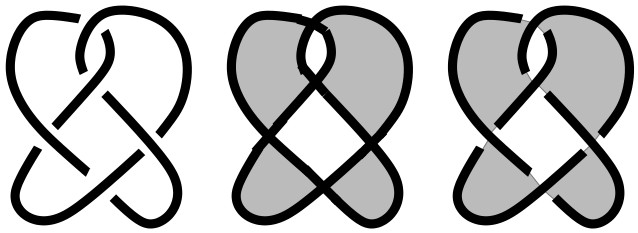
We are now going to use surfaces to study knots, so from now on they will tend to be embedded in $\mathbb{R}^3$. This certainly helps to visualise them, but remember that the way a surface is tangled inside $\mathbb{R}^3$ does not affect its homeomorphism type. All surfaces will be assumed to be combinatorial, despite being drawn "smoothly".

7.1. **Seifert surfaces.**

Definition 7.1.1. If F is a subspace of $\mathbb{R}^3$ which is a compact surface with one boundary component then its boundary is a knot K, and we say that K *bounds the surface* F

Lemma 7.1.2. *Any knot K bounds some surface F.*

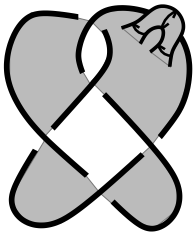
Proof. Draw a diagram D of K, and then chessboard-colour the regions of D in black and white (let's suppose the outside unbounded region is white). Then the union of the black regions, glued together using little half-twisted bands at the crossings, forms a surface with boundary K.



□

Exercise 7.1.3. Why is it possible to chessboard-colour a knot projection in two colours, as we did above?

Remark 7.1.4. Of course, any knot bounds lots of different surfaces. Different diagrams will clearly tend to give different surfaces, and in addition one can add handles to any surface, increasing its genus arbitrarily without affecting its boundary.



One problem with this construction is that the resulting surface may be non-orientable, which makes it harder to work with. Fortunately we can do a different construction which always produces an orientable surface.

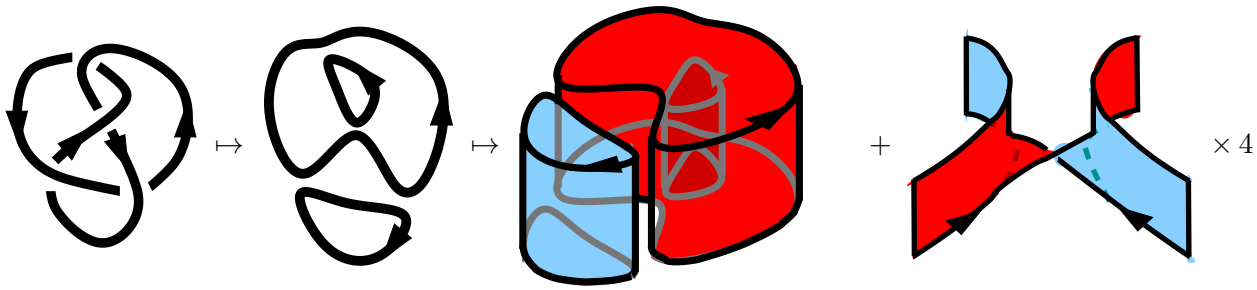
Definition 7.1.5. A *Seifert surface* for K is just a connected orientable surface in $\mathbb{R}^3$ bounded by K.

Lemma 7.1.6. *Any knot has a Seifert surface.*

Proof. (Seifert's algorithm.) Pick a diagram of the knot and choose an orientation for it. Smooth all the crossings in the standard orientation-respecting way (illustrated in the figure below) to obtain a disjoint union of oriented loops called *Seifert circles*. The claim is that if we now fill in each of

these circles with a disc, and reconnect them by attaching half-twisted bands at the crossings, then the result will be an orientable surface.

In order to make the (in general, nested) circles bound disjoint discs in $\mathbb{R}^3$, it's convenient to attach a vertical cylinder to each and then add a disc on top. The height of the vertical cylinders can be adjusted to make the resulting surfaces disjoint (innermost circles in a nest have the shortest cylinders, outermost the tallest). To show that the resulting surface is orientable, move around each Seifert circle in the direction specified by its orientation, colouring the vertical cylinder red on the right-hand side and blue on the left-hand side, and extending this colouring onto the top disc. (This makes the upper side of the disc red if the Seifert circle is overall an anticlockwise-rotating one, and blue if it's overall clockwise. An example is shown below.) Because of this colouring rule, the half-twisted bands we glue on always connect like-coloured sides of the surface, so it really does end up having two distinct sides.



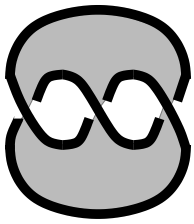
□

Because of this theorem we can immediately define a useful new invariant of knots.

Definition 7.1.7. The *genus* $g(K)$ of a knot K is the minimal genus of any Seifert surface for K.

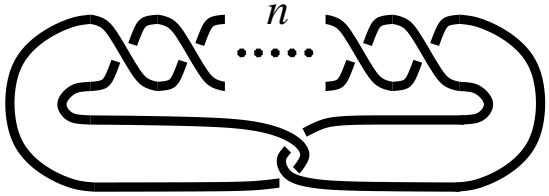
Example 7.1.8. A knot has genus 0 if and only if it is the unknot. This is because having genus 0 is equivalent to bounding a disc in $\mathbb{R}^3$. If a knot bounds a disc, the triangles making up the disc give a sequence of Δ-moves that deform the knot down to a single triangle.

Example 7.1.9. The trefoil has genus 1, because it certainly bounds a once-punctured torus (with genus 1) but is distinct from the unknot, therefore doesn't bound a disc.



Exercise 7.1.10. By viewing the Seifert surface constructed from Seifert's algorithm as a disc-and-band surface (exercise 6.8.10), show that the genus of any knot is bounded in terms of its crossing number by the formula $g(K) \leq c(K)/2$.

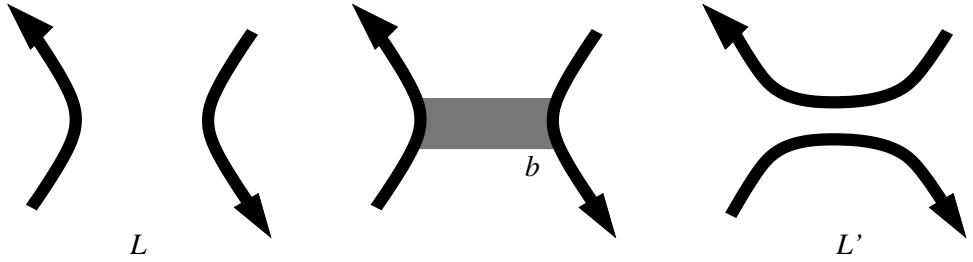
Exercise 7.1.11. Show that all the knots in the family of twisted doubles of the unknot shown below have genus 1.



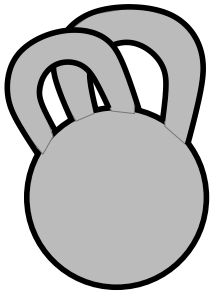
Exercise 7.1.12. (1998F) (i). Show that any knot diagram D can be turned into a diagram of the unknot by changing some of its crossings.

(ii) Define the unknotting number of a knot, and show that it is less than or equal to half the crossing number of the knot.

(iii). If L is any oriented knot or link in $\mathbb{R}^3$, and b is a band meeting L only at its two ends, then we may perform *surgery* on L to produce a new link L', as shown below. (This operation could also be described as "connect-summing L with itself".)



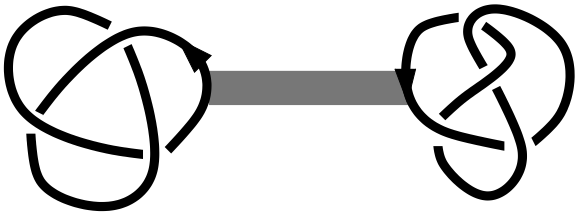
Define the *surgery number* to be the minimum number of surgeries needed to reduce L to an unknot. Suppose K is a knot with genus 1; show that the surgery number of K is less than or equal to 2. *Hint:* Consider the surface below.



(iv). Show similarly that a knot with genus g has surgery number less than or equal to $2g$.

7.2. **Additivity of the genus.**

Definition 7.2.1. If K_1, K_2 are oriented knots then their *connect-sum* $K_1 \# K_2$ is defined as follows. Take any small band in $\mathbb{R}^3$ which meets the knots only in its ends, such that the induced orientations on the ends of the band circulate the same way around its boundary. Then cut out these two arcs from the knots, and join in the other two boundary edges of the band. The resulting knot is then naturally oriented. (The condition on orientations at the end of the band ensures this.)



Remark 7.2.2. The operation is well-defined on equivalence classes of knots, regardless of where the band goes. Even if it is itself highly tangled, the idea of retracting it back and shrinking one of the knots relative to the other makes this clear. Additionally, the operation is commutative and associative.

Remark 7.2.3. If K is a connect-sum, it is possible to find a 2-sphere S contained in $\mathbb{R}^3$ which is disjoint from K except at two points, so that S is a sphere separating the two *factors* of the knot. In the usual picture of the connect-sum, the existence of this sphere is clear. In general, K is a very tangled-up version of this picture; it is *equivalent* to a knot whose two factors are far away and connected by two long strands, but it doesn't actually look like this. However, a separating sphere S will always exist – consider going from the "nice" picture to the "tangled" one via Δ-moves, pushing the sphere along as you go. (Alternatively recall the definition of equivalence of knots in terms of ambient isotopy from section 2.1, which makes it very clear.)

Theorem 7.2.4. *The genus of knots is additive:* $g(K_1 \# K_2) = g(K_1) + g(K_2)$.

Proof. The only thing we really know about the genus is how to bound it from above by just exhibiting *some* Seifert surface for a knot. Consequently, the way to prove this theorem is in two stages, as follows.

($\leq$). Take Σ_1, Σ_2 minimal genus Seifert surfaces for K_1 and K_2. (Imagine the knots far apart so that these surfaces are disjoint in $\mathbb{R}^3$.) Taking the union of $\Sigma_1 \amalg \Sigma_2$ with the band used to construct the connect-sum (this operation, not surprisingly, is called *band-connect-sum* of surfaces) gives a connected orientable (hence Seifert) surface for $K_1 \# K_2$. Using the addititivity property of the Euler characteristic gives

$$\chi(F) = \chi(\Sigma_1) + \chi(\Sigma_2) + 1 - 1 - 1,$$

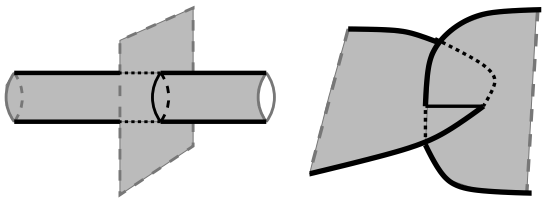
because the Euler characteristic of the band is 1, and its intersection with $\Sigma_1 \amalg \Sigma_2$ consists of two arcs, each with Euler characteristic 1. Therefore, using the formula $\chi = 2 - 2g - 1$ relating the genus and Euler characteristic of an orientable surface with one boundary component, we see that $g(F) = g(\Sigma_1) + g(\Sigma_2)$, and hence

$$g(K_1 \# K_2) \leq g(F) = g(\Sigma_1) + g(\Sigma_2) = g(K_1) + g(K_2).$$

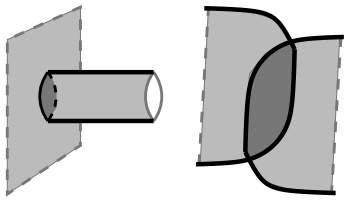
($\geq$). This is a bit harder, as we have to start with a minimal-genus Seifert surface for $K = K_1 \# K_2$ and somehow split it to obtain Seifert surfaces for K_1 and K_2 separately. The argument involves studying the intersection of two overlapping surfaces in $\mathbb{R}^3$, for which we will need the following facts. (Compare with fact 2.2.5 which discusses the perturbations of knots to get regular projections.)

Fact 7.2.5. If F is a surface in $\mathbb{R}^3$, then an *ϵ-perturbation* of F is one obtained by moving the vertices distances less than ϵ (and moving the triangles accordingly). If Σ_1 and Σ_2 are two surfaces contained in $\mathbb{R}^3$, then by an arbitrarily small perturbation of Σ_2 (say) we can arrange that Σ_1, Σ_2 *meet transversely*: that $\Sigma_1 \cap \Sigma_2$ consists of a union of circles disjoint from the boundaries of both surfaces, and arcs whose interiors are disjoint from there boundaries of both surfaces but whose endpoints lie on the boundary of one surface. (The proof of this fact is simply based on what happens for a pair of triangles in $\mathbb{R}^3$.) Some transverse and non-transverse intersections are shown below.

Transverse (good!):



Non-transverse (bad!):

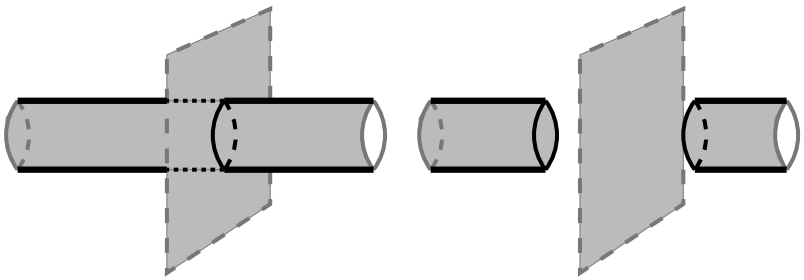


Recall then that F is a *minimal genus* Seifert surface for $K = K_1 \# K_2$, and let S be a separating sphere (remark 7.2.3). Let us make S and F transverse, as explained above. Then their intersection is a union of circles and a single arc, which runs between the two points of $K \cap S$. (All arcs have to end on ∂F since $\partial S = \emptyset$, but $\partial F \cap S = K \cap S$ is just those two points.)

The idea is to repeatedly alter F so that eventually all the circles in $F \cap S$ are eliminated, and it meets S only in the arc.

Consider just the system of circles $F \cap S$ on S (ignore the arc). They should be pictured as nested inside each other in a complicated way. Cutting along them all gets a union of manifolds with non-empty boundary, the sum of whose Euler characteristics is 2 (because they glue along circles to make the whole sphere – compare exercise 6.8.14). Therefore one of them must have positive Euler characteristic, and in fact since $\chi = 2 - 2g - n$ for an orientable surface, this can only happen with $g = 0, n = 1$, i.e. a disc. Let C be its boundary curve: what we have shown is that C is an *innermost* circle amongst those of $F \cap S$, meaning that one component of $S - C$ (the "inside") contains no other circles of $F \cap S$.

Near C the picture of F and S is as shown below on the left. We do surgery on F along C to turn it into F', shown on the right. This procedure can only be carried out when C is innermost, because otherwise the surgery would make F intersect itself.



What kind of a surface is F'? It is certainly orientable since F was. If C had been non-separating in F then F' would be connected, but $\chi(F') = \chi(F) + 2$ means that $g(F') = g(F) - 1$ which would

contradict the minimality of F. Therefore C is separating, and F' has two components: it is *not* a Seifert surface for K. But F' has the same boundary as F, so only one of its two components (call it F'') has a boundary, and the other (call it X) must be closed. We can throw away X and just keep F'', which (being connected and orientable) *is* a Seifert surface for K. The Euler characteristic shows that

$$\chi(F) + 2 = \chi(F') = \chi(F'') + \chi(X).$$

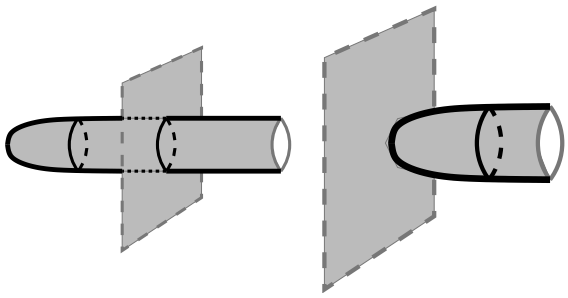
But $\chi(X) \leq 2$ (by lemma 6.4.10) and again by minimality of F, $\chi(F'')$ cannot be bigger than $\chi(F)$. Hence $\chi(X) = 2$, X is a sphere, $\chi(F'') = \chi(F)$, and so F and F'' have the same genus. Note that $F'' \cap S$ is some proper subset of $F \cap S$; at least one (possibly more, when we throw out X) circles of intersection have been eliminated.

Repeat this whole procedure (finding an innermost circle, doing surgery, throwing away spheres) until eventually we have a Seifert surface G with the same genus as the original F and with $G \cap S$ consisting of a single arc. Cutting G along the arc gives a disjoint union $\Sigma_1 \amalg \Sigma_2$ (one part inside the sphere S and one part outside), where Σ_1, Σ_2 are connected orientable surfaces with boundaries (equivalent to) K_1, K_2. Therefore they are Seifert surfaces, and since the sum of their genera is $g(G)$ (another simple Euler characteristic computation just as in the ($\leq$) part), we have our bound:

$$g(K_1) + g(K_2) \leq g(\Sigma_1) + g(\Sigma_2) = g(G) = g(F) = g(K_1 \# K_2).$$

□

Remark 7.2.6. This proof actually shows something a bit better, if we appeal to the *Schönflies theorem* (a three-dimensional analogue of the Jordan curve theorem) that *any sphere in* $\mathbb{R}^3$ *bounds a ball.* The surface X, which is a sphere, must bound a ball, and so each transformation from F to F'' can actually be done by just moving the position of the surface F in $\mathbb{R}^3$ (*isotopy*) rather than by surgery. Therefore the theorem shows that any minimal genus Seifert surface for $K_1 \# K_2$ is a band-connect-sum of minimal surfaces for K_1 and K_2, a much stronger result than the above.



Exercise 7.2.7. Use genus to show that there are infinitely-many distinct knots.

Definition 7.2.8. A knot K is *composite* if there exist non-trivial K_1, K_2 such that $K = K_1 \# K_2$. Otherwise (as long as it isn't the unknot, which like the number 1 isn't considered prime) it is *prime.*

Exercise 7.2.9. Show that any genus-1 knot is prime.

Corollary 7.2.10. *Any non-trivial knot K has a* prime factorisation*, in other words there exist $r \geq 1$ and prime knots $K_1, K_2, \ldots, K_r$ such that $K = K_1 \# K_2 \# \cdots \# K_r$.*

Proof. The proof is basically obvious. If K is prime then we're done: otherwise K is composite, so has a non-trivial splitting $K = K_1 \# K_2$; repeat with K_1 and K_2. The only problem is that the process might never stop. Fortunately, additivity of the genus means that a knot of genus g can't be written as the connect-sum of more than g non-trivial knots, so it does in fact terminate. □

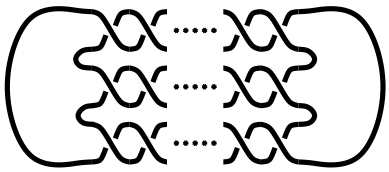
Corollary 7.2.11. *If K is a non-trivial knot, then K connect-summed with any knot J is still non-trivial.*

Proof. By additivity $g(K\#J) \geq g(K) \geq 1$. □

These results demonstrate the similarity between the semigroup of equivalence classes of knots under connect-sum and that of positive integers under multiplication. The last corollary shows that the only element in the knot semigroup which has an inverse is the unknot.

Remark 7.2.12. In fact it can be shown that prime decompositions are unique, in the sense that if $K = K_1\#K_2\#\cdots\#K_r$ and $K = J_1\#J_2\#\cdots\#J_s$ are two prime decompositions of K, then $r = s$ and $K_i = J_i$ (probably after some reordering!).

Exercise 7.2.13. (1998F) Given integers $p, q, r \geq 0$ we define the (p, q, r)-*pretzel link* $P_{p,q,r}$ by the picture below; the three rows consist of p,q and r crossings.



(i). Show that the number of components of $P_{p,q,r}$ depends only on p, q and r modulo 2, and that $P_{p,q,r}$ is a knot if and only if at most one of p, q, r is even.

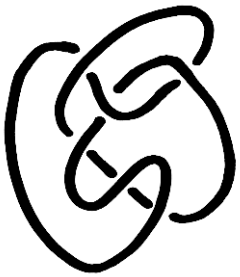
(ii). Show that the number of 3-colourings $\tau(P_{p,q,r})$ depends only on p, q and r modulo 3, and evaluate $\tau(P_{7,13,19})$.

(iii). Show that when p, q and r are odd, $P_{p,q,r}$ is a knot of genus 1.

(iv). State precisely a theorem relating the breadth of the Jones polynomial of an alternating knot with the crossing number of a diagram of it. Use this to show that the knots $P_{3,5,r}$ (for odd r) are all distinct.

Exercise 7.2.14. (1999F) (i). Let K be a knot. Define a *Seifert surface* for K, the *genus* of K, and say what it means for K to be *prime*.

(ii). Use Seifert's algorithm to construct a Seifert surface for the knot diagram shown below, and identify it using the theorem on classification of surfaces.

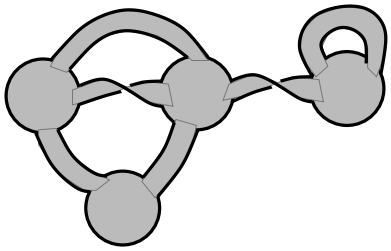


(iii). Compute the genus of the knot. (You may use any theorem provided you state it clearly.)

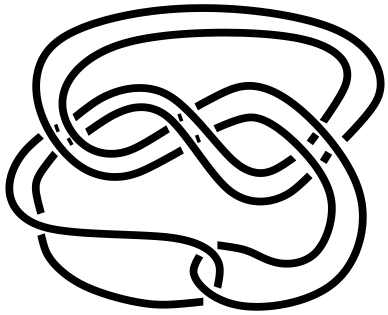
(iv). Is the knot prime? (Again, you may use any theorem provided it is stated clearly.)

Exercise 7.2.15. (1997F) (i). Let K be an oriented knot represented by a diagram D. Describe Seifert's algorithm for constructing a Seifert surface for K from D. Explain how to colour the resulting surface with red and blue to demonstrate that it has two sides, and why this means it is orientable.

(ii) Suppose a connected surface F is built by starting with $v \geq 1$ disjoint closed discs and attaching e bands to them (each end of each band is attached to a small interval in the boundary of one of the discs, but otherwise they are disjoint from the discs, as well as each other). What is the Euler characteristic of F, and why? (The picture shows a typical example of such an F.)



(iii) Apply Seifert's algorithm to the knot diagram shown below and compute the genus of the resulting surface. What bound does this place on the genus of the knot depicted? Is this reasonable, or can you come up with a better bound on the genus of the knot?



Exercise 7.2.16. (2000F) Suppose that A is an annulus embedded (and possibly knotted) in $\mathbb{R}^3$, and call its two boundary components K_1, K_2. Each can be considered by itself as a knot in $\mathbb{R}^3$, but we can also consider both together as a two-component link $L = K_1 \cup K_2$. Suppose that L is a *split link*, that is there exists a sphere S embedded in $\mathbb{R}^3$, disjoint from K_1 and K_2, with K_1 inside it and K_2 outside. (Informally, this means that the two components of L may be disentangled from one another, though they can still be non-trivial knots.)

(i). Explain what it means to say that A and S *meet transversely*. Explain briefly why we can assume that that this is so.

(ii). Now suppose that C is a circle in $A \cap S$. Define what it means to say that C is *innermost* in S, and prove that there is always an innermost circle of $A \cap S$ in S.

(iii). Explain how to change the surface A to a new surface A', again embedded in $\mathbb{R}^3$ and with boundary $\partial A' = K_1 \cup K_2$, but which has the property that $A' \cap S$ has fewer components than does $A \cap S$. What are the two possibilities for the homeomorphism type of A'?

(iv). Use this operation to show that L is in fact a two-component *unlink*, that is that K_1, K_2 are both unknots.

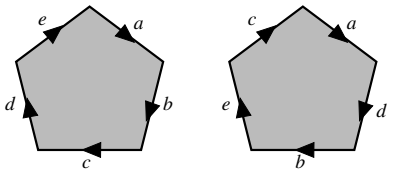
Therefore: if any 2-parallel of a knot is a split link, then the knot is the unknot.

Exercise 7.2.17. (2007F) Use Euler characteristic and orientability to identify each of the following surfaces as one of the standard ones $M_{g,n}$ (orientable, genus g, n boundary circles) or $N_{h,n}$ (non-orientable, genus h, n boundary circles). Explain your reasoning.

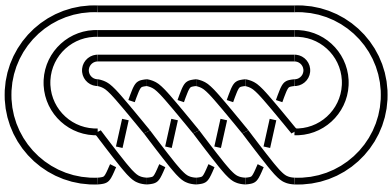
(i). The surface made by gluing two Möbius strips along their boundary circles.

(ii). The connect-sum of a Klein bottle and a torus.

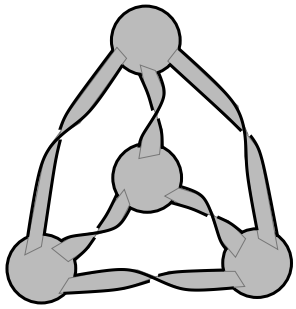
(iii). The surface obtained by gluing two solid pentagons as shown below:



(iv). The Seifert surface obtained by applying Seifert's algorithm to the following knot diagram:



(v). The "discs and bands" surface shown below:



Exercise 7.2.18. (2001F) For each of the following statements, say whether it is true or false. *No proof or explanation is required.* You score two marks for each correct answer and *lose two marks for each wrong answer*, so think carefully before answering: you don't have to answer each part. (The minimum number of marks for this question is zero!)

(1). The surface obtained by taking a torus, removing a disc, and identifying antipodal points of the boundary circle is a Klein bottle.

(2). For every oriented knot K, the value of the Jones polynomial $V_K(t)$, evaluated at $t = 1$, is 1.

(3). For any n, the number of knots with crossing number at most n is finite.

(4). If K is any oriented knot and rK is the knot with reversed orientation, then their Jones polynomials satisfy $V_K(t) = V_{rK}(t^{-1})$.

(5). Any two oriented two-component links with equal linking numbers are equivalent.

(6). The unknotting number u satisfies $u(K_1 \# K_2) \leq u(K_1) + u(K_2)$, for any knots K_1, K_2.

(7). Every closed compact connected combinatorial surface which is non-orientable must contain a 1-sided curve.

(8). There is no knot K with $\tau_3(K) = 6$.

(9). Any two reduced alternating diagrams of the same knot have the same number of crossings.

(10). All closed compact connected combinatorial surfaces with Euler characteristic equal to 1 are homeomorphic.

Exercise 7.2.19. (2007F) Say whether each of the following statements is true or false. (You don't have to give any explanation, or write anything other than "T" or "F".)

(1). A 1-sided curve in a surface is always non-separating.

(2). A 2-sided curve in a surface is always separating.

(3). Any surface with Euler characteristic $\chi \geq 3$ must be disconnected.

(4). Surgery on a curve in a surface always increases the Euler characteristic.

(5). Any non-orientable surface has odd Euler characteristic.

(6). The Jones polynomial of a knot is independent of the orientation on the knot.

(7). The HOMFLY polynomial determines the Jones polynomial of a knot.

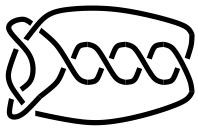
(8). For any knot K, the unknotting number $u(K)$ is less than or equal to half of the crossing number $c(K)$.

(9). The number of 3-colourings τ satisfies the connect-sum formula $\tau(K_1 \# K_2) = \tau(K_1)\tau(K_2)$.

(10). A knot with 27 3-colourings must have unknotting number at least 2.

(11). Let L be a two-component link with linking number 0. If one of the components has its orientation reversed, then the linking number is still 0.

(12). The knot represented by the following diagram has crossing number equal to 8:



Exercise 7.2.20. (2000F) For each of the following statements, say whether it is true or false. *No proof or explanation is required.*

(1). The surface obtained by taking a torus, removing a disc, and identifying antipodal points of the boundary circle is a Klein bottle.

(2). For every knot K, the value of the Jones polynomial $V_K(t)$, evaluated at $t = 1$, is 1.

(3). There are exactly four homeomorphism classes of connected compact combinatorial surfaces (with or without boundary) which have Euler characteristic zero.

(4). If K is any oriented knot and rK is the knot with reversed orientation, then their Jones polynomials satisfy $V_K(t) = V_{rK}(t^{-1})$.

(5). Any two oriented two-component links with equal linking numbers are equivalent.

(6). The unknotting number u satisfies $u(K_1 \# K_2) \leq u(K_1) + u(K_2)$, for any knots K_1, K_2.

(7). Every closed compact connected combinatorial surface with odd Euler characteristic must contain a 1-sided curve.

(8). If J is a knot with the property that $J \# K$ is equivalent to K, for any knot K, then J must be the unknot.

(9). Any two reduced alternating diagrams of the same knot have the same number of crossings.

(10). If F, G, H are three closed compact connected combinatorial surfaces, and $F \# G$ is homeomorphic to $F \# H$, then G must be homeomorphic to H.

8. Van Kampen's theorem and knot groups

In this last section we will study knots by algebraic methods. The main idea is that the fundamental group of the complement of a knot in $\mathbb{R}^3$ gives lots of information about the knot. We will study van Kampen's theorem, a technique for computing fundamental groups of spaces. Since it gives the answer in the form of a presentation, we will have to consider these first.

8.1. **Presentations of groups.**

Definition 8.1.1. If S is a set of symbols $a, b, c, \ldots$, let $\bar{S}$ denote the set of symbols $\bar{a}, \bar{b}, \bar{c}, \ldots$. Define the set of *words in* S, $W(S)$, to be the set of all finite strings of symbols from $S \cup \bar{S}$, including the empty word $\emptyset$. If w_1, w_2 are two words we can concatenate them in the obvious way to make a new word $w_1 w_2$. Also, any word can be written backwards, with all bars and unbars exchanged, giving an operation $w \mapsto \bar{w}$.

Definition 8.1.2. Given a set of *generators* S and a set of *relators* $R \subseteq W(S)$, we can define a group π as follows.

As a set, $\pi = W(S)/\sim$, where $\sim$ is an equivalence relation defined by $w \sim w'$ if and only if there is a finite sequence of words $w = w_0, w_1, \ldots, w_n = w'$ such that each word differs from its predecessor by one of the two operations:

(1). Cancellation: $w_1 a \bar{a} w_2 \leftrightarrow w_1 w_2 \leftrightarrow w_1 \bar{a} a w_2$ (for w_1, w_2 any words and a any generator in S). This allows the insertion or deletion of a bar-unbar pair of generators at any point in a word.

(2). Relation: $w_1 r w_2 \leftrightarrow w_1 w_2$ (for w_1, w_2 any words and r any element of R). An element of R can be inserted or deleted from any point of a word.

Let us write $[w]$ for the equivalence class (element of π) represented by a word w. The multiplication operation is induced by concatenation of words: $[w_1][w_2] = [w_1 w_2]$, the identity is $[\emptyset]$ (denoted by 1 of course!) and the inverse of an element $[w]$ is $[\bar{w}]$.

We say that π has a *presentation* $\langle S : R \rangle$. The only cases we will consider in this section are ones where both S, R are finite sets (π is called *finitely-presented*).

Lemma 8.1.3. *The above procedure really does define a group structure.*

Proof. It should be clear what we have to prove: that the operation of multiplication is actually well-defined (since it's expressed using representatives of equivalence classes), that it is associative, and that the identity and inverse work properly. The whole thing is of course utterly straightforward and boring, but here it is anyway in case you don't believe me. First note that for any words, $u \sim v$ implies both $uw \sim vw$ and $wu \sim wv$, just by sattaching w at the start or finish of all words in a sequence relating u and v. Therefore if w_1, w_1' are representatives for $[w_1]$ and w_2, w_2' for $[w_2]$ then $w_1 w_2 \sim w_1 w_2' \sim w_1' w_2'$ and so $[w_1][w_2] = [w_1'][w_2']$, as required. Associativity is obvious because concatenation of words is associative. Concatenating with the empty word obviously leaves everything unchanged. Inversion is well-defined because any sequence of cancellations and relations also works "when barred" (in particular note that $r \sim \emptyset \implies \bar{r} r \sim \bar{r} \implies \emptyset \sim \bar{r}$, so inverses of relators can also be considered as relators). And finally, for any word w we have $w\bar{w} \sim \emptyset \sim \bar{w} w$ by repeated cancellation of opposite pairs from the middle of those words, therefore $[w][\bar{w}] = 1$. □

In order to give some recognisable examples, we need to have a method of writing down homomorphisms from groups given by presentations to other groups. Suppose $\pi = \langle S : R \rangle$ be a group given by a presentation, and G be some other group.

Lemma 8.1.4. *There is a bijective correspondence between* functions $f : S \to G$ *and* functions $\hat{f} : W(S) \to G$ *which satisfy* $\hat{f}(w_1 w_2) = \hat{f}(w_1)\hat{f}(w_2)$ *for all words* $w_1, w_2 \in W(S)$.

Proof. This is very simple: any $\hat{f}$ defined on $W(S)$ defines an f on S by restricting it to the words of length 1, which include single symbols of S. Conversely, given an f on S, first extend it to $\bar{S}$ by setting $f(\bar{a}) = f(a)^{-1}$ (the inverse is the inverse in G), and then define $\hat{f}$ on a word w by breaking the word into its constituent generators in $S \cup \bar{S}$, taking f of these, and multiplying the resulting elements of G together. Such an $\hat{f}$ obviously satisfies the multiplicative property (note that this identity also implies that $\hat{f}(\bar{w}) = \hat{f}(w)^{-1}$ and $\hat{f}(\emptyset) = 1_G$). These two operations $f \leftrightarrow \hat{f}$ are mutually inverse, giving a bijection. □

Lemma 8.1.5. *Let $\pi = \langle S : R \rangle$ be a group given by a presentation, and G be some other group. Then there is a bijective correspondence between homomorphisms $\theta : \pi \to G$ and functions $f : S \to G$ whose associated $\hat{f}$ functions satisfy $\hat{f}(r_1) = \hat{f}(r_2)$ for any relation $r_1 = r_2$ in R.*

Proof. Any homomorphism $\theta : \pi \to G$ determines a function $f : S \to G$ by setting $f(a) = \theta([a])$, for any generator $a \in S$. Clearly the associated $\hat{f} : W(S) \to G$ in this case is given by $\hat{f}(w) = \theta([w])$, if one carries out the above construction and uses the fact that θ is a homomorphism. Therefore it satisfies $\hat{f}(r_1) = \hat{f}(r_2)$ for any relation, because $[r_1] = [r_2]$ in π.

Conversely, any function $f : S \to G$ determines an $\hat{f} : W(S) \to G$ by the previous lemma. This function satisfies $\hat{f}(w_1 a \bar{a} w_2) = \hat{f}(w_1 w_2) = \hat{f}(w_1 \bar{a} a w_2)$ automatically, because of the way $\hat{f}$ is defined. If the $\hat{f}$ satisfies the extra hypothesis in the statement of the lemma then it also satisfies $\hat{f}(w_1 r_1 w_2) = \hat{f}(w_1 r_2 w_2)$ for each relation. Therefore $\hat{f}$ induces a function $\theta : \pi (= W(S)/\sim) \to G$. Because of the definition of $\hat{f}$, this is a homomorphism.

Again, the two operations are mutually inverse, giving a bijection. □

Remark 8.1.6. There are many notational simplifications to be made. *Relators* are not always terribly convenient, and it is often better to think of *relations*: a relation is an expression of the form $r_1 = r_2$, which is interpreted as meaning that one can replace r_1 anywhere in a word by r_2. Using the relation $r_1 = r_2$ is equivalent to using the relator $r_1 \bar{r}_2$ (in particular, any relator r is equivalent to the relation $r = 1$). Additionally, we usually replace the bars by inverses when writing down relations. The bars used above were simply formal symbols emphasising the distinction between the set of words and the set of equivalence classes (group elements). Once we are happy with the definition there's little need to distinguish between them. Instead of writing $aaaaa$ and $\bar{a}\bar{a}\bar{a}$ we can obviously write a^5 and a^{-3}, and similarly with powers of arbitrary words $w^3 = www$, etc. Finally, we will tend not to bother writing the square brackets after the next couple of lemmas.

Example 8.1.7. In each case below we will define a map from a group π given by a presentation to a group G we understand already, and show that it's an isomorphism, thereby identifying the thing given by the presentation. To define a homomorphism, in view of the above lemma, all we have to do is send each generator of π to an element of G such that the relations are satisfied by these elements in G. Showing surjectivity is usually easy, as we only need to check that the chosen elements of G generate it. But injectivity is trickier, and the alternative, defining a map $G \to \pi$, is also not very easy.

(1). $\langle a \rangle \cong \mathbb{Z}$. We send a to $1 \in \mathbb{Z}$, satisfying all relations (there aren't any). It's onto since 1 generates $\mathbb{Z}$, and injective because any word in $a, \bar{a}$ which maps to 0 must have equal numbers of a's and $\bar{a}$'s, and therefore (by repeated cancellation) is equivalent to the empty word.

(2). $\langle a : a^5 = 1 \rangle \cong \mathbb{Z}_5$. Send a to $1 \in \mathbb{Z}_5$. Now the relation is satisfied, as $aaaaa$ maps to $1+1+1+1+1 = 0$ in $\mathbb{Z}_5$. As in the previous example, this is obviously onto. Again, any word is equivalent (by cancellation alone) to a word of the form a^n, and if this maps to 0 then n must be divisible by 5, and hence the word is actually equivalent to the empty word, using the relation to remove generators five at a time.

(3). $\langle a, b : ab = ba \rangle \cong \mathbb{Z}^2$. Send a, b to $(1, 0), (0, 1)$. Using cancellation and the commutation relation, any word can be made equivalent to something of the form $a^m b^n$, from which we can see the injectivity again.

(4). $\langle a, b : aba^{-1}b^{-1} = 1 \rangle \cong \mathbb{Z}^2$. This just demonstrates that relations can be written in various equivalent ways. Replacing $aba^{-1}b^{-1}$ by the empty word is equivalent to replacing ab by ba (simply post-multiply the equivalence by ba).

(5). $\langle a, b, c : a = 1, b = 1, c = 1 \rangle \cong 1$. Obviously all words are equivalent to the empty word! Note that the same kind of thing with different numbers of generators shows that this number is not any kind of isomorphism-invariant associated with the group. (The *minimal* number of generators over all presentations of a group π is an invariant of π, however.)

(6). $\langle a, b : a^5 = 1, b^2 = 1, bab = a^{-1} \rangle \cong D_{10}$. Send a to the 72 degree rotation of the plane about the origin, and b to the reflection in the x-axis: these elements of the dihedral group do indeed satisfy the relations, and they generate D_10 therefore the map is onto. To show injectivity it is enough to show that there are at most 10 equivalence classes of words, because if a set with 10 or fewer elements surjects onto a 10-element one then the map must be a bijection. Any word can be made equivalent to one made up of alternating symbols a^t $(1 \leq t \leq 4)$ and b, by collecting up adjacent a's and adjacent b's, and using the first two relations to make all powers positive and in the range shown. Then use the third relation to shorten any word with two or more b's into one of the these ten:

$$b, ba, ba^2, ba^3, ba^4, ab, a^2b, a^3b, a^4b, 1$$

to finish.

(7). $\langle a, b \rangle = F_2$ is the *free group* on two generators. This is a group we have not previously encountered. Its elements are simply words in a, b, a^{-1}, b^{-1} of arbitrary finite length, subject only to the equivalence relation of cancellation of adjacent opposites. Thus one can start listing all its elements in order of word-length:

$$1; \quad a, b, a^{-1}, b^{-1}; \quad ab, ab^{-1}, a^2, ba, ba^{-1}, b^2, a^{-1}b, a^{-1}b^{-1}, a^{-2}, b^{-1}a, b^{-1}a^{-1}, b^{-2}; \quad \dots$$

It is an infinite group, because one may define a surjection to $F_2 \to \mathbb{Z}$ by sending a, b to 1. It is non-abelian: one can define a homomorphism to S^3 by sending a to a 3-cycle and b to a transposition, and since these images of a and b do not commute, neither do a and b. The free group is really a very strange group indeed: for example, it contains subgroups which are free groups on arbitrarily many generators, a fact which seems quite counterintuitive!

(8). $\langle a, b : a^2 = 1, b^3 = 1, (ab)^5 = 1 \rangle \cong A_5$. View A_5 as the group of rotations preserving a regular dodecahedron. Send a to the 180 degree rotation about the midpoint of some edge and b to the 120 degree rotation about one of the end vertices of that edge. Then their product is rotation about the centre of a face, with order 5. Proving injectivity is not so easy!

(9). $\langle a, b : a^2 = 1, b^3 = 1, (ab)^7 = 1 \rangle$ is isomorphic to the group of orientation-preserving symmetries of the hyperbolic plane preserving a tiling by congruent hyperbolic triangles with angles $(\pi/2, \pi/3, \pi/7)$. (See the picture by Escher!)

Exercise 8.1.8. Show that the alternating group A_4 has a presentation

$$\langle a, b : a^2 = 1, b^3 = 1, (ab)^3 = 1 \rangle.$$

(Define a map from the group with this presentation to A_4 and check that it's an isomorphism.)

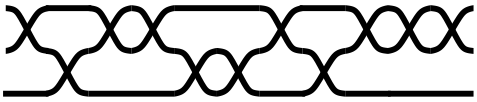
Exercise 8.1.9. Show that the symmetric group S_3 has a presentation

$$\langle a, b : a^2 = 1, b^2 = 1, aba = bab \rangle.$$

Consider the *braid group on 3 strings* B_3 given by the presentation

$$\langle x, y : xyx = yxy \rangle.$$

Show that there is a homomorphism $B_3 \to S_3$, and that B_3 is an infinite group. See if you understand why the following pictures are relevant!



Exercise 8.1.10. Consider the identity braid in the group B_n. Now rotate the top end of the braid through 360 degrees anticlockwise (imagine that the strings are attached to a metal plate and that you rotate the whole plate to do this). What you get is an interesting braid T_n, called the full twist. Find an expression for the full twist in terms of the braid group's generators. (It helps to draw explicit pictures for small values of n first). Then show that T_n commutes with every element of B_n (There are two ways to do this: one entirely pictorial, and one algebraic.)

Remark 8.1.11. Note the big drawback about presentations: in general they reveal no useful information about the group at all. Who would suspect that examples (6), (8) are finite, but (9) is infinite? A presentation is about the least one can know about a group. To get more understanding one usually needs to find something that the group acts on as a group of symmetries. (e.g. the dodecahedron, in example (8).)

8.2. **Reminder of the fundamental group and homotopy.**

This section is a reminder of the definition and properties of the fundamental group of a space. By "map" we mean always "continuous map" in this section. (People who are not comfortable with point-set topology can ignore all this, because van Kampen's theorem, the thing we actually need to understand knot groups, can be applied without understanding any of this.)

Definition 8.2.1. Suppose X, Y are topological spaces. Two maps $f_0, f_1 : X \to Y$ are *homotopic* (written $f_0 \simeq f_1$) if there exists a map $F : X \times I \to Y$ such that F restricted to $X \times \{0\}$ coincides with f_0, and F restricted to the $X \times \{1\}$ coincides with f_1. The *homotopy* F can be thought of as a time-dependent continuously-varying family of maps $f_t : X \to Y$ (where $t \in I$) interpolating between f_0 and f_1. If A is a subspace of X, we can consider *homotopy rel* A, in which f_t restricted to A is always the identity. (Thus two maps can be homotopic rel A only if they already coincide on A.) Homotopy is an equivalence relation.

Definition 8.2.2. If X is a topological space and x_0 some *basepoint* in X, then the *fundamental group* $\pi_1(X, x_0)$ is the set of homotopy classes, rel $\{0, 1\}$, of maps $I \to X$ which send $0, 1$ to x_0. These maps can be thought of as loops in X, starting and ending at x_0, and the relation of homotopy rel $\{0, 1\}$ means that all deformations of loops must keep both ends anchored at x_0. The group multiplication is induced by concatenation of paths (and rescaling the unit interval), and inversion is induced by reversing the direction of loops. The identity element is represented by the *constant loop* $I \to \{x_0\}$.

Example 8.2.3. (1). The fundamental group of $\mathbb{R}^n$ (based anywhere) is trivial, because all maps into $\mathbb{R}^n$ are always homotopic using a *linear homotopy* $f_t(x) = (1-t)f_0(x) + tf_1(x)$, which indeed works rel $\{0, 1\}$.

(2). The fundamental group of $S^n, n \geq 2$ is also trivial, by a Lebesgue covering lemma argument ensuring that any loop is homotopic to one missing the north pole, and therefore to one into $\mathbb{R}^n$, which is homotopic to the constant loop.

(3). For $n = 1$ this "pushing away" argument fails, and indeed $\pi_1(S^1) \cong \mathbb{Z}$ (with any basepoint). To prove this one uses the covering map $x \mapsto e^{2\pi i x}$ from $\mathbb{R}$ to S^1: any map $I \to S^1$ can be lifted into a unique map to $\mathbb{R}$, given a lift of its starting point, and the lift of any loop will end at a value n more than its starting point, where $n \in \mathbb{Z}$. This integer is the *winding number* of the loop, and defines the isomorphism to $\mathbb{Z}$.

(4). If X is a path-connected space then $\pi_1(X, x_0) \cong \pi_1(X, x_1)$, i.e. the isomorphism class of group is independent of the basepoint. The isomorphism is defined by picking a connecting path $\gamma : x_0 \to x_1$ in X, and then sending any loop α at x_0 to the loop $\gamma.\alpha.\gamma^{-1}$, which goes back along γ from x_1 to x_0, around α, then forwards along γ from x_0 to x_1 again. Clearly this is reversible up to homotopy. For this reason we tend to ignore the basepoint when referring to "the fundamental group" of a path-connected space, but it should not be completely forgotten about!

The fundamental group has many important *functorial* properties, describing how maps between spaces induce maps between fundamental groups. These are standard, and state in the lemma below.

Lemma 8.2.4. (1). *If $f : X \to Y$ takes x_0 to y_0 then composing it with loops in X induces a homomorphism $f_* : \pi_1(X, x_0) \to \pi_1(Y, y_0)$. The identity map $X \to X$ induces the identity homomorphisms, and if $g : Y \to Z$ takes y_0 to z_0 then $g_* f_* = (gf)_*$.*

(2). *If $f, g : X \to Y$ both take x_0 to y_0 and are homotopic rel $\{x_0\}$ then $f_* = g_*$ (this is easy).*

(3). *If $f, g : X \to Y$ have $f(x_0) = y_0, g(x_0) = y_1$ not necessarily equal, and they are homotopic, then one has to let γ be the path $t \mapsto F(x_0, t)$ around which the image of x_0 moves during the homotopy $F : f \simeq g$: then $g_*(x) = \gamma f_*(x)\gamma^{-1}$ (compare (4) in the previous example).*

Definition 8.2.5. Two spaces X, Y are *homotopy-equivalent* if there exist maps $f : X \to Y, g : Y \to X$ such that both composites are homotopic to the identity: $fg \simeq 1_Y, gf \simeq 1_X$. A space homotopy-equivalent to a point is called *contractible*.

Lemma 8.2.6. *If X, Y are homotopy-equivalent and path-connected then their fundamental groups (the basepoint being irrelevant) are isomorphic.*

Proof. Since $gf \simeq 1_X$ we have $g_*f_*(x) = (gf)_*(x) = \gamma(1_X)_*(x)\gamma^{-1} = \gamma x \gamma^{-1}$, and therefore g_*f_* is an isomorphism from $\pi_1(X, x_0)$, via $\pi_1(Y, f(x_0))$, to $\pi_1(X, gf(x_0))$. Similarly f_*g_* is an isomorphism from $\pi_1(Y, f(x_0))$, via $\pi_1(X, gf(x_0))$, to $\pi_1(X, fgf(x_0))$. The same g_*'s occur in both these compositions (careful: the f_*'s are actually different, as different basepoints are involved. This is an abuse of notation!), the first being a surjection and the second an injection, so this is an isomorphism. □

8.3. **Van Kampen's theorem.**

The statement of this theorem is rather long-winded, but it's easier than it sounds:

Theorem 8.3.1. *Let X be a topological space containing subsets U, V such that $U, V, W = U \cap V$ are all open and path-connected, and $U \cup V = X$. Let x_0 be a basepoint in W (therefore in U, V too). Let the fundamental groups of U, V, W be given by presentations:*

$$\pi_1(U, x_0) = \langle S_U : R_U \rangle, \quad \pi_1(V, x_0) = \langle S_V : R_V \rangle, \quad \pi_1(W, x_0) = \langle S_W : R_W \rangle.$$

Consider the inclusions $i^U : W \hookrightarrow U, i^V : W \hookrightarrow V$ and their induced maps of fundamental groups i^U_*, i^V_*. For each $g \in S_W$, pick a word $j_U(g) \in W(S_U)$ representing the element $i^U_*(g)$, and a word $j_V(g) \in W(S_V)$ representing the element $i^V_*(g)$. Then $\pi_1(X, x_0)$ has a presentation

$$\langle S_U \cup S_V : R_U \cup R_V \cup \{j_U(g) = j_V(g) : \forall g \in S_W\} \rangle.$$

Remark 8.3.2. In English, what this says is that one starts by taking the union of the presentations of the fundamental groups of the two open subsets U, V. However, any loop in $W = U \cap V$ is then represented by a word in the S_U generators (if one thinks of it as a loop in U) as well as a word in the S_V generators (if one thinks of it as living in V), and since the presentation so far has no relations mixing up the two types of generators, these words represent distinct elements. In the actual fundamental group of X they should represent the same element. Consequently one has to add new relations saying that these two words are equivalent, in order to eliminate the duplication. Fortunately, it is enough to add such a new relation for each *generator* of the fundamental group of W, rather than for each loop, so provided $\pi_1(W)$ is finitely-generated, only finitely-many new relations are added.

Example 8.3.3. Let X be the join of two circles. Let U (V) be the left (right) circle union a small open neighbourhood of the vertex. Each of U, V is homotopy-equivalent to its circle (shrink the extra bits). Then W is a small open cross shape, which is contractible. We can take the presentations $\langle a \rangle, \langle b \rangle$ for the fundamental groups of U, V, and can use the empty presentation for W since its group is trivial. Then the theorem shows that $\pi_1(X)$ is the free group on two generators.

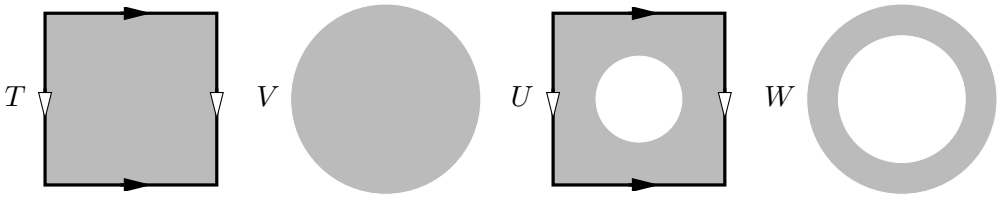
Sketch proof of van Kampen's theorem. (Gilbert and Porter has a full proof). The first stage is to define a homomorphism from the free group $\langle S_U \cup S_V \rangle$ to $\pi_1(X)$, which is done in the obvious way: the generators in S_u, S_V correspond to loops in U and V, and a word in the generators can be mapped to the product of the corresponding loops. That this is a surjection follows from the Lebesgue covering lemma (dissect any path in X based at x_0 into a finite number of smaller paths, each one lying completely inside at least one of U and V) and the path-connectedness of X (at each point of dissection, which lies in X, insert an extra journey (inside X) to the basepoint and then reverse along it before continuing along the next small segment – now the path is visibly a composite of loops, each inside at least one of U or V, which is represented by some word in the generators). Certainly all the relations R_U, R_V and the extra ones of the theorem are satisfied by this map, and it therefore induces a surjective homomorphism

$$\langle S_U \cup S_V : R_U \cup R_V \cup \{j_U(g) = j_V(g) : \forall g \in S_W\} \rangle \to \pi_1(X, x_0).$$

The remainder of the proof is devoted to proving injectivity. One assumes some word in $W(S_U \cup S_V)$ maps to a null-homotopic loop in X and dissects the null-homotopy into small parts (using a similar Lebesgue lemma idea) each of which represents a homotopy in U or in V (which we can already account for). The added relations account for the "change of coordinates" between U and V which can occur on the overlap, and that is all. □

Example 8.3.4. The fundamental group of the torus, computed by van Kampen's theorem. Represent the torus as the square with identification. Let V be a smaller open square, and U be the

whole figure minus a closed square a bit smaller than V, so that the overlap W is a (squareish) open annulus. Let x_0 be in this annulus on one of the diagonals of the big square.



Then U is homotopy-equivalent, via radial projection, to its boundary, which is the figure-of-eight space used above. We may take $S_U = \{a, b\}$ corresponding to the labelled loops (the basepoint of U is the vertex). V is contractible so has trivial fundamental group. W is homotopy-equivalent to a circle by squashing it to its centreline, and so has one generator, a loop g that runs once around the annulus. Including this loop g into V makes it null-homotopic, represented by the empty word. Including it into U makes it homotopic to the path running right round the boundary of the square, which in terms of the "coordinates" S_U is the word $aba^{-1}b^{-1}$. Therefore van Kampen's theorem gives a presentation:

$$\langle a, b : aba^{-1}b^{-1} = 1\rangle,$$

which is of course just the group $\mathbb{Z}^2$.

Exercise 8.3.5. Give an alternative calculation of the fundamental group of the torus by first showing that $\pi_1(X \times Y, (x_0, y_0)) \cong \pi_1(X, x_0) \times \pi_1(Y, y_0)$ for arbitrary spaces X, Y.

Example 8.3.6. A presentation of the fundamental group of the orientable surface M_g is calculated in exactly the same way. This surface may be represented by a solid regular $4g$-gon with its sides identified in pairs according to the scheme (reading around the boundary)

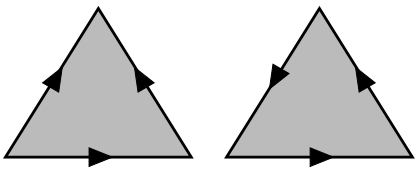
$$a_1 b_1 a_1^{-1} b_1^{-1} a_2 b_2 a_2^{-1} b_2^{-1} \cdots a_g b_g a_g^{-1} b_g^{-1}.$$

(Using this gluing scheme certainly gives an orientable surface, by the "circulation" argument of exercise **??**. It also makes all vertices of the polygon equivalent, and therefore the Euler characteristic of the resulting closed surface, which consists of the disjoint union of an open disc, a vertex and $2g$ edges is $1 - 2g + 1 = 2 - 2g$, proving that this surface is M_g.) Applying exactly the same method as above gives

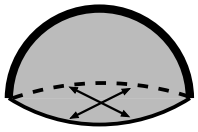
$$\pi_1(M_g) \cong \langle a_1, b_1, a_2, b_2, \ldots, a_g, b_g : \prod_{i=1}^{g} [a_i, b_i] = 1\rangle,$$

where $[a, b]$ denotes the *commutator* $aba^{-1}b^{-1}$.

Exercise 8.3.7. Compute a presentation of the fundamental group of the "dunce cap", a solid triangle whose three edges are all glued together according to the arrows shown. What is the group? Do the same computation for the second space shown below.



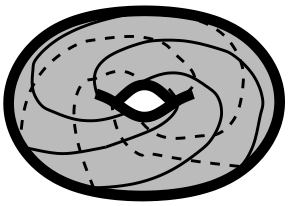
Exercise 8.3.8. Compute the fundamental group of the projective plane (shown below as a hemisphere with antipodal boundary points identified) by applying van Kampen's theorem.



Exercise 8.3.9. Show that the non-orientable surface N_h has a fundamental group

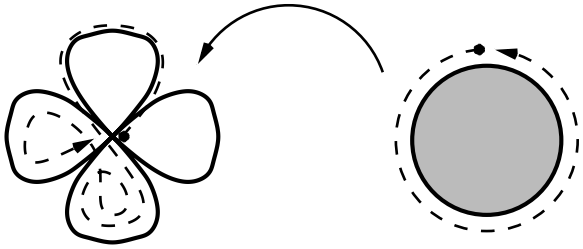
$$\pi_1(N_h) \cong \langle a_1, a_2, \ldots, a_h : \prod_{i=1}^{h} a_i^2 = 1 \rangle,$$

Exercise 8.3.10. Let p, q be coprime positive integers. Compute the fundamental group of the space $L_{p,q}$ formed by attaching two discs to a torus, one along each of the curves drawn in the picture (one is a meridian curve, the other is a (p, q) curve as in example 1.6.1).



Exercise 8.3.11. Compute the fundamental group of an orientable surface M_g^1 of genus g and with one boundary component. What happens to the group when another disc is removed?

Exercise 8.3.12. Suppose X is a bouquet (join or one-point union) of g circles, with basepoint x_0. Let γ be a loop based at x_0. Form a space $X \cup_\gamma D^2$ by starting with $X \amalg D^2$ and identifying $x \in \partial D^2$ with $\gamma(x) \in X$. Let w_γ be a word in the a_i's representing the homotopy class $[\gamma] \in \pi_1(X, x_0)$. Show that the fundamental group of this space has a presentation $\langle a_1, \ldots, a_g : w_\gamma = 1 \rangle >$.



Exercise 8.3.13. What happens if more discs are attached to the bouquet? Deduce that associated to any finite presentation of a group π is a space whose fundamental group is isomorphic to π.

Exercise 8.3.14. Compute the number of homomorphisms from $\pi_1(M_g)$ to $\mathbb{Z}_2$, and conclude that different-genus orientable surfaces have non-isomorphic fundamental groups. Do the same with the groups $\pi_1(N_h)$. Why does this not show that the $\pi_1(N_h)$ and $\pi_1(M_g)$ are *all* pairwise distinct? Show that considering the homomorphisms to $\mathbb{Z}_3$ as well *does* prove all these groups distinct, thereby finally completing the classification of surfaces!

Exercise 8.3.15. The *commutator subgroup* $[\pi, \pi]$ of a group π is the subgroup generated by all commutators (elements of the form $[a, b] = aba^{-1}b^{-1}$) in π. The *abelianisation* π^{ab} of π may be defined intrinsically as the quotient $\pi/[\pi, \pi]$. If $\pi = \langle S : R \rangle$ then the abelianisation has a presentation

$$\pi^{ab} = \langle S : R \cup \{ab = ba : \forall a, b \in S\} \rangle.$$

Compute the abelianisations of the fundamental groups of all closed surfaces. Can you prove they are pairwise non-isomorphic?

Exercise 8.3.16. Show that the commutator subgroup of π lies in the kernel of any homomorphism $\theta : \pi \to A$ between a group π and an *abelian* group A. Deduce that there is a bijection between the set of such homomorphisms and the set of homomorphisms $\psi : \pi^{ab} \to A$. Compute, for each closed surface Σ, the set of homomorphisms $\pi_1(\Sigma) \to \mathbb{Z}$.

8.4. **The knot group.**

Definition 8.4.1. Let K be a knot in $\mathbb{R}^3$. Let X be the *complement* or *exterior* $\mathbb{R}^3 - K$. This is a path-connected (non-compact) 3-manifold. The knot group $\pi(K)$ is defined to be the fundamental group of X. (By path-connectedness, the basepoint is irrelevant).

Remark 8.4.2. There are two ways in which the definition of the knot complement may differ. One is that often people think of knots as lying in S^3, the 3-*sphere*, which is $\mathbb{R}^3$ union a point at infinity. This makes no difference to the knot theory, because knots and sequences of deformations of knots may always be assumed not to hit ∞. Secondly, a small open ϵ-neighbourhood of a knot is homeomorphic to an open solid torus. Removing this neighbourhood gives us a 3-manifold X' *with boundary a torus*. If both these modifications are performed then the result is a *compact* version of the knot complement, which is easier to work with in various ways (the torus boundary is useful too). However, all of these different complements have the same fundamental group, so it's not really important which we actually use (as long as we're consistent).

Remark 8.4.3. (1). "The knot determines the complement". This slogan means that equivalent knots have homeomorphic complements: if one considers the effect of a Δ-move, it should be clear that the complement's homeomorphism type is unchanged under such an operation.

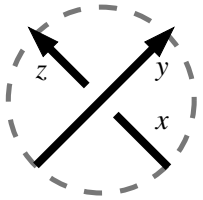
(2). "The knot group is an invariant of knots", because equivalent knots have homeomorphic complements which therefore have isomorphic fundamental groups (it is the isomorphism class of the group which is really considered as the invariant here.)

(3). Much more surprising is the converse theorem: "knots are determined by their complements". This theorem was proved by Gordon and Luecke in 1987, though it had been a conjecture that everybody believed for a very long time. It states, more precisely, that if two knots have homeomorphic complements then they are equivalent (possibly only up to mirror-imaging). (This ambiguity can be removed if one requires an orientation-preserving homeomorphism between the complements.) If you think this is obviously true, think harder until you see why it might not be! The analogous theorem for *links* is immediately false (see example 8.4.8 below).

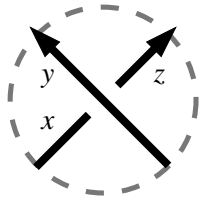
(4). Another surprising thing is that "the knot group determines the knot". Whitten proved that if two *prime* knots have isomorphic groups then their complements are homeomorphic, and hence by the Gordon-Luecke theorem they are equivalent (possibly up to mirror-imaging). The first part of this statement is definitely false for composite knots: example 8.4.10 gives two distinct composite knots with isomorphic groups.

If x, y are two group elements, let x^y denote $y^{-1}xy$, the element obtained from x by *conjugating* it with y. For notational convenience I will use $\bar{y}$ to denote y^{-1} occasionally.

Theorem 8.4.4 (The Wirtinger presentation)**.** *An oriented diagram D of K determines a presentation of $\pi(K)$, having one generator for each arc and one relation for each crossing, as follows. Label the arcs $a_1, a_2, \ldots a_k$, and let $S = \{a_1, a_2, \ldots a_k\}$. Each crossing's relation depends on its sign and incident labels, as shown below. Let R be the set of such relations; then $\pi(K) \cong \langle S : R \rangle$.*



$x^y = z$



$x^{\bar{y}} = z$

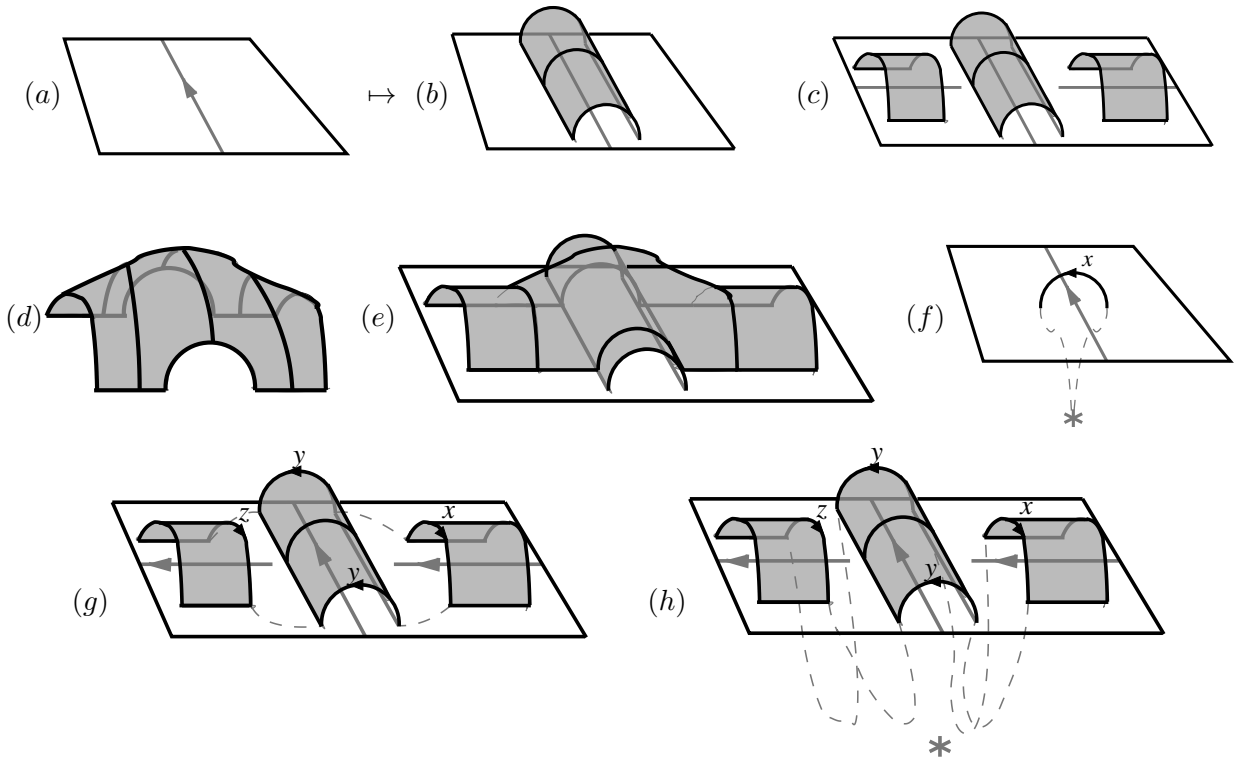
(The relations say "the output label is the input label conjugated by the overpass label (if the overpass crosses from left to right) or its inverse (if right to left)".)

Proof. To make a concrete block containing a knotted hole, you can't make a solid block and then remove the knot! You have to pour concrete in stages: first make the solid bottom half of the block, then carefully build some tunnels so as to "enclose the hole", then pour the rest. Here is a way to do this which allows us to apply van Kampen's theorem in the form of example 8.3.13 to obtain the desired generators and relations.

Make the lower block, then draw the knot diagram D on its top surface. Over each arc of the diagram, glue on a half-cylindrical canopy, as shown in (a) and (b) below; the point of this is to create a tunnel corresponding to the arc when we eventually bury the whole thing. These canopies end where the arcs end, so that near each crossing we get the configuration shown in (c).

We can roof over the gaps at the crossings, thereby completely enclosing the tunnel, by gluing on 'pavilions' of the shape shown in (d), so as to get (e). Having made this guide shape, we now just pour concrete on top to make the a solid block with knot-shaped hole.

The block with the half-cylinder canopies attached is homotopy equivalent to the block with only the middle semicircles of the canopies attached, as in (f). Therefore its fundamental group is free on generators corresponding to the oriented arcs shown; we orient them to run anticlockwise around the oriented arcs of the knot diagram. Each pavilion is a disc, attaching along the sort of curve shown in (g). This curve is homotopic, by pulling down each of its four flat sections to touch the basepoint as in (h), and homotoping its semicircular parts along the canopy to the middle semicircle, to a product of four generators, in this case $y^{-1}xyz^{-1} = 1$, which is the Wirtinger relation $x^y = z$ for a positive crossing like the one shown. Finally, the concrete lying on top of everything is irrelevant (as is the fact that our knot is in a cube rather than all of $\mathbb{R}^3$), since the spaces with it and without it are homotopy equivalent.

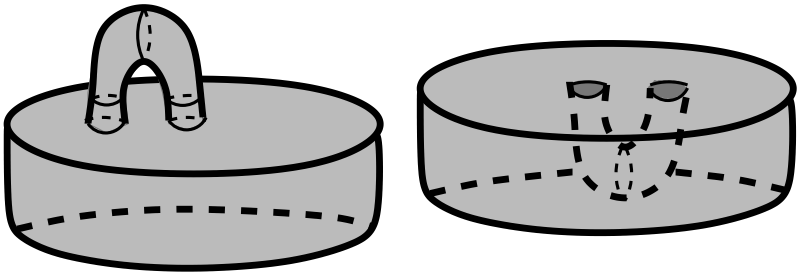


Did you notice the deliberate mistake in the above? The hole we built was actually in the shape of the *mirror-image* $\bar{K}$, and not K! I did this 'wrongly" because it made the perspective in the pictures look a lot better (to me); because the complements of K and $\bar{K}$ are obviously homeomorphic, they have the same knot group, so it doesn't actually make any difference. □

Example 8.4.5. Applying the theorem to the standard picture of the right trefoil (the one whose writhe is +3) gives the presentation

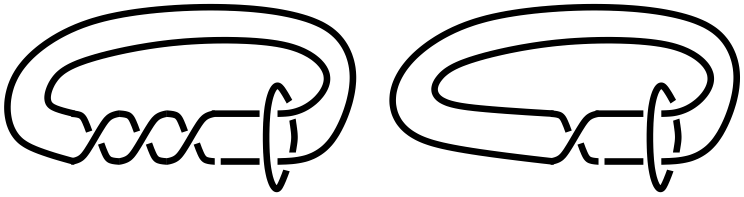
$$\langle x, y, z : x^y = z, y^z = x, z^x = y \rangle.$$

Exercise 8.4.6. Let X be the closed upper half-space with g handle-shaped holes removed from it, and let Y be the same space with g solid handle-shaped protrusions added to it. Show that these spaces are homeomorphic, and further that they are both homotopy-equivalent to a bouquet of g circles.



Exercise 8.4.7. The fundamental group of a link $L \subseteq \mathbb{R}^3$ is defined as the fundamental group of its complement $\mathbb{R}^3 - L$ with respect to some base point. Calculate the fundamental groups of the two-component unlink and the Hopf link.

Exercise 8.4.8. Show that the two links L_1, L_2 shown below have homeomorphic exteriors, thus demonstrating that the statement "*links* are determined by their complements" is false (except of course in the case of 1-component links, i.e. *knots*, where it is true by the Gordon-Luecke theorem).



Exercise 8.4.9. Show that the knot groups of any knot and its mirror-image are isomorphic (explaining the problem with Whitten's result).

Exercise 8.4.10. (Hard!) Write down presentations for the knot groups of the square and reef knots from sheet 2, and show that these groups are isomorphic. (In fact the knot complements are *not* homeomorphic. This counterexample demonstrates that composite knots aren't determined by their groups.)

We can prove knots are distinct by showing that their groups are not isomorphic. In fact, if one appeals to the above theorems, then distinguishing prime knots is exactly as hard as distinguishing their groups! The natural question is: how can we do this? The groups are infinite and we can't make much sense of them just by looking at their presentation s. The simplest answer has already been hinted at in example 8.3.14 when distinguishing the fundamental groups of surfaces: *count homomorphisms into some finite group G* to get an invariant of groups. This idea also also finally explains what our p-colouring invariants really were!

Lemma 8.4.11. *If π, G are groups, let* $\mathrm{Hom}(\pi, G)$ *denote the set of homomorphisms from π to G. If π has a presentation with finitely-many generators and G is finite then* $\mathrm{Hom}(\pi, G)$ *is finite.*

Proof. By lemma 8.1.5, homomorphisms from $\pi = \langle S : R \rangle$ to G are in bijective correspondence with functions $f : S \to G$ such that the associated $\hat{f}$ satisfies $\hat{f}(r_1) = \hat{f}(r_2)$ for each relation $r_1 = r_2$. There can only be finitely-many such f's if S and G are finite. □

Definition 8.4.12. If G is any finite group then we can define an *invariant* of finitely-presented groups $\lambda(-, G)$ by $\lambda(\pi, G) = |\operatorname{Hom}(\pi, G)|$; this is finite by the lemma. Such an invariant $\lambda(\pi, G)$ is computable from any finite presentations of π but doesn't depend on it.

Remark 8.4.13. As usual, it may be that one invariant $\lambda(-, G)$ fails to distinguish two inequivalent groups where another $\lambda(-, H)$ succeeds. Taken together, all such finite-group invariants form a very powerful system, but it is still possible for two inequivalent groups to have equal invariants $\lambda(-, G)$ for all finite groups G.

Remark 8.4.14. In practice, counting the homomorphisms is just a matter of *solving equations in a group* G. For example, to count homomorphisms from the trefoil group to S_3 requires us just to count all solutions $(x, y, z) \in (S_3)^3$ of the "simultaneous equations"

$$x^y = z, y^z = x, z^x = y.$$

This kind of computation can be easily programmed as a quick algorithm on a computer.

Remark 8.4.15. In the case of knots, we can abuse notation and write $\lambda(K, G)$ for the knot invariant $\lambda(\pi(K), G)$. There is an alternative interpretation of $\lambda(K, G)$ in terms of *labellings* of the knot diagram by elements of G. Suppose D is a diagram of K, giving rise to a Wirtinger presentation $\pi = \langle S : R \rangle$ as in theorem 8.4.4. Homomorphisms $\pi \to G$ are simply assignments of elements of G to the arcs of the diagram, satisfying an equation of the form $x^y = z$ at the crossings. Thus $\lambda(K, G)$ is rather like a number of 3-colourings or p-colourings, with group elements replacing the colours.

Remark 8.4.16. There is an additional refinement of the invariant $\lambda(K, G)$. Suppose that we have a labelling of the diagram satisfying the conditions at the crossings. Run around the knot from an arbitrary basepoint, looking at how the labels change. Each time one goes under another strand, the outgoing label is a conjugate of the ingoing one. Therefore (running right around) all labels appearing are conjugate; they lie in some fixed conjugacy class $C \subseteq G$. The set of all such labellings by elements of G is therefore partitioned into subsets according to this conjugacy class. We can therefore define an invariant $\lambda(K, G, G)$ counting just those labellings by elements of the conjugacy class C. Because of the partition one has a sum over all conjugacy classes:

$$\lambda(K, G) = \sum_C \lambda(K, G, C).$$

Theorem 8.4.17. *The number of* 3*-colourings* $\tau(K)$ *of a knot* K *is just the invariant* $\lambda(K, S_3, C)$*, where* C *is the conjugacy class comprising the three transpositions in* S_3*.*

Proof. The three transpositions $a, b, c \in S_3$ have the property that any element conjugated by itself is itself, and conjugated by a different element is the third. Therefore the labellings counted by $\lambda(K, S_3, C)$ are just labellings of the arcs of the diagram by these three transpositions such that at each crossing one sees either a single transposition three times, or each one once. This is exactly the 3-colouring condition. □

Exercise 8.4.18. Show that $\lambda(K, S_3) = 3 + \tau(K)$.

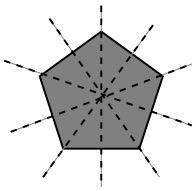
Exercise 8.4.19. Suppose A is a finite abelian group. Show that the number of labellings $\lambda(K, A)$ equals the order of A, regardless of the knot K.

Exercise 8.4.20. How many conjugacy classes are there in the symmetric group S_5, and how many elements are there in each?

Exercise 8.4.21. The *dihedral group* D_{2p} is the group of symmetries (rotations and reflections) of a regular p-sided polygon in the plane (let's assume $p \geq 3$). It has $2p$ elements: how many are reflections? Suppose R_θ is a reflection in a line at angle θ to the x-axis. Show that

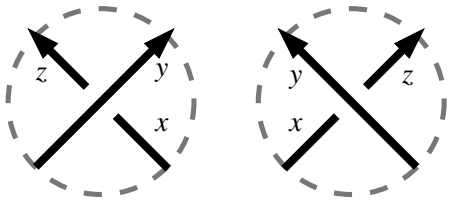
$$R_\theta^{-1} R_\phi R_\theta = R_{2\theta - \phi}.$$

(A geometric rather than coordinate-geometry proof might be easiest.) Show that when p is odd, the set of all reflections in D_{2p} forms a conjugacy class.



Exercise 8.4.22. Let $p \geq 3$ be *prime*. Consider $\lambda(K, D_{2p}, C)$, where C is the conjugacy class of reflections, in other words the number of labellings of a knot diagram by elements of the dihedral group D_{2p} such that every label is a reflection. Suppose the labels at a crossing are written as below, with a label "x" (an integer between 0 and $p-1$) denoting the reflection $R_{2\pi x/p}$. What is the condition on x, y, z for the labelling to satisfy the Wirtinger equation at the crossing? Deduce that this invariant is just the number of p-colourings:

$$\lambda(K, D_{2p}, C) = \tau_p(K).$$



Exercise 8.4.23. Show that $\lambda(K, G)$ does not depend on the orientation of the knot.

Exercise 8.4.24. Compute the number of labellings of the trefoil knot by 3-cycles from the symmetric group S_4.

Exercise 8.4.25. Show that the abelianisation of any knot group π (see exercise 8.3.15) is isomorphic to $\mathbb{Z}$.

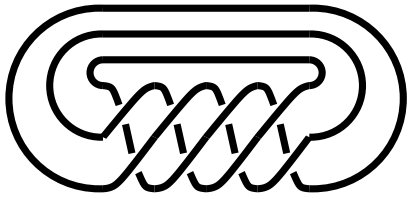
Exercise 8.4.26. Using the notation $x^y = y^{-1}xy$ for conjugation, show that

$$(x^y)^z = x^{yz} \quad \text{and} \quad (x^y)^{(z^y)} = x^{zy}.$$

Write down a presentation for the knot group of the torus knot $T_{3,4}$ (see example 1.6.1) shown below, and show that it is isomorphic to the group

$$\langle p, q : p^3 = q^4 \rangle.$$

Can you see how you might obtain this presentation directly using van Kampen's theorem, and then generalise it to get a presentation of $\pi(T_{p,q})$ for a general torus knot? (This is somewhat related to the earlier exercise about counting p-colourings of torus knots.)



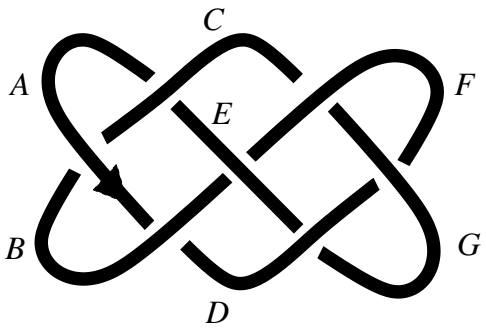
Exercise 8.4.27. (2005F) Let S_3 denote the symmetric group on three symbols $\{1, 2, 3\}$.

(i). Explain why the transpositions in S_3 form a conjugacy class C, and write down the "conjugation table" expressing what happens when one element of C is conjugated by another.

(ii). Suppose K is a knot with knot group $\pi(K)$. Explain why the number of homomorphisms $\lambda(K, S_3, C)$ equals the number of 3-colourings $\tau_3(K)$.

(iii). Give a formula expressing the *total* number $\lambda(K, S_3)$ of homomorphisms $\pi(K) \to S_3$ in terms of $\tau_3(K)$.

Exercise 8.4.28. (1997F) (i). An (oriented) diagram of the knot 7_4, its arcs labelled by the letters $A, B, \ldots, G$, is shown below. Write down a presentation of the knot group π of 7_4, and explain briefly how the group is actually constructed from its presentation (i.e. describe how its elements are represented as equivalence classes and what multiplication, inverses and the identity are - you don't have to verify the group axioms).



(ii). What is meant by a *valid labelling* of the diagram by elements $a, b, \ldots, g$ in S_3? Let Λ be the set of valid labellings: explain briefly why $|\Lambda|$ equals the number of homomorphisms $\pi \to S_3$.

(iii). Let $\Lambda' \subseteq \Lambda$ be the set of valid labellings such that $a, b, \ldots, g$ are transpositions in S_3. Show that the elements of Λ' are exactly the 3-colourings of the knot.

(iv). Compute the number of 3-colourings of the knot 7_4, and deduce that it is not equivalent to the unknot.

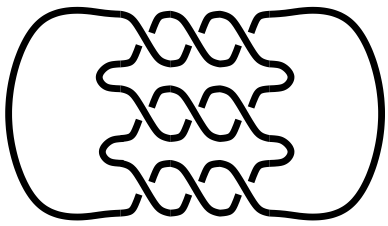
Exercise 8.4.29. (2003F) Say whether each of the following statements is true or false. You don't have to give any explanation, or write anything other than "true" or "false".

(1). For any n, the number of knot diagrams with n crossings is finite.

(2). For any knot K, $\tau_2(K) = 2$.

(3). There exists a knot K with $\tau_{61}(K) = 3721$.

(4). The number of 3-colourings of the following *pretzel knot* is 27.

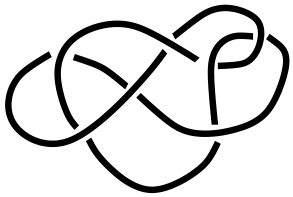


(5). If two knots have the same Jones polynomial then they are equivalent.

(6). The linking number of a link is independent of the choice of orientation of the components of the link.

(7). The braid groups B_n are all abelian groups.

(8). The crossing number of the knot represented by the following diagram is equal to 7.

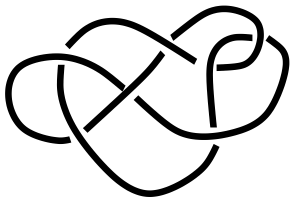


(9). The unknotting number of the trefoil knot is 1.

(10). There exists a knot whose Jones polynomial has breadth 1000.

(11). The number of 3-colourings of a knot with unknotting number 1 must be 3 or 9.

(12). The crossing number of the knot represented by the following diagram is equal to 7.



Exercise 8.4.30. (2005F) Say whether each of the following statements is true or false. You don't have to give any explanation, or write anything other than "true" or "false".

(1). For any knot K, the numbers of 3-colourings of K and its mirror image $\bar{K}$ are equal.

(2). The number of homomorphisms from the braid group B_3 to the group $\mathbb{Z}_7$ (the integers mod 7) is 7.

(3). Let L be a 2-component oriented link with linking number 0. If one component of L has its orientation reversed, then the linking number remains 0.

(4). Let L be a 3-component oriented link with (total) linking number 0. If one component of L has its orientation reversed, then the (total) linking number remains 0.

(5). If two knots K_1, K_2 have equal Jones polynomials, then K_1 and K_2 are equivalent.

(6). For any knot K, the Jones polynomials of K and its mirror image $\bar{K}$ are equal.

(7). The minimal number of crossings of an alternating knot K is equal to the span of its Jones polynomial.

(8). The number of crossings in an alternating knot diagram D equals the span of the Jones polynomial of the knot represented by D.

(9). If two knots K_1 and K_2 have isomorphic knot groups, then they have $\tau_p(K_1) = \tau_p(K_2)$ for all p.

(10). If two knots K_1 and K_2 have $\tau_p(K_1) = \tau_p(K_2)$ for all p then their knot groups are isomorphic.

(11). The HOMFLY polynomial of a link determines the Jones polynomial of that link.

(12). For a knot K with crossing number c, $\tau_3(K) \geq c$.

Exercise 8.4.31. (1999F) For each of the following statements, say whether it is true or false, and give a proof or counterexample.

(i). If F is a combinatorial surface, C is a curve in it, and F_C is the surface obtained by doing surgery on C, then $\chi(F_C) > \chi(F)$.

(ii). The group π of the trefoil, which has a presentation

$$\pi \cong \langle x, y, z : x^y = z, y^z = x, z^x = y \rangle$$

is abelian.

(iii). The unknotting number u satisfies $u(K_1 \# K_2) \leq u(K_1) + u(K_2)$, for any two knots K_1, K_2.

(iv). A connected combinatorial surface with Euler characteristic 0 must be orientable.

9. Appendix: point-set topology

This section is a brief guide to the concepts of topological spaces, continuous functions, and the other basic aspects of point-set topology which we will need during the course.

Point-set topology is not very interesting to teach; it's a language with which to work, rather than an end in itself. In addition, most of the proofs of the theorems "do themselves": there's only really one way to start, in most cases, and it's just a matter of joining the dots, or more precisely, of linking the relevant definitions. These proofs tend to look complicated when written down, because they involve lots of small steps and lots of notation, rather than a single idea which can be expressed in an English sentence. It is therefore usually easier to construct them oneself than to read them from a book. What I'm getting at is: *I'm going to write down very few proofs in this section. Instead, most things are left as exercises, including many standard results which are worth knowing in their own right.* The easiest way to learn them is by doing the exercises.

The primary object of study in algebraic topology is the *topological space*. It is the most general kind of space in which one can do sensible analysis, by which I mean that the notions of continuity, limit etc. make sense. Let's begin working towards the definition by reciting the time-honoured definition of continuity for a real-valued function of a real variable:

Definition 9.0.32. A function $f : \mathbb{R} \to \mathbb{R}$ is *continuous at* $a \in \mathbb{R}$ if for every $\epsilon > 0$, there exists a $\delta > 0$ such that $|x - a| < \delta$ implies that $|f(x) - f(a)| < \epsilon$.

The intuition is, of course, that the function does not jump about locally: if one looks at a sufficiently small range of values about a, then the values of the function may be confined to be arbitrarily close to $f(a)$. Of course, a function is said without qualification to be *continuous* if it is continuous everywhere, i.e. for at all $a \in \mathbb{R}$.

9.1. Metric spaces.

Suppose now that we want to generalise to more complicated kinds of function, such as (let's not get carried away) a real-valued function of two real variables. Obviously the correct thing to do is repeat the same definition with the Euclidean distance $\|x - a\|$ replacing $|x - a|$ now that x, a are points of $\mathbb{R}^2$.

In fact the same principle will work to give a sensible definition of continuity of any function between subsets of a Euclidean space $\mathbb{R}^n$; all that is needed is the notion of distance between pairs of points. In this way, one can quite happily start talking about continuous functions between spheres of arbitary dimensions, because the n-sphere is usually thought of as simply the unit sphere inside the Euclidean space $\mathbb{R}^{n+1}$.

If we want to escape the confines of Euclidean space, it is necessary to abstract away the really essential aspects of Euclidean distance. It turns out that the most important thing is the *triangle inequality*: if you start writing out proofs of the simplest properties of continuous one-variable functions, you will need it pretty quickly.

Let's quickly recall the triangle inequality for $\mathbb{R}^n$ and its proof. We want to prove that for any three vectors x, y, z,

$$\|x - y\| \leq \|x - z\| + \|z - y\|.$$

In other words we need to prove that for every pair of vectors a, b,

$$\|a + b\| \leq \|a\| + \|b\|.$$

Squaring this equation and writing it out in terms of coordinates, it becomes

$$\sum a_i b_i \leq \sqrt{\sum a_i^2} \sqrt{\sum b_i^2}$$

which is just the Cauchy-Schwarz inequality. The proof of this is easy: the quadratic function of λ given by $\sum(a_i - \lambda b_i)^2$ is non-negative, so its discriminant "$b^2 - 4ac$" must be non-positive.

Definition 9.1.1. A *metric space* is a set X equipped with a *metric* function $d : X \times X \to \mathbb{R}_{\geq 0}$ such that

(1). $d(x, y) = 0$ if and only if $x = y$

(2). $d(x, y) = d(y, x)$ for any x, y (symmetry)

(3). $d(x, y) \leq d(x, z) + d(z, x)$, for any x, y, z (triangle inequality).

With this notion, we can make an obvious definition of continuity:

Definition 9.1.2. A function $f : X \to Y$ between metric spaces is *continuous at* a if for every $\epsilon > 0$, there exists a $\delta > 0$ such that $d_X(x, a) < \delta$ implies that $d_Y(f(x), f(a)) < \epsilon$.

Example 9.1.3. Here are some very simple examples of metric spaces and continuous functions.

(1). The basic example is obviously $\mathbb{R}^n$ with the Euclidean distance function, as described above.

(2). The *product* of any two metric spaces becomes a metric space, using the sum of the two metrics:

$$d_{X\times Y}((x_1, y_1), (x_2, y_2)) = d_X(x_1, x_2) + d_Y(y_1, y_2).$$

The "product" $f_1 \times f_2 : X_1 \times X_2 \to Y_1 \times Y_2$ of two continuous functions $f_1 : X_1 \to Y_1$ and $f_2 : X_2 \to Y_2$ is continuous.

(3). For any metric space X, the *identity map* $\mathrm{id}_X : X \to X$ and the *diagonal map* $X \to X \times X$ (given by $x \mapsto x$ and $x \mapsto (x, x)$ respectively) are continuous. The metric itself, as a function $X \times X \to \mathbb{R}$, is continuous.

(4). If $f : X \to Y$ and $g : Y \to Z$ are continuous maps between metric spaces, then their *composite* $g \circ f : X \to Z$ is also continuous.

(5). The functions $+ : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n$ and $\cdot : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n$ defining the sum of two vectors and the product of a vector with a scalar, are continuous. The norm function $\| - \| : \mathbb{R}^n \to \mathbb{R}$ is continuous.

(6). Combining all these facts gives slick ways to prove that explicit functions (which we will sometimes need to write down) are continuous. One simply has to factorise them as composites of "elementary" functions, which one knows to be continous. For example, the radial projection function $\mathbb{R}^n - \{0\} \to \mathbb{R}^n - \{0\}$ given by $x \to x/\|x\|$ is continuous: it can be written as the composite

$$x \mapsto (x, x) \mapsto (x, \|x\|) \mapsto (x, \|x\|^{-1}) \mapsto x/\|x\|$$

where we use the diagonal to duplicate x, then the product of the identity with the norm function, then the product of the identity with the inversion function on $\mathbb{R}-\{0\}$, then the scalar multiplication function.

(7). The metric associated to the usual Euclidean norm $\|x\|_2 = (\sum x_i^2)^{1/2}$ on $\mathbb{R}^n$ is not the only way of measuring distance. Analysts often use the metric associated to the *ℓ_p-norm* (for $p \geq 1$) given by

$$\|x\|_p = (\sum |x_i|^p)^{1/p},$$

in particular the cases $p = 1$ (the norm of a vector is just the sum of the absolute values of the coordinates) and $p = \infty$ case, which denotes the limiting case

$$\|x\|_\infty = \max |x_i|.$$

A simple way to get a feel for these norms is to draw their "unit circles" in the case of $\mathbb{R}^2$: for $\| - \|_1$ one gets a "diamond" (a square with vertices at plus and minus the usual basis vectors) and for $\| - \|_\infty$ one gets the square with vertices $(\pm 1, \pm 1)$.

One might expect that continuity of functions into and out of $\mathbb{R}^n$ would depend on which of these ℓ_p-metrics is used. But this is not the case. Two metrics d, d' on a fixed metric space X are said to be *Lipschitz-equivalent* if there exist constants $K, k \geq 1$ such that for all points x, y,

$$\frac{1}{k} d(x,y) \leq d'(x,y) \leq K d(x,y).$$

The condition means that the metrics distort each others' distances by a bounded amount. If this is the case then the identity map of X, considered as a map between the different metric spaces (X,d) and (X,d') or vice versa, is continuous. Then for example, if $f : (X,d) \to Y$ is a continuous map of metric spaces, so is f thought of as a map starting from (X,d'), because this is the composite of the original f with the continuous identity map $(X,d') \to (X,d)$.

It is easy to see that all the ℓ_p metrics on $\mathbb{R}^n$ are Lipschitz equivalent; geometrically this is just the fact that the unit sphere of any of them can be sandwiched between two (positive radius) unit spheres of any other. (One could work out the best possible constants k, K using this picture, though it isn't necessary to do so. They will depend on n as well as the two choices of p.)

A similar example arises if one tries to decide what is the "natural" metric on the sphere S^n. Probably the first thing one thinks of is to use Euclidean distance between points of $\mathbb{R}^{n+1}$, obtaining the *chordal metric*. (It measures the length of a straight-line chord joining the points inside the sphere.) On reflection, this is quite tasteless: to define it we used geometry external to the sphere. A better choice is to use the great-circle distance on the sphere's surface, measuring "as the crow flies". Fortunately the two metrics are Lipschitz equivalent, so if we are only interested in continuity of functions, it is irrelevant whether we have taste or not.

This last exercise highlights a problem with the use of metric spaces as a foundation of the theory of continuity — it shows that the actual metric itself contains *far more information* than we need when simply thinking about continuity. The notion of a *topological space* will be more economical: it will incorporate only what we actually need.

Another reason for being unhappy with metric spaces is that not all the constructions we hope to perform with spaces work well. We can certainly take subspaces and products of metric spaces and get sensible induced metrics. But the notions of *quotient* is hopeless. Even *disjoint union* of two metric spaces is unpleasant: in $X \amalg Y$, we have perfectly sensible ways of measuring distance between pairs of points of X, and between pairs of points of Y. But what should be the distance between a point of X and a point of Y? Of course ad hoc definitions are available, but there is no canonical, choice-free method.

9.2. **Topological spaces.**

To work towards the definition of a topological space, it helps to rephrase the metric space definition of continuity, avoiding explicit dependence on the metric (which we are trying to get rid of).

Definition 9.2.1. Given a point x of a metric space X and a real number $\epsilon > 0$, let us define the *ball of radius ϵ at x* as

$$B_\epsilon(x) = \{y \in X : d(y,x) < \epsilon\}.$$

Definition 9.2.2. A set N is called a *neighbourhood* of a point $x \in X$ if it contains some ball $B_\epsilon(x)$ of positive radius about x.

Definition 9.2.3. A set U is set to be *open* if it is a neighbourhood of each of its points. Such a set can then be written in the form (check each direction of containment if this seems puzzling!)

$$U = \bigcup_{x \in U} B_{\epsilon(x)}(x).$$

This is a bit of a silly expression, from one point of view: we are writing a set as a union of small balls about all of its points in a very redundant way. However, the intuition that every open set can be expressed as some huge union of special kinds of standard small open sets is a valuable one.

Lemma 9.2.4. (*Local form*) *A function $f : X \to Y$ is continuous at $a \in X$ if and only if, for each neighbourhood N of $f(a)$, the inverse image $f^{-1}(N)$ is a neighbourhood of a.*

(*Global form*) *A function $f : X \to Y$ is continuous if and only if, for each open set U in Y, the inverse image $f^{-1}(U)$ is open in X.*

Proof. It's straightforward to check necessity and sufficiency in each case. □

This lemma then, removes the explicit dependence on the metric, as we desired — the *open sets* of a metric space provide enough information for us to talk about continuity. The conceptual leap to a topological space is then simply the realisation that we may as well only specify these open sets, rather than a metric. Remarkably, a few simple axioms suffice to make the structure behave (for the most part) in the way we have come to expect.

Definition 9.2.5. A *topological space* is a set X together with a set τ_X (its *topology*) of subsets of X, whose elements are referred to as the *open sets*, satisfying the axioms:

(1). the whole space X, and the empty set $\emptyset$ are open

(2). the union of an *arbitrary* family of open sets is again open

(3). the intersection of *finitely many* open sets is again open.

Definition 9.2.6. A function $f : X \to Y$ between topological spaces is *continuous* if for every open U in Y (i.e. $U \in \tau_Y$), the inverse image $f^{-1}(U)$ is open in X (i.e. $f^{-1}(U) \in \tau_X$).

I have chosen to write the "slogan" versions here, instead of emphasising the set-theoretic notation, which require one to be very careful not to confuse the symbols $\in$ and $\subseteq$.

It is convenient to define a *neighbourhood* of a point x in a topological space X to be any set which contains an open set containing x. (When X is a metric space, this coincides with our original definition.) It is then possible to define continuity of a function locally (that is, at a point) in terms of neighbourhoods, just as we did for metric spaces.

These definitions are somewhat frightening, and not just because all the geometry appears to have gone out of the window. The structures involved (topologies) can be absolutely enormous, and the whole apparatus appears unmanageable. Fortunately, the intuition developed by thinking with metric spaces is surprisingly helpful for understanding topological spaces, and after working through analogues of the basic theorems (and playing with some of the standard counterexamples) they begin to seem quite visualisable. As for the amount of structure being carried around — well, the metric on a metric space carries more information than the topology it defines (see below); it's just that it somehow seems "smaller".

Example 9.2.7. Here are some standard examples of topological spaces.

(1). Any metric space (X, d) can be considered as a topological space, letting τ_X be the set of d-open sets in X. The second and third axioms for a topology require checking, and it's worth

doing this explicitly to illustrate why one deals with infinitely many sets and the other with finitely many.

If $U = \bigcup U_i$ is any union of d-open sets and $x \in U$, then x lies in at least one of the U_i's. Because this particular U_i is open, it contains some $B_\epsilon(x)$, which therefore lies inside U. This proves that U is a neighbourhood of x, and therefore (because the argument works for each x) is open.

On the other hand, if $U = \bigcap U_i$ is an intersection of open sets $U_1, U_2, \ldots, U_n$ and $x \in U$ then we can find a collection of balls $B_{\epsilon_1}(x) \subseteq U_1, B_{\epsilon_2}(x) \subseteq U_2, \ldots, B_{\epsilon_1}(x) \subseteq U_1$. The intersection of these balls, which is $B_\epsilon(x)$ where $\epsilon = \min\{\epsilon_1, \epsilon_2, \ldots, \epsilon_n\}$ (a positive number), lies inside each U_i and therefore inside U. So again we see that U is open. Note however that if there were infinitely-many U_i's then their associated ϵ_i's might converge to zero, and the intersection of the balls could be just the set $\{x\}$, which wouldn't have to be open.

Two metrics on a set are said to be *equivalent* if they define the same topologies. (Lipschitz equivalence is a special case.)

(2). Any set has a *discrete topology* in which all subsets are open, and an *indiscrete topology* in which only X and $\emptyset$ are open sets. All functions out of a discrete space, and into an indiscrete one, are continuous.

(3). The collection of subsets $\{\emptyset, \{a\}, \{a, b\}, X\}$ of the set $X = \{a, b, c\}$ is a topology. An n-element finite set has 2^n subsets, and therefore at most 2^{2^n} possible topologies. Finding a formula for the number (or better, the number considered up to automorphisms (permutations) of the set) is, as far as I know, a hard and unsolved (but fairly irrelevant) combinatorial problem.

(4). In algebraic geometry one uses the *Zariski topology*, in which the open sets are the complements of subsets defined by algebraic equations. For example in $\mathbb{C}$, the algebraic subsets are just the zero-sets of polynomials, and therefore just the finite subsets. The open sets of the Zariski topology are the (rather big) sets $\mathbb{C} - \{x_1, x_2, \ldots, x_n\}$, where the x_i's are points of $\mathbb{C}$.

9.3. **Hausdorff spaces.**

Definition 9.3.1. A topological space is said to be *Hausdorff* if, for any pair of distinct points x, y, one can find disjoint open sets U, V containing x, y respectively.

This definition is part of a family of "separation axioms" dealing with whether points and/or open sets can always be "insulated" from one another by means of larger open sets. Hausdorffness (Hausdorffitude?) is the only one worth bothering with here (at all?), for the following simple reason. Any metric space is automatically Hausdorff: if x, y are distinct then $d(x, y) > 0$, and balls of radius $d(x, y)/3$ at x, y are disjoint, by the triangle inequality. In contrast, the Zariski topology described above has no disjoint open sets at all, so it certainly isn't Hausdorff. Topological spaces, then, form a strictly larger class than metric spaces.

9.4. **Homeomorphism.**

Definition 9.4.1. Two topological spaces are *homeomorphic* if there exists a pair of mutually-inverse continuous maps between them.

The term *isomorphic* would be just as good: an isomorphism between mathematical structures (topological spaces, groups, vector spaces, ...) can always be defined as a pair of mutually-inverse "structure-preserving" maps, where structure-preserving is interpreted appropriately: that is, as "homomorphism" in the case of groups, "linear map" in the case of vector spaces, and "continuous map" in the case of topological spaces. The language of *category theory*, which will be explained in the next section, encapsulates this idea neatly.

Example 9.4.2. (1). The open unit interval $(-1, 1)$ and the real line $\mathbb{R}$ are homeomorphic. Just use the map $x \mapsto x/(1 - x^2)$ and its inverse.

(2). Generalising this, we have that the open unit ball Int B^n and the space $\mathbb{R}^n$ are homeomorphic. The map $x \mapsto x/(1 - \|x\|^2)$ is perhaps the nicest choice of homeomorphism.

(3). The map $t \mapsto e^{2\pi it}$ is a continuous bijection between the interval $[0, 1)$ and the circle S^1. However, its inverse is *not* continuous, and therefore the circle is not homeomorphic to an interval (which is just as well, as topology would be boring if it were!)

9.5. **Open maps.**

A function $f : X \to Y$ is called an *open map* if for each open U in X, $f(U)$ is open in Y. (Notice that the requirement here is on the "pushforwards" $f(U)$ of open sets U in X, rather than on the "pullbacks" $f^{-1}(U)$ of open sets U in Y, as in the definition of continuity). Open maps aren't especially important, but they are useful in constructing homeomorphisms. We often have a situation as in (3) above, where we can construct a continuous bijection $f : X \to Y$ and want to know whether its inverse is continuous. The key point is that when f is a bijection, f^{-1} is continuous *if and only if* f *is open* (write down the definitions: they're the same).

The map in example (3) is not an open map because for example the open set $[0, \frac{1}{2})$ does not get sent to an open set. An even simpler example of a continuous but non-open map is $x \mapsto x^2$, mapping $\mathbb{R}$ to itself.

9.6. **Bases.**

It's worth noting that the idea of *generators* for a group has an analogue in the world of topological spaces, and this is sometimes a convenient time-saving device in proofs. Take any collection ρ of subsets of a set X whose union is all of X. It won't in general be a topology, but it is easy to construct a "smallest" topology (one with the fewest open sets) containing all those subsets, as follows. If we close ρ under finite intersections (by adjoining all sets which are intersections of finitely-many elements of ρ), we obtain a larger collection σ which obviously satisfies the third axiom for a topological space, and also contains the empty set. If we now close σ under arbitrary unions (by — what else? — adjoining all unions of elements of σ) we get a collection τ which satisfies the second axiom, contains X as well as $\emptyset$, and (check) still satisfies the third axiom; it is a genuine topology.

Any collection of open sets such as σ which, when closed under unions, generates τ, is called a *base* for τ. Any collection such as ρ, which requires closure under both finite intersections and arbitrary unions to generate τ ,is called a *sub-base*. In a metric space, for example, the collection of all open balls $B_\epsilon(x)$ is a base for the topology. It's easy to see that just the balls of radius $1/n$ (for positive integer n) will do. In $\mathbb{R}^n$, we may actually use balls of radius $1/n$ based at rational points (that is, points whose coordinates are all rational). A space such as $\mathbb{R}^n$ with a countable base is called *second countable*; this property is technically part of the definition of a manifold, which we will see later.

9.7. **Interiors, closures, accumulation points and limits.**

The concept of the *interior* of a subset A of a topological space X is quite a natural one. There are two different formulations of the notion: a local and global one.

Definition 9.7.1. The *interior* Int(A) of a subset A of X is the subset of all points $a \in A$ such that A is a neighbourhood of a.

Exercise 9.7.2. Show that U is an open set if and only if $U = \text{Int}(U)$.

Exercise 9.7.3. Show that $\mathrm{Int}(A)$ is the union of all the open subsets of X contained by A.

A *closed set* in a topological space X is one whose complement is open. This is not a terribly interesting definition, but it does suggest, correctly, that one can could reformulate all statements about topological spaces in terms of closed sets rather than open ones.

Exercise 9.7.4. Give examples of subsets of $\mathbb{R}$ which are open but not closed; closed but not open; both; neither.

Exercise 9.7.5. Show that the intersection of arbitrarily many closed sets is closed, and the union of finitely-many closed sets is closed.

Exercise 9.7.6. Show that a function $f : X \to Y$ between topological spaces is continuous if and only if $f^{-1}(F)$ is closed in X, whenever F is closed in Y.

Exercise 9.7.7. A *closed map* is something which pushes forwards closed sets to closed sets. Show that a continuous bijection which is a closed map is a homeomorphism.

A more interesting characterisation of closed sets comes from considering what the *closure* of a set should be. As with the interior, there are two versions of the definition.

Definition 9.7.8. An *accumulation point* of a set A in a topological space X is any point $x \in X$, each of whose punctured neighbourhoods (things of the form $N - \{x\}$, where N is a neighbourhood of x) contains a point of A. The *closure* $\bar{A}$ of A is the union of A and its set of accumulation points; thus, it is the set of points $x \in X$, each of whose neighbourhoods contains a point of A.

Exercise 9.7.9. What are the accumulation points of the following subsets of $\mathbb{R}$: $\mathbb{Z}$, $\mathbb{Q}$, I, $(0, 1)$?

Exercise 9.7.10. Show that F is a closed set if and only if $F = \bar{F}$.

Exercise 9.7.11. Show that $\bar{A}$ is equal to the intersection of all the closed subsets of X which contain A.

Definition 9.7.12. The *boundary* ∂A of a subset A of X is defined to be its closure minus its interior.

Exercise 9.7.13. Consider the operations of closure, interior, and complement as functions C, I, N which take subsets of $\mathbb{R}$ to subsets of $\mathbb{R}$. Show, by exhibiting a subset A of $\mathbb{R}$ for which $CI(A) \neq IC(A)$, that the operations C and I are not commutative. Prove that $C^2 = C, I^2 = I, N^2 = 1$. Are there any other relations in the semigroup generated by C, I, N? See if you can prove that it is finite, and find its order!

A particularly useful idea in topology is the idea of an *open dense subset* U of a space X: this is an open set whose closure is all of X. This concept typically appears whenever one has some kind of space parametrising geometric configurations, and is intimately tied to the ideas of *genericity* and *stability* of such configurations. As an example, let X be the set of all oriented lines in $\mathbb{R}^2$. It is a 2-dimensional space, in fact homeomorphic to $S^1 \times \mathbb{R}$, because lines can be parametrised by their direction and (signed) distance to the origin. It's easy to check that the set of lines which are not horizontal (parallel to the x-axis) forms an open dense subset: *dense*, because if we have a line parallel to the x-axis, then we can perturb it by an arbitrarily small amount to make it non-parallel; and *open* because if we limit ourselves to small enough perturbations, then non-horizontal lines remain non-horizontal. We could say that horizontality is an *unstable* property, but non-horizontality is *stable*. (Non-horizontal lines are also *generic* in the space of all lines — meaning that a randomly chosen line will be non-horizontal with probability 1 — though here we are using measure-theoretic language which is unwarranted without further work.)

In analysis, the idea of the limit of a sequence of points is just as important as the idea of continuity. It is straightforward to give a definition that works for topological spaces.

Definition 9.7.14. A point x of a topological space X is a *limit point* of the sequence of points $(x_n)_{n\in\mathbb{N}}$ if for any neighbourhood N of x, there exists some integer m such that $x_n \in N$ for all $n \geq m$.

This agrees with the usual definition in a metric space: for example, the limit of the sequence $x_n = 1/n$ in $\mathbb{R}$ is 0, while the sequence $x_n = n$ has no limit. In any Hausdorff space, the limit of a sequence is unique if it exists. In a non-Hausdorff space this need not be the case, and the intuitive picture breaks down. For example, let X be the "real line with a double zero" made by adding a new point $0'$ to $\mathbb{R}$, and adding open sets which are copies of the existing ones containing 0, but with $0'$ inserted in its place. Then the same sequence $1/n$ has two limit points!

Although the idea of limit of a sequence appears rather similar to the idea of accumulation point of a set, they are actually rather different. If you try to prove the theorem that the accumulation points of a set are given by the limit points of sequences of elements of that set, you will find that it can't be done (and isn't true) without additional information about the topological space: you need the existence of a countable base of neighbourhoods for each point (something which is true, for example, for a metric space).

9.8. **Constructing new spaces from old.**

There are four standard constructions in set theory which can be enhanced to constructions with topological spaces.

(1). If $A \subseteq X$ is a subset of a topological space X, it can be given the *subspace topology*

(2). If X, Y are topological spaces then the set $X \times Y$ can be given the *product topology*

(3.) If $\sim$ is an equivalence relation on a topological space X, then the set of equivalence classes $X/\sim$ may be given the *quotient* (or *identification space*) *topology*. An alternative way to describe this situation is simply to consider that we are given a surjective map of sets $q : X \twoheadrightarrow Y$ and a topology on X; we then produce one on Y. (Any equivalence relation defines a surjection $q : X \twoheadrightarrow X/\sim$, and conversely, any surjection q defines an equivalence relation whose equivalence classes are inverse images of points.)

(4). If X, Y are topological spaces then their *disjoint union* $X \amalg Y$ may be made into a topological space. (This might be termed "coproduct" if one were thinking category-theoretically.)

The precise definitions of these new topologies are easy to understand in terms of the following simple principle. In each of the above constructions, there are certain basic "structural maps" relating the old and new sets, and the new topologies should be chosen so as to make these continuous. Here are the structure maps in the four cases:

(1). An injective inclusion map $i : A \hookrightarrow X$.

(2). Two surjective projection maps $\pi_X : X \times Y \to X, \pi_Y : X \times Y \to Y$.

(3). A surjective quotient map $X \to Y$.

(4). Two injective inclusion maps $i_X : X \hookrightarrow X \amalg Y, i_Y : Y \hookrightarrow X \amalg Y$.

Consider case (1). If we are to make i continuous then we certainly require that all sets of the form $i^{-1}(U)$, where U is open in X, are included in the topology on A. In fact these sets form a topology on A (easy check) and so we are done! Notice that we could have used a *larger* topology on A (the cheat's answer would be to give A the discrete topology, making *all* maps out of it continuous!) but that what we have here is the *minimal* topology on A which makes i continuous.

A similar principle does case (2). We are required to put all sets of the form $U \times Y$ and $X \times V$ (where U, V are open sets of X, Y) into the topology on $X \times Y$. For it to be closed under intersections, we must then add all sets of the form $U \times V$'s. Then to be closed under union, we must add all unions of such "box-shaped" sets. The result of this is the product topology; again, it is the minimal possible choice. Notice that sets of the form $U \times V$ form a base for the product topology.

The third and four cases are duals of the first and second, in the sense that the directions of the structure maps are simply reversed, and injections become surjections. (This sort of duality pervades topology and category theory.) In these cases we aren't being *forced* to add sets to the topology to achieve continuity; rather, since the structure maps go *into* our new spaces, we are being *limited* in how much we can add; there is an *upper bound* to the choice of the topology, rather than a lower bound. We can certainly "cheat" again by using the indiscrete topology on the new spaces, but the sensible thing to do is put in the *maximal* topology which will make the structure maps continuous.

In the case of the quotient space we therefore define U to be open in Y if and only if $q^{-1}(U)$ is open in X. For the disjoint union, the open sets are all sets of the form $U \amalg V$, for open sets U, V of X, Y.

Some of the properties of subspaces, products and disjoint unions are covered by the following exercises. (Quotient spaces are sufficiently important to get the next section all to themselves!)

Exercise 9.8.1. 1. Let X be a topological space. Let Y be a subset of X, equipped with the subspace topology, and let A be a subset of Y. Show that if A is closed in Y and Y is closed in X, then A is closed in X. Show that this statement is still true if both "closeds" are replaced by "opens". Give counterexamples for the two "mixed" cases.

Exercise 9.8.2. If A is a subspace of X, and $f : X \to Y$ is continuous, then the restriction of f to A is continuous.

Exercise 9.8.3. Let $i : A \hookrightarrow X$ be the inclusion of a subspace A in a topological space X. Let Z be another space, and $f : Z \to A$ a function. Show that f is continuous if and only if $i \circ f$ is continuous.

Exercise 9.8.4. Suppose that a topological space X is written as the union of finitely-many closed sets F_i, and that we are given functions $f_i : F_i \to Y$ (to some other space Y) which agree on the overlaps $F_i \cap F_j$. Prove that the function $f : X \to Y$ defined piecewise by the f_i's is a continuous function if and only of the individual f_i's are continuous. (This "gluing lemma" is very handy when dealing with piecewise-defined functions.) What is the corresponding statement when we decompose X into open sets U_i?

Exercise 9.8.5. The product of topological spaces $\mathbb{R} \times \mathbb{R}$ is homeomorphic to the space $\mathbb{R}^2$ (with topology coming from the Euclidean metric).

Exercise 9.8.6. The diagonal map $X \to X \times X$ is always a continuous function, for any space X.

Exercise 9.8.7. A function $f : Z \to X \times Y$ is continuous if and only if its coordinate functions $\pi_X \circ f, \pi_Y \circ f$ are both continuous.

Exercise 9.8.8. Suppose X, Y are topological spaces and $A \subseteq X, B \subseteq Y$ are subsets. Show that the two ways of topologising $A \times B$ (as a subspace of a product, or as a product of subspaces) are homeomorphic.

Exercise 9.8.9. Suppose X, Y are topological spaces and $A \subseteq X, B \subseteq Y$ are subsets. Show that the following Leibniz rule holds:

$$\partial(A \times B) = (\partial A \times \bar{B}) \cup (\bar{A} \times \partial B).$$

Exercise 9.8.10. Suppose $\{X_i\}_{i\in I}$ is an infinite family of topological spaces. What is the correct definition of the product topology on the product of sets $\prod X_i$?

Exercise 9.8.11. A function $X \amalg Y \to Z$ is continuous if and only if its restrictions to X and Y are continuous.

9.9. **Quotient spaces.**

Quotient spaces are very useful in topology. Recall the definition given above: if X is a space and $\sim$ is an equivalence relation on X, then the set $Y = X/\sim$ of equivalence classes becomes a space: if $q : X \to X/\sim$ is the natural map taking a point x to its equivalence class $[x]$, then the open sets are those U such that $q^{-1}(U)$ is open in X.

Let's work out an example. Consider the equivalence relation on $\mathbb{R}$ given by $x \sim y$ if $x - y$ is an integer. The equivalence classes are *lattices* of the form $[x] = x + \mathbb{Z}$; they are subsets which are translates of the subset $\mathbb{Z}$ of integers.

The quotient space is meant to parametrise such lattices, in the sense that each point of the quotient "is" a lattice, and two points are "close" in the quotient topology if their lattices are "close" inside $\mathbb{R}$. If one starts with the lattice $\mathbb{Z}$ and pushes it gradually to the right, it returns to its initial position (as a subset, not pointwise) after moving a distance of one unit, having in the process taken on every possible position of a lattice in $\mathbb{R}$. This makes it intuitively clear that the quotient space is a *circle*; let us prove this rigorously, as an example of how to work with quotient spaces.

Let $\mathbb{R}/\mathbb{Z}$ denote the quotient space and $q : \mathbb{R} \to \mathbb{R}/\mathbb{Z}$ be the quotient map $x \mapsto [x] = x + \mathbb{Z}$. Let S^1 be the standard circle of unit complex numbers, equipped with the subspace topology from $\mathbb{C}$. To prove that $\mathbb{R}/\mathbb{Z}$ is homeomorphic to S^1 we need to construct a map $f : \mathbb{R}/\mathbb{Z} \to S^1$ which is a bijection, continuous, and has a continuous inverse.

Simply at the level of *sets* (ignoring continuity), there is a bijective correspondence between the set of functions $f : \mathbb{R}/\mathbb{Z} \to S^1$ and the set of functions $\tilde{f} : \mathbb{R} \to S^1$ having the property that $\tilde{f}(x) = \tilde{f}(y)$ whenever $x \sim y$. The correspondence is given by $\tilde{f} \leftrightarrow f \circ q$, or more explicitly by the formula $f([x]) = \tilde{f}(x)$. (If we're given $\tilde{f}$, this formula can be taken as the *definition* of the corresponding f; we see that the property "$\tilde{f}(x) = \tilde{f}(y)$ whenever $x \sim y$" is precisely what's needed to make $f([x])$ *well-defined*, that is, independent of the choice of element x representing the equivalence class $[x]$. When we put back the topology, there is nothing to worry about: the definition of the quotient topology ensures that under this correspondence, *f is continuous if and only if $\tilde{f}$ is.*

We therefore define $\tilde{f} : \mathbb{R} \to S^1$ to be the map $x \mapsto e^{2\pi i x}$. This satisfies $\tilde{f}(x) = \tilde{f}(y)$ whenever $x - y$ is an integer, and it's obviously continuous. So it induces a well-defined continuous map $f : \mathbb{R}/\mathbb{Z} \to S^1$ by the formula $[x] \mapsto e^{2\pi i x}$.

The map f is surjective because $\tilde{f} = f \circ q$ is. It's easy to check that f is also an injection: if $[x]$ and $[y]$ are two points of $\mathbb{R}/\mathbb{Z}$ such that $f([x]) = f([y])$, then $e^{2\pi i x} = e^{2\pi i y}$, x and y must differ by an integer, and we see that actually $[x] = [y]$.

All that remains is to check that f has a continuous inverse, which we do by showing that it's an open map. (This neat trick avoids having to actually write down the inverse of f, which could be messy and confusing). If U is open in $\mathbb{R}/\mathbb{Z}$ then $f(U)$ can also be written as $f \circ q(q^{-1}(U)) = \tilde{f}(q^{-1}(U))$. Since $q^{-1}(U)$ is open in $\mathbb{R}$, by definition of the quotient topology, and $\tilde{f}$ is an open map (by inspection), $f(U)$ is open, and we're done.

(An alternative trick that's sometimes useful at the end of proofs like this is the "compact space to Hausdorff space" property, number (9) in an exercise in the next section.)

Here is a more sophisticated view of the above example. We say that a group G *acts* on a set X if we are given a homomorphism $\rho : G \to \mathrm{Aut}(X)$. Thus, we associate to each group element $g \in G$ an invertible function $\rho(g) : X \to X$ such that $\rho(gh) = \rho(g) \circ \rho(h)$ and $\rho(1) = \mathrm{id}_X$. Instead of writing $\rho(g)(x)$ for the point to which g's function carries x, we usually just call it gx. In the above example, the group $\mathbb{Z}$ is acting on $\mathbb{R}$; the element n acts by "translation of $\mathbb{R}$ by n".

(In topology we usually want to talk about *continuous* actions of *topological* groups on *topological spaces*, but this is irrelevant for now.)

When a group G acts on a set X, there is a natural equivalence relation on X: define $x \sim y$ if there is an element g such that $gx = y$. The equivalence classes are called *orbits*, and the set of orbits $X/\sim$ is usually written X/G. If X is a topological space (not just a set) then X/G becomes the *space* of orbits via the quotient topology. (This explains the reason for the notation $\mathbb{R}/\mathbb{Z}$ above.)

As a further example, the group $\mathbb{Z}^2$ acts on $\mathbb{R}^2$ by integer translations, and the quotient is the 2-torus $S^1 \times S^1$. (The proof is as in example (1), we would take $\tilde{f} : \mathbb{R}^2 \to S^1 \times S^1$ given by $(x, y) \mapsto (e^{2\pi ix}, e^{2\pi iy})$.) Similarly $\mathbb{Z}^n$ acts on $\mathbb{R}^n$ via translations, and we get the n-torus $T^n = (S^1)^n$.

Quotient spaces arising from group actions are perhaps the nicest "naturally occurring" examples. But the most common kind of quotient in topology is somehow less sophisticated: we simply want to glue together or identify various existing spaces in some way to make a new one, and we define an equivalence relation to achieve this.

For example, consider the unit square $X = I \times I$ with the top edge glued to the bottom and the left edge glued to the right: define $(x, 0) \sim (x, 1)$ for each x and $(0, y) \sim (1, y)$ for each y. The equivalence class of a point in $I \times I$ then contains one, two or four elements according to whether the point is in the interior of the square, interior of an edge, or is one of the corners. You can imagine $X/\sim$ as being like the square, except that you can "go off one side and come back on the opposite side" as in the videogame "Asteroids". Alternatively you can imagine actually pasting the edges of a (stretchy, rubbery) square together: after gluing one pair we'd have a cylinder, and after gluing the remaining pair a torus. We conclude that $X/\sim$ (like the universe in "Asteroids") is a torus.

To actually prove that $X/\sim \cong S^1 \times S^1$ we follow the same method as before. Define a map $\tilde{f} : I \times I \to S^1 \times S^1$ by $(x, y) \mapsto (e^{2\pi ix}, e^{2\pi iy})$; check that it respects the equivalence relation in the right way to induce a map $f : X/\sim \to S^1 \times S^1$; check that this is a bijection, and then show its inverse is continuous.

Here are some further examples of quotient spaces. It would take a lot of effort to describe each of these in the detail it deserves (and also lots pictures, which I am too lazy to do right now) so I will give up and just give you the idea.

(1). Take $I \times I$ and identify $(0, y)$ with $(1, 1 - y)$ for each y. This gives the *Möbius strip*.

(2). Take $I \times I$ and identify $(0, y) \sim (1, 1 - y)$ for each y and also $(x, 0) \sim (x, 1)$ for each x; this gives the *Klein bottle*.

(3). Take S^2 and identify antipodal points. (Or, take the quotient of S^2 by the group Z_2, whose non-trivial element acts via the antipodal map $x \mapsto -x$). The result is the *projective plane* $\mathbb{R}P^2$.

(4). Take a regular octagon and identify its opposite edges in pairs (make them correspond via a translation) just as we did for a torus. The result is a *closed orientable surface of genus* 2, which looks like a torus with two "holes". In fact all connected closed 2-manifolds can be obtained by gluing pairs of edges of polygons in this way.

(5). There is a huge class of spaces which can be built by merely gluing together collections of balls B^n (of varying dimension). These spaces are called *CW complexes* and their behaviour is so nice that they are the main class of spaces with which people actually work in algebraic topology. (Arbitrary topological spaces can be very "pathological", and one often needs to make additional assumptions in order to make much progress understanding their topology.)

(7). *Real projective n-space* $\mathbb{R}P^n$ is the space of lines through the origin in $\mathbb{R}^{n+1}$. To define it we consider the equivalence relation on $\mathbb{R}^{n+1} - \{0\}$ given by $x \sim y$ if x is a non-zero scalar multiple of y. The equivalence classes are lines through the origin with their zero points missing (if we didn't remove zero, everything would be equivalent to everything else) and so the quotient space is a space whose points correspond to lines through the origin, as required. We could express the same space more nicely as the quotient of $\mathbb{R}^{n+1} - \{0\}$ by the action of the multiplicative group $\mathbb{R}^* = \mathbb{R} - \{0\}$ of dilations.

A slight variation on this gives a different construction of the sphere S^n. If we take $\mathbb{R}^{n+1} - \{0\}$ modulo the action of the multiplicative group $\mathbb{R}_{>0}$ of *positive* dilations, the orbits are half-lines (or *rays*) emanating from (though once more not including) the origin, so the quotient space is the "celestial sphere" (of directions from which light rays can approach the observer at zero) of *directions* in $\mathbb{R}^{n+1}$.

Another variation is to repeat the process with the *complex* vector space $\mathbb{C}^{n+1}$ and the multiplicative group $\mathbb{C}^*$ to give $\mathbb{C}P^n$, the complex projective space. These spaces are of fundamental importance in algebraic geometry.

Exercise 9.9.1. Prove that the product of two Hausdorff space is Hausdorff, and that a subspace of a Hausdorff space is Hausdorff. Give an example to show that a quotient of a Hausdorff space need not be Hausdorff.

Exercise 9.9.2. The group $\mathbb{Z}$ acts on the space $\mathbb{C}^2 - \{0\}$ as a group of dilations, according to the formula

$$n.(z, w)) = (2^n z, 2^n w).$$

Show that the space of orbits is homeomorphic to the space $S^1 \times S^3$.

9.10. **Compactness.**

In the study of complex vector spaces, one learns that the correct notion of "finiteness" is actually *finite-dimensionality* rather than finiteness as a set. (Only the 0-dimensional space is actually a finite set!) Similarly, when studying groups, rings or modules (vector spaces being a special case), the notion of being *finitely-generated* is useful, in the sense that such "finite" objects have nice properties not shared by their "infinite"relations.

Compactness is in some sense the topological space analogue of being finitely-generated. We will see that for subspaces of Euclidean spaces $\mathbb{R}^n$, it is exactly the same as being closed and bounded, and that intuition about the properties of such sets can often be carried into the general case.

Definition 9.10.1. An *open cover* of a space X is a family of open sets $\{U_i\}_{i \in I}$ whose union is X. A *subcover* of such a cover is any subcollection $\{U_j\}_{j \in J}$, where $J \subseteq I$, whose union is still all of X. A space is *compact* if every open cover has a finite subcover.

Note immediately that this is a *topological property*: that changing a space by a homeomorphism preserves its compactness (or lack of it).

To say that a *subset* A of a space X is compact means that A, when given the subspace topology (and viewed as a space in its own right), is compact in the above sense. Equivalently, we can define an open cover of A to be a family of open sets of X whose union contains A; then A is compact if each of its open covers has a finite subcover.

Example 9.10.2. The open unit ball in any $\mathbb{R}^n$ is not compact, because the family of open balls of radius $1 - 1/k$ (for integers $k \geq 2$) forms a cover, any of whose finite subcovers has a largest element with radius strictly less than 1.

Theorem 9.10.3 (Heine-Borel)**.** *The (closed) unit interval I is compact.*

Proof. This is a standard proof by contradiction. Suppose we have a cover with no finite subcover. By restriction, it gives a cover of each of the two half-intervals $[0, \frac{1}{2}]$ and $[\frac{1}{2}, 1]$. At least one of these covers cannot have a finite subcover, or we could combine the two to get a finite subcover of I. So pick this half-interval (if neither has a finite subcover, just choose one of them), and repeat the argument. We obtain a nested sequence of closed intervals of successively halving length whose left endpoints form a bounded, monotonically increasing sequence, and whose right endpoints form a bounded, monotonically decreasing sequence. Each of these sequences therefore has a limit (it supremum/infimum - that these exist is the defining property of the real numbers) but because of the halving lengths they must agree. We obtain therefore that the intersection of the family of halving intervals is a single point $x \in I$. This x must lie in at least one of the sets of the cover, so choose one: because the set is open, it extends a positive distance to either side of x, and must therefore eventually contain the tiny halving intervals of the sequence, contradicting the fact that they do not have finite subcovers. □

Exercise 9.10.4. Prove the following sequence of statements. (At each stage, you can assume the previous ones.)

(1). The product of two compact spaces is compact. (It follows by induction that any finite product of compact spaces is compact, but in fact an *arbitrary* (infinite) product of compact spaces is compact; this is *Tychonoff's theorem.*)

(2). A closed subspace of a compact space is compact.

(3). A closed, bounded subspace of $\mathbb{R}^n$ is compact.

(4). A quotient space of a compact space is compact.

(5). A compact subspace of a Hausdorff space is closed.

(6). A continuous real-valued function on a compact space is bounded and attains its bounds.

(7). A compact subspace of $\mathbb{R}^n$ is closed and bounded.

(8). The union of finitely many compact sets is compact. (Is this true for finite intersections?)

(9). A continuous map from a compact space to a Hausdorff space is a closed map; and consequently that a continuous bijection from a compact space to a Hausdorff space is a homeomorphism.

Exercise 9.10.5 (The *Lebesgue lemma*)**.** Given any open cover $\{U_i\}$ of a compact metric space, there is some constant $\delta > 0$ such that for any set of diameter less than δ, one can find one of the U_i's which contains it. (The *diameter* of a subset of a metric space is the supremum of the set of pairwise distances between its points.)

Exercise 9.10.6. An infinite subset of a compact space must have an accumulation point. Deduce that a discrete subset (one which, as a subspace, inherits the discrete topology) of a compact space must be finite.

9.11. **Mapping spaces.**

There is one simple way to create "new spaces from old" which was omitted from the earlier list: the *mapping space* construction. If X and Y are topological spaces we can consider the set $\mathrm{Map}(X, Y)$ of all continuous functions $f : X \to Y$. It would be very nice to have a topology on

this, so that we could talk about functions being "close", sequences of functions converging, and so on. How can we do this?

If X is a compact space and Y is a metric space then there is a very straightforward way to make $\mathrm{Map}(X, Y)$ into a metric space: we can define

$$d(f, g) = \sup_{x \in X} d_Y(f(x), g(x)),$$

because the compactness of X guarantees that the supremum exists. Clearly, functions are considered "close" if their values are close *everywhere*.

Given arbitrary topological spaces X, Y we can define a topology on $\mathrm{Map}(X, Y)$ as follows. For each compact subset K of X and open subset U of Y, let $S_{K,U}$ denote the set of all functions $f : X \to Y$ such that $f(K) \subseteq U$. We can take these sets $S_{K,U}$ as a *sub-base* for a topology which is called the *compact-open* topology. It's not hard to check that it agrees with the metric topology in the special case of the previous paragraph.

It's a little hard to get a feel for this topology in general, but a key fact is that for three spaces X, Y, Z (where Y is locally compact) the natural bijection between set-theoretic functions induces a bijection between the spaces of *continuous* functions

$$\mathrm{Map}(X \times Y, Z) \leftrightarrow \mathrm{Map}(X, \mathrm{Map}(Y, Z));$$

that is, a function $X \to \mathrm{Map}(Y, Z)$ is continuous if and only if its "adjoint" $X \times Y \to Z$ is continuous. With a few more assumptions on the spaces (locally compact, Hausdorff? - I need to look this up!) this "exponential law" (try rewriting it using the notation Z^Y instead of $\mathrm{Map}(Y, Z)$!) is actually a *homeomorphism*.

9.12. **Connectedness and path-connectedness.** There are two sensible notions of connectedness in common use. The more obvious one, *path-connectedness*, measures whether any two points of a space may be joined by a continuous path (a continuous image of the unit interval). Thus it studies the space using continuous maps *into* it. It is very common in topology to find a duality between "maps in" and "maps out" notions, and this is no exception; the notion of *connectedness* measures whether there exist continuous maps out of a space onto a discrete space with two points.

Definition 9.12.1. A space is *path-connected* if for each pair of points $x, y \in X$, it is possible to find a *path* joining x and y, that is a continuous map $\gamma : I \to X$ such that $\gamma(0) = x, \gamma(1) = y$. I will write such a path as $\gamma : x \to y$, hoping it will not be cause confusion with the notation for functions.

Example 9.12.2. Any Euclidean space $\mathbb{R}^n$ is path-connected; the formula for the straight-line path γ joining two vectors x, y is the "weighted linear combination"

$$\gamma(t) = (1 - t)x + ty \quad 0 \leq t \leq 1,$$

which will come in handy on many occasions. On the other hand, the space $\mathbb{R} - \{0\}$ is not path-connected. This follows from the intermediate value theorem of basic real analysis: any continuous map $I \to \mathbb{R}$ whose initial value is negative and whose final value is positive must take the value 0; therefore there are no continuous paths connecting any negative real with any positive one.

The property of path-connectedness is, like compactness, a topological property: homeomorphism preserves path-connectedness (or lack of it).

Corollary 9.12.3. *The spaces $\mathbb{R}$ and $\mathbb{R}^2$ are not homeomorphic.*

Proof. If $f : \mathbb{R} \to \mathbb{R}^2$ were a homeomorphism, then f would restrict to a homeomorphism between $\mathbb{R} - \{0\}$ and $\mathbb{R}^2 - \{f(0)\}$. However, the former is not path-connected, whereas the latter one is (easy check), so no such homeomorphism exists. □

Unfortunately this idea does not generalise to distinguish (meaning, to prove non-homeomorphic) Euclidean spaces of dimension greater than 1 from one another. While it is indeed true that $\mathbb{R}^m \not\simeq \mathbb{R}^m$ for distinct m, n (this fact being called "Brouwer's invariance of domain"), we need more subtle algebraic-topological tools to show this.

Let us consider now the other notion of connectedness.

Definition 9.12.4. A space is *disconnected* if it is possible to write it as the union of two non-empty disjoint open sets. Therefore, a space X is *connected* if, whenever X is written as a union $X = U_1 \cup U_2$ of disjoint open sets U_i, one of them must be empty. Equivalently, X is connected if the only subsets which are both open and closed are X and $\emptyset$.

This definition is the standard one, but it can be immediately rephrased in a more illuminating way as follows.

Lemma 9.12.5. *A space X is disconnected if and only if there exists a continuous surjection $X \to T$, where $T = \{0, 1\}$ denotes the two-point space with the discrete topology. So a space is connected if and only if there is no such surjection.*

Proof. The correspondence should be clear: a map $X \to T$ is continuous precisely when the preimages $U_0 = f^{-1}(0), U_1 = f^{-1}(1)$ are open, and is a surjection precisely when they are both non-empty. Thus existence of such a function disconnects a space, and conversely a disconnected space has such a function. □

The most common relations between path-connectedness and connectedness are summarised by the following statements.

Lemma 9.12.6. *A path-connected space X is connected.*

Proof. A continuous function $X \to T$ must be constant along the image of any continuous path $\gamma : I \to X$, by the intermediate value theorem. So for a path-connected space it takes the same value everywhere, and therefore cannot be surjective. □

Example 9.12.7. A connected space need not be path-connected. An example is the subspace X of $\mathbb{R}^2$ consisting of the graph of the function $\sin(1/x)$, for $x \neq 0$, together with a single point at the origin.

To show it is connected: clearly each side of the graph is a path-connected, hence connected, set. So any continuous function $f : X \to T$ must be constant for $x > 0$ and for $x < 0$. But since there are sequences of points on each side converging to the origin, and by continuity the value $f(0)$ is the limit of each sequence, the two values are equal, and the function f is constant on X. Therefore X is connected.

To show it is not path-connected: suppose there were a path $\gamma : I \to X$ joining the points with x-coordinates ± 1. Let π be the projection from $\mathbb{R}^2$ onto the x-axis. Then $\pi\gamma(I)$ is a subset of the x-axis containing the interval $[-1, 1]$, because of the intermediate value theorem and the fact that it is a path-connected set. Since π restricted to X is injective, $\gamma(I)$ contains all points of X with x-coordinates in $[-1, 1]$. However, this is not a closed subset of $\mathbb{R}^2$, whereas $\gamma(I)$, which is a continuous image of a compact space, must be. This contradiction finishes the proof.

Exercise 9.12.8. A connected subset of a disconnected space $X = U_1 \cup U_2$ lies either completely inside U_1, or completely inside U_2.

Exercise 9.12.9. The product of connected spaces is connected.

Exercise 9.12.10. A quotient of a connected space is connected.

Exercise 9.12.11. If two connected sets intersect non-trivially, then their union is connected.

Exercise 9.12.12. Which of the above statements is true when path-connected replaces connected?

The relation of being joined by a path is easily seen to be an *equivalence relation*: it is reflexive (the constant path $\gamma(t) = x$ joins x to x!), symmetric (define $\gamma^{-1}(t) = \gamma(1-t)$ to turn a path $\gamma : x \to y$ into $\gamma^{-1} : y \to x$) and, most importantly, transitive: compose paths $\gamma : x \to y$ and $\delta : y \to z$ using the formula below (when composing paths, it seems madness not to write them from left to right, as I do here):

$$(\gamma.\delta)(t) = \begin{cases} \gamma(2t) & 0 \leq t \leq \frac{1}{2} \\ \delta(2t-1) & \frac{1}{2} \leq t \leq 1. \end{cases}$$

Definition 9.12.13. The *set of path-components* of a space X, denoted $\pi_0(X)$, is the set of equivalence classes under the above relation. Note that there is a map $X \to \pi_0(X)$ sending each point to its equivalence-class, or *path-component*.

Example 9.12.14. An open subset U of $\mathbb{R}^n$ which is connected is path-connected.

Proof. Suppose y is a point in the path-component of x. Since U is open, we may find a small ball $B_\epsilon(y)$ contained in U. Each point of this ball is joined to y by a linear path, and therefore the whole ball lies in the path-component of x; we've therefore shown that any path-component of U is open. Now observe that the fact that the path-components partition U, together with the fact that U is connected, shows that there can be only one. □

(This proof would work more generally inside a space X which, like $\mathbb{R}^n$, is *locally path-connected*, in that every point possesses some neighbourhood which is path-connected.)

It is possible to define *connected components* (or simply *components*) of a space. Like path-components, they partition the space.

Definition 9.12.15. The *connected component* of a point x of X is the union of all connected subsets of X which contain x.

Exercise 9.12.16. If $\{C_i\}$ is any family of connected sets whose intersection is non-empty, then the union $\bigcup C_i$ is connected. Hence "connected components" are (as the name certainly suggests, though in mathematics this kind of logical reasoning can be disastrous) themselves connected!

Exercise 9.12.17. The closure of a connected set is connected. Hence components are closed.

Example 9.12.18. Let X be the subspace of $\mathbb{R}$ consisting of the numbers $\{1/n\}$ (for all $n \in \mathbb{N}$) together with the point $\{0\}$. Each point $\{1/n\}$ is its own component, and therefore (since components partition) so is $\{0\}$. However, $\{0\}$ is not open: this shows that components, though always closed, do *not* have to be open. (However, in a space which is *locally connected* – every point possessing a neighbourhood which is connected – components are both open and closed.)

Although the notion of path-connectedness is not very subtle or complicated, it lies at the heart of algebraic topology — and for two distinct reasons. Firstly, we will see later how the subject revolves around the families of topological invariants known as *homotopy groups* and *cohomology*

groups. There are various possible ways to define these invariants for a space X, but perhaps the slickest (at least conceptually) is in terms of the (path-)connectedness of certain "auxiliary" spaces built naturally out of X. Homotopy classes of maps $X \to Y$, for example, are just the path-components of the mapping space $\mathrm{Map}(X, Y)$; the homotopy groups of Y result when take X to be a sphere.

Secondly, all these invariants are *functors*: they are "machines" which convert topological information (spaces and continuous maps between them) into algebraic information (usually groups, and homomorphisms between groups). The idea of a functor will be explained properly in the next section, but it is worth illustrating it here.

Recall that the set $\pi_0(X)$ is the quotient of X by the equivalence relation which identifies all pairs of points joined by paths. It's easy to see that a continuous map $f : X \to Y$ induces a well-defined map $\pi_0(f) : \pi_0(X) \to \pi_0(Y)$. So we can view π_0 as a machine which converts topological spaces into sets, and maps between spaces into functions between those sets.

Moreover, we can see that $\pi_0(\mathrm{id}_X) = \mathrm{id}_{\pi_0(X)}$ for any X, and that if $g : Y \to Z$ is also continuous, then $\pi_0(g \circ f) = \pi_0(g) \circ \pi_0(f)$. If we pretend that f, g are elements of a group, in which composition is the multiplication, and that $\pi_0(f), \pi_0(g))$ are elements of a different "group", again with composition as multiplication, then π_0 looks like a homomorphism. Such a "homomorphism" is actually called a *functor* from the *category of topological spaces* to the *category of sets*.

Exercise 9.12.19. If $f : X \to Y$ is a homeomorphism then $\pi_0(f)$ is a bijection between $\pi_0(X)$ and $\pi_0(Y)$.

A "dual" approach would be to define $H^0(X)$ to be the set of (not necessarily continuous) functions $X \to \mathbb{Z}$ which are constant along paths in X. The natural operation of pointwise addition of functions turns $H^0(X)$ into a group, and fairly clearly it is just the set of integer-valued functions on the set path-components of X (which is why it is in some sense "dual" to $\pi_0(X)$.) For example, if X is path-connected then $H^0(X) \cong \mathbb{Z}$. Now given a continuous function $f : X \to Y$, we can compose the functions in H^0 with it so as to obtain an induced map $H^0(f) : H^0(Y) \to H^0(X)$; notice that this map goes *the wrong way*, because functions $Y \to \mathbb{Z}$ are *pulled back* (pre-composed with f) to functions on X. The map $H^0(f)$ is obviously a homomorphism of groups. Because of the wrong-wayness, H^0 is an example of a *contravariant functor* from spaces to sets (wherease something like π_0 is said to be *covariant*.)

In the next section we will describe the *fundamental group* π_1 and *higher homotopy groups* π_n of a space. They are in a sense just special cases of π_0. Similarly, we will go on to study *singular cohomology groups* H^n of a space. These generalise the zeroth group H^0 above. All these operations π_n, H^n are functors, satisfying laws like the above. The π_n are *covariant* whereas the H^n are *contravariant*.

There are in fact various different kinds of cohomology we can define for spaces. *Čech cohomology* generalises a slightly different definition of $H^0(X)$ as the abelian group of *locally constant* functions $X \to \mathbb{Z}$. (A locally constant function on X is one such that each point possesses a neighbourhood on which the function is constant; when $\mathbb{Z}$ is the target, these are actually the same as continuous functions $X \to \mathbb{Z}$.) If X is locally connected then this group is the same as the group of $\mathbb{Z}$-valued functions on the set of components of X; if X is also locally path-connected then it agrees with the singular cohomology group $H^0(X)$ we defined above. The higher Čech cohomology also agree with the singular cohomology groups when the space X is nice enough (for example, when it is a CW-complex) but otherwise measure slightly different topological information.