KNOTS KNOTES

JUSTIN ROBERTS

Contents

Date: March 15th 1999. Lecture notes from Edinburgh course Maths 415.

1. Motivation, basic definitions and questions

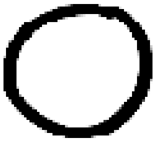
This section just attempts to give an outline of what is ahead: the objects of study, the natural questions (and some of their answers), some of the basic definitions and properties, and many examples of knots.

1.1. Basic definitions.

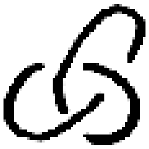
Definition 1.1.1 (Provisional). A *knot* is a closed loop of string in $\mathbb{R}^3$; two knots are *equivalent* (the symbol $\cong$ is used) if one can be wiggled around, stretched, tangled and untangled until it coincides with the other. Cutting and rejoining is *not* allowed.

Example 1.1.2.

unknot



left trefoil



right trefoil

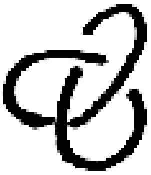
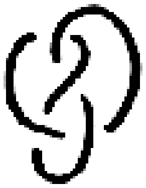
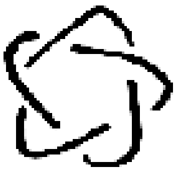


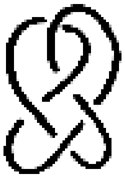
figure-eight



5_1



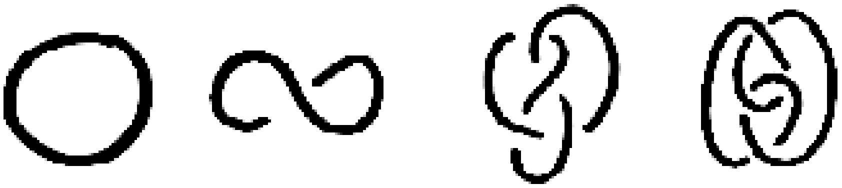
5_2



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 two pictures of the unknot and two of the figure-eight knot. (Convince yourself of the latter using string or careful redrawing of pictures!)



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*).

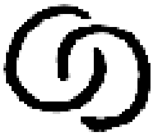
Question 1.2.3. Can one produce a table of the simplest knot types (a *knot type* means an equivalence class of knots, in other words a *topological* as opposed to *geometrical* knot: often we will simply call it "a knot"). "Simplest" is clearly something we will need to define: how should one measure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

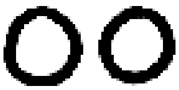
Definition 1.2.4. A *link* is simply a collection of (finitely-many) disjoint closed loops of string in $\mathbb{R}^3$; each loop is called a *component* of the link. Equivalence is defined in the obvious way. A knot is therefore just a one-component link.

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.

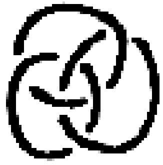
Hopf link



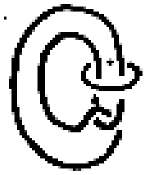
an unlink



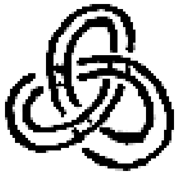
Borromean rings



Whitehead link



doubled trefoil



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.

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.

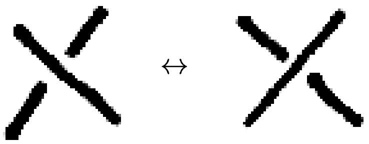
Exercise 1.2.9. Prove that there are no knots with crossing number 1 or 2 (just draw the possible diagrams and check).

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.



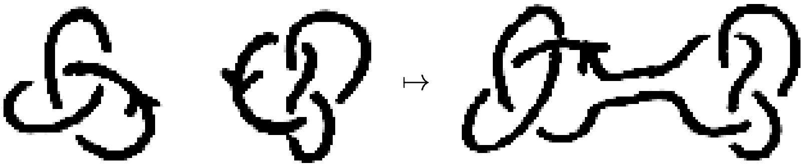
This is evident if one considers reflecting in the plane of the page.

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).

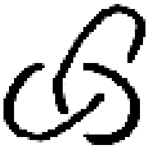
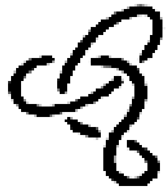
Definition 1.3.5. If K_1, K_2 are *oriented* knots, one may form their *connect-sum* $K_1 \# K_2$ by removing a little arc from each and splicing up the ends to get a single component, making sure the orientations glue to get a consistent orientation on the result. (If the knots aren't oriented, there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

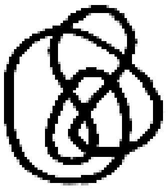
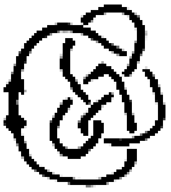
1.4. **Alternating knots.**

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  ↦ alternating diagram 

Exercise 1.4.3. Why does this never give a contradiction when one returns to a crossing for the second time?

If one carries this out it seems that the results "really are knotted". In fact 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.

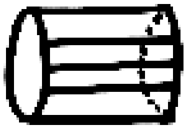
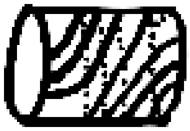
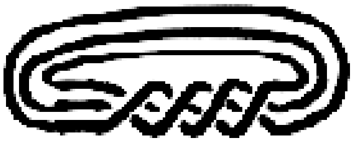
1.5. **Unknotting number.** If one repeats the "random knot" construction above but puts in the crossings so that the first time one reaches any given crossing one goes *over* (one will eventually come back to it on the underpass), one produces mainly unknots. In fact there is always a way of assigning 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 $p > 0, q$, 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.

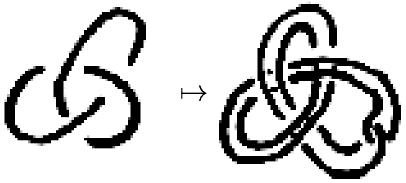
cylinder  $\mapsto$ twisted up  $\mapsto$ $T_{3,4}$ 

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

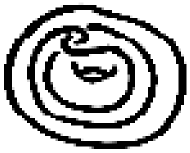
Exercise 1.6.2. How many components does the torus link $T_{p,q}$ have? Show in particular that it is a knot if and only if p, q are coprime.

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.

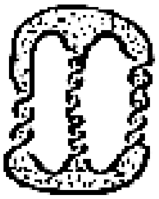
Example 1.6.4. Any knot may be *Whitehead doubled*: one replaces the knot by two parallel copies (there is a degree of freedom in how many times one twists around the other) and then adds a "clasp" to join the resulting two components together (in a non-unravelling way!).



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 top and bottom).



Exercise 1.6.7. How many components does a pretzel link have? In particular, when is it a knot?

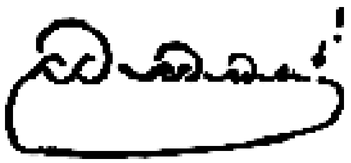
1.7. **Methods.** There are three main kinds of method which ware used to study knots. *Algebraic* methods are those coming from the theory of the fundamental group, algebraic topology, and so on (see section 7). *Geometric* methods are those coming from arguments that are essentially nothing more than 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!)

2. Formal definitions and Reidemeister moves

2.1. **Knots and equivalence.** How should we formulate the notion of deformation of a knot?

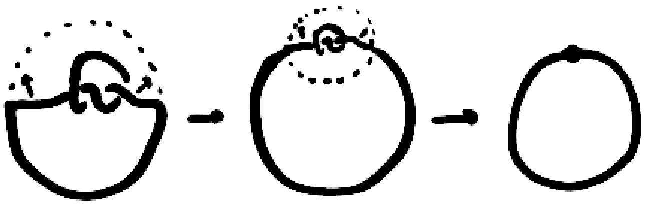
If you studied basic topology you will be familiar with the notion of *homotopy*. We could consider (continuous) maps $S^1 \to \mathbb{R}^3$ as our knots. Two such maps $f_0, f_1 : S^1 \to \mathbb{R}^3$ are called *homotopic* if there exists a (continuous) map $F : S^1 \times I \to \mathbb{R}^3$ with F restricting to f_0, f_1 on $S^1 \times \{0\}$, $S^1 \times \{1\}$. This is obviously no good as a definition, as *all* such maps are homotopic – the string is allowed to pass through itself! (Also, it might intersect itself to start with - we didn't say that the f's should be injective!)

We can solve these problems if we also consider only *injective* maps $f : S^1 \to \mathbb{R}^3$, and require of F that each $f_t = F|_{S^1 \times \{t\}}$ is injective: this relation is called *isotopy.* Unfortunately, this is *not* a very good definition. Firstly, it allows "wild" knots like the one below, which really are continuous (compare with $x \sin(\frac{1}{x})$!) but have infinitely complicated knotting that we can't hope to understand well.



Worse, all knots turn out to be isotopic, albeit for a more subtle reason than they are all homotopic:

Exercise 2.1.1. Check that "gradually pulling the string tight" (see picture below) so that the knot shrinks to a point is a valid isotopy between any knot and the unknot, so this is also no good!



An alternative method is to forget about functions $S^1 \to \mathbb{R}^3$ and just think of a knot as a *subspace* of $\mathbb{R}^3$ which is homeomorphic to the circle; two such knots are *ambient isotopic* if there exists an (orientation-preserving) homeomorphism $\mathbb{R}^3 \to \mathbb{R}^3$ carrying one to the other. This definition works (not all knots are equivalent to one another), but it still doesn't rule out wild knots: the best way of doing this is to declare that all knots should be *polygonal* subspaces of $\mathbb{R}^3$, with *finitely many edges*, thereby ruling out the kind of wildness pictured above.

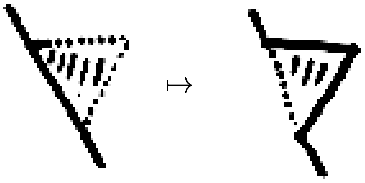
But in practice, once we have decided that a knot should be a knotted polygon, we might as well go the whole hog and use a similar polygonal notion of equivalence, as below. This approach makes the whole subject a lot simpler technically. While we always consider knots to be polygonal, we may as well carry on thinking about and drawing them smoothly, because any smooth (non-wild) knot can always be approximated by a polygonal one with very many short edges.

Definition 2.1.2. A *knot* is a subset of $\mathbb{R}^3$, homeomorphic to the circle S^1, and expressible as a disjoint union of finitely-many points (vertices) and open straight arc-segments (edges).

Remark 2.1.3. The definition really gives a knotted polygon which doesn't intersect itself. For example, the closure of each (open) edge contains exactly two vertices.

Definition 2.1.4. Suppose a closed triangle in $\mathbb{R}^3$ meets a knot K in exactly one of its sides. Then we may replace K by "sliding part of it across the triangle" to obtain a new knot K'. Such a move, or its reverse, is called a (*Delta-*)Δ*-move.* It is clearly the simplest kind of polygonal "wiggle" that we should allow.

Definition 2.1.5. Two knots K, J are *equivalent* (or *isotopic*) if there is a sequence of intermediate knots $K = K_0, K_1, K_2, \dots, K_n = J$ of knots such that each pair K_i, K_{i+1} is related by a Δ-move.



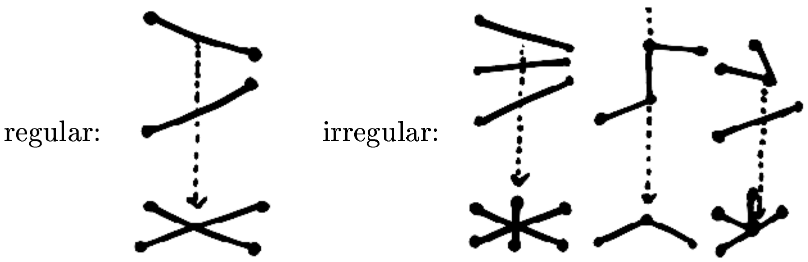
Remark 2.1.6. This is clearly an equivalence relation on knots. We will often confuse knots in $\mathbb{R}^3$ with their equivalence classes (or *knot types*), which are the things we are really interested in topologically. For example, the *unknot* is really the equivalence class of the boundary of 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.

Example 2.1.7. 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 (polygonal subset of the plane homeomorphic to the circle) separates it into two pieces, one of which is homeomorphic to a disc. (The full Jordan curve theorem, which states that *any* embedded subset homeomorphic to the circle separates the plane, is much harder to prove. See Armstrong for some information about both these theorems.) Dividing the component that's homeomorphic to a disc gives a sequence of Δ-moves that shrinks the polygon down to a triangle.

Exercise 2.1.8. Prove the first part of the polygonal Jordan curve theorem as follows. Pick a point p far away from the curve and not collinear with any two vertices of the curve. "Define" a "colouring" function $f : \mathbb{R}^2 - C \to \{0, 1\}$ by $f(x) = |[p, x] \cap C| \mod 2$ (i.e. the number of points of intersection of the arc segment $[p, x]$ with C, taken mod 2). Explain why f is not quite well-defined yet, and what should be added to the definition to make it so. Then show that f is continuous and surjective, proving the "separation" part. Finally, show that there couldn't be a continuous surjective $g : \mathbb{R}^2 - C \to \{0, 1, 2\}$, proving the "two components" part.

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. Thus 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. Thus, 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.

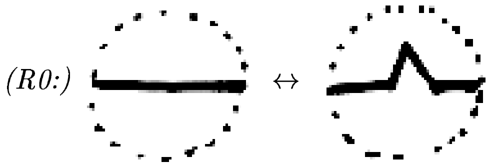
Thus, knots with irregular projections are very rare: the first proposition implies that if one constructed knots by randomly picking their vertices, they would be regular with probability 1. In particular, any knot with an irregular projection need only be wiggled a tiny amount in space to make its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

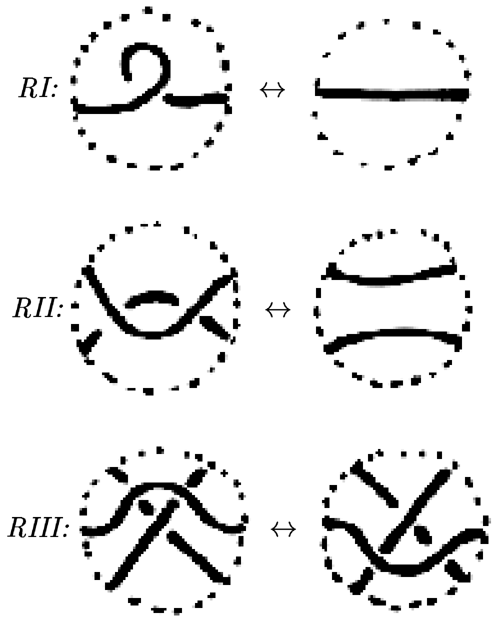
Corollary 2.2.6. *Any knot has a diagram. From a diagram one can reconstruct the knot up to equivalence. Any knot having a diagram with no crossings is an unknot.*

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 adjacency matrix when thought of as a planar graph, its number of regions, 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.

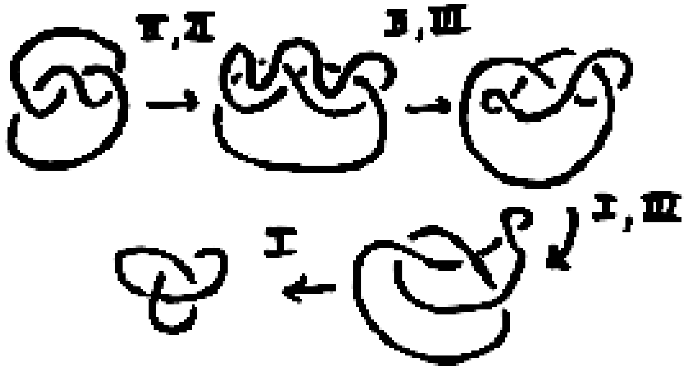
Theorem 2.3.1 (Reidemeister's theorem). *Two knots K, K' with diagrams D, D' are equivalent if and only if their diagrams are related by a finite sequence $D = D_0, D_1, \ldots, D_n = D'$ of intermediate diagrams such that each differs from its predecessor by one of the following three* (*really four, 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 explain 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. Exhibiting a sequence of moves relating two diagrams therefore constitutes a proof that they represent the same knot. (But it is tedious to do, and tricky unless one uses chalk!)



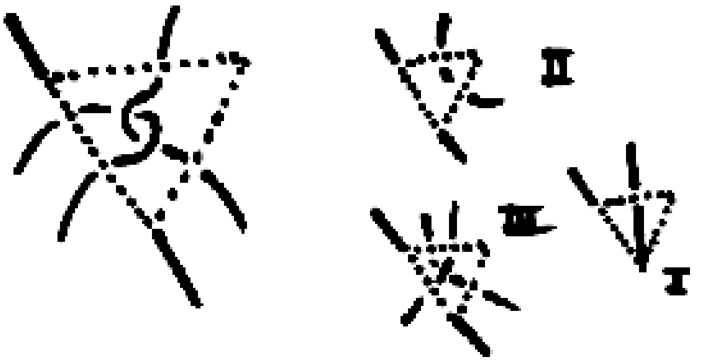
(2). The "zeroth" move is just planar isotopy of diagrams, in other words allows wiggling and stretching of diagrams without changing their combinatorial structure.

(3). Once we have this theorem, we can forget about all the previous technical stuff and simply think of a knot as being an equivalence class of diagrams under Reidemeister moves. This is in fact what Gilbert and Porter do in their book, but it seems a bit artificial to start with that definition.

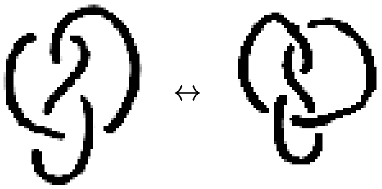
(4). The main way we will use the theorem is to produce invariants of knots. We will construct functions, computable from knot diagrams, which take the same value on all diagrams of a given knot. The way to prove that a function of diagrams is a knot invariant is simply to check that it takes the same 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). One might wonder whether the theorem makes classification of knots by computer possible. A computer can certainly enumerate the finitely many *diagrams* with N crossings or fewer: all we have to do to produce a table of the *knots* with N crossings or fewer is to group these diagrams into Reidemeister-move-equivalent classes. The trouble is that sequences of Reidemeister moves may necessarily increase (at least temporarily) the number of crossings: for example, Adams' book shows a 7-crossing diagram of the unknot, which cannot be reduced to the standard circular diagram without passing through something with more than 7 crossings. Therefore looking for pairs of diagrams on the $\leq N$-crossing table related by a single move is not enough: one is forced to work with diagrams with more than N crossings in order just to classify those with $\leq N$. It is very difficult to bound the number of crossings that might be needed, and this is where the finiteness of the task the computer is undertaking becomes unclear.

Proof of Reidemeister's theorem (sketch). As noted above, the "if" part is trivial, so we consider the "only if" part. Suppose that K, K' with diagrams D, D' are equivalent. Then there is a sequence of Δ-moves getting from K to K'. If one watches these happening in a projection (we can assume all the intermediate knots have regular projections, without much effort) one sees a sequence of diagrams, each obtained from its predecessor by replaced a straight edge by two other sides of a triangle (or vice versa). The projection of the triangle may contain lots more of the knot diagram. If so, subdivide it into smaller triangles so that each contains either a single crossing of the diagram or a single arc-segment. (This corresponds to viewing the Δ-move as the composition of a lot of Δ-moves on smaller triangles.) Then analyse the different possibilities in each case: one sees only Reidemeister moves (see remark below). □



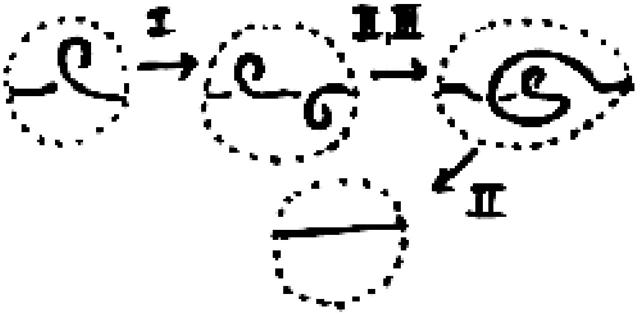
Exercise 2.3.2. Draw a sequence of Reidemeister moves which sends the diagram of the *figure-eight knot* below to its mirror image.



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.

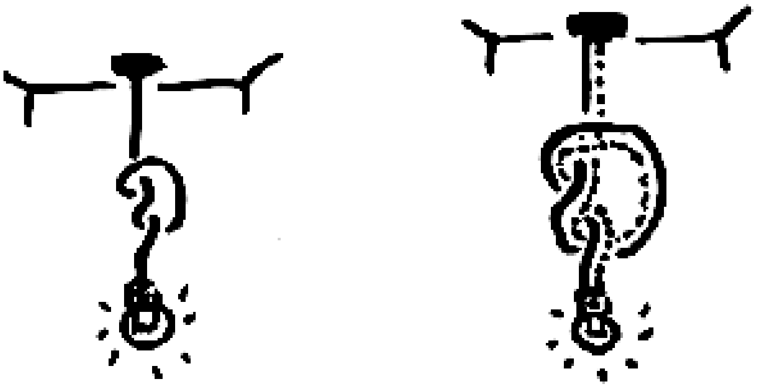
Remark 2.3.5. The statement of Reidemeister's theorem given above is economical in its list of moves (this will be useful in the next chapter.) Suppose for example that one has a knot diagram containing a kink like the one shown above on the left of move RI but with the crossing switched. Move RI does *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 RI, RII and RIII.



In fact RIII 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 RI, one RII and four RIII's). But by similar comositions 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 RI, RII, RIII 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 RI's, four RII's and eight RIII'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?



3. Simple invariants

3.1. **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 3.1.1. A *knot invariant* is any function i of knots which depends only on their equivalence classes.

Remark 3.1.2. 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 3.1.3. 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.

Remark 3.1.4. 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 3.1.5. Link invariants, oriented link invariants, and so on (for all the different types of knotty things we might consider) are defined and used similarly.

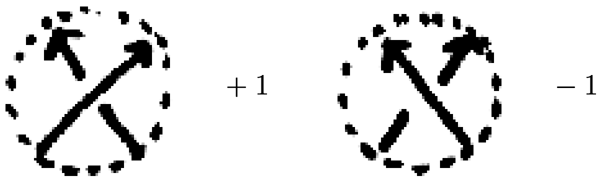
Example 3.1.6. The *crossing number* $c(K)$ is the minimal number of crossings occurring in *any* diagram of the knot K. This is an invariant by definition, but at this stage the *only* crossing number we can actually compute is that of the unknot, namely zero!

Example 3.1.7. 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 3.1.8. 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?

3.2. **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 3.2.1. 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 3.2.2. 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 3.2.3. *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 em 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 RI, one side has an extra crossing but it is a self-crossing, so contributes nothing extra. In RII, 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 RIII, 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 3.2.4. 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.

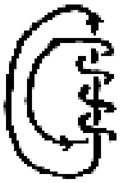
Since for knots, the total linking number is always 0 (all crossings are self-crossings) this invariant is totally useless as a knot invariant.

Exercise 3.2.5. Compute the linking number of the Borromean rings by first choosing an orientation for each component. Does the choice matter?

Exercise 3.2.6. 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 3.2.7. 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 3.2.8. 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 one unknot winds n times (the sign of n denoting the direction of winding) about another. Thus the linking number has a nice visual interpretation as a *winding number* (c.f. complex analysis).



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 k arcs. Then there are 3^k 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 3.2.3 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 RI, 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 RII, 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 RIII 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. □

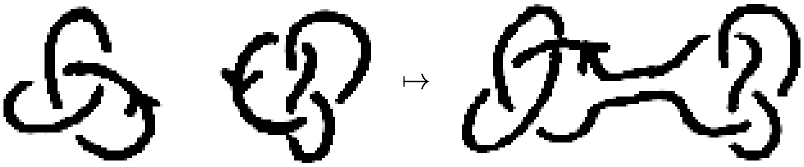
Exercise 3.3.5. Compute $\tau(5_1)$ from its usual diagram. Observe that this invariant does not distinguish it from the unknot.

Exercise 3.3.6. 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?

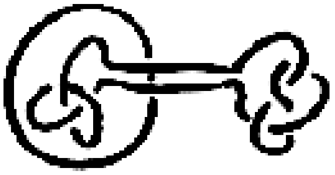
Exercise 3.3.7. Compute the number of 3-colourings τ for the figure-eight knot and the two five-crossing knots.

Exercise 3.3.8. Show that linking number fails to distinguish the Whitehead link from the unlink, but that τ succeeds.

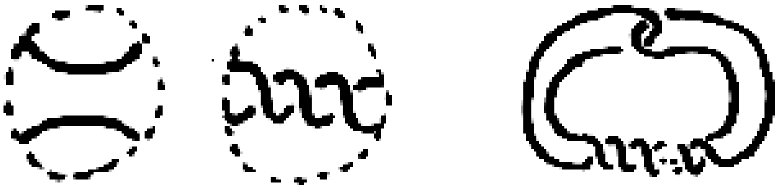
Exercise 3.3.9. The *connect-sum* $K_1 \# K_2$ of two oriented knots K_1, K_2 may be defined by a diagrammatic example like the one below.



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 3.3.10. 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.



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 k arcs $A_1, A_2, \ldots, A_k$, and l crossings $C_1, C_2, \ldots, C_l$.

Exercise 3.3.11. Looking at the knot table one sees that $k = l$ 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

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

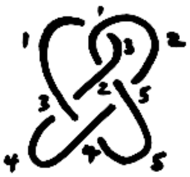
Thus, $T(D)$ is the set of solutions of l homogeneous linear equations in k unknowns over the field $\mathbb{F}_3$.

Theorem 3.3.12. *$T(D)$ 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 $l \times k$ matrix A (over $\mathbb{F}_3$) encoding the l equations from crossings, and want to find the dimension of the space of solutions of $Ax = 0$ ($x \in \mathbb{F}_3^k$). This space is just the kernel, its dimension is just the nullity of the matrix, and we can calculate it by Gaussian elimination.

Example 3.3.13. 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 $\tau(5_2) = 3$.

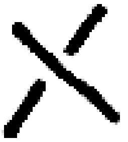
Remark 3.3.14. The most common mistake in computing using this method is to forget that everything is performed mod 3! You have been warned!

Remark 3.3.15. 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.

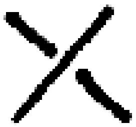
Remark 3.3.16. Livingston talks not about the number of 3-colourings but about "3-colourability of a knot". His definition of a 3-colourable knot is, in our language, one with more than three 3-colourings. Actually counting the number gives more information though, so we will not use his definition.

Exercise 3.3.17. (Harder) Suppose K_+ and K_- are two knots having diagrams D_+ and D_- which are identical except at one crossing, as shown below.

D_+ :



D_- :

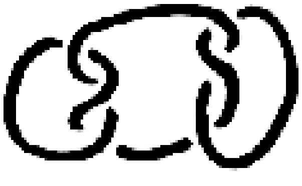


Let T_+, T_- be the vector spaces of 3-colourings of these knots: show that they can be written in the form $T_+ = W \cap V_+$, $T_- = W \cap V_-$, where each of V_+, V_- is the space of solutions of a single equation, and W is some other subspace.

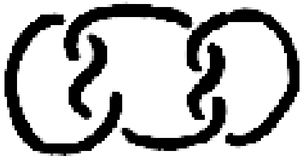
$$\dim(P+Q) = \dim(P) + \dim(Q) - dim(P \cap Q)$$

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. Deduce that the unknotting number of a knot satisfies $u(K) \geq \log_3(\tau(K)) - 1$, and use this to show that both the *reef* and *granny* knots below have unknotting number 2.

Square (reef) knot:



Granny knot:



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 $\{0, 1, \dots, p-1\}$ 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 equation

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

is satisfied.

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.

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".

3.5. **Unknotting number.** We have so far seen five examples of invariants. One (the number of link components), was obvious and not interesting. Three (the linking number, τ and τ_p) 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 other (the crossing number) was an invariant by definition but we have no simple means of computing it at all. The best we seem to be able to do is produce upper bounds (by just drawing diagrams), 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.

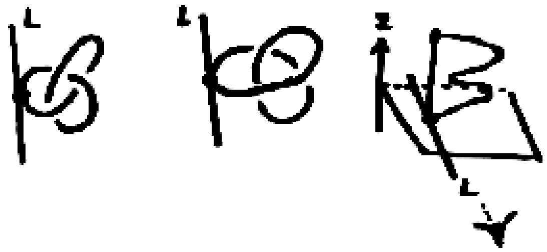
Definition 3.5.1. The *unknotting number* $u(K)$ of a knot K is the minimum, over all diagrams D of K, of the minimal number of crossing changes required to turn D into a diagram of the unknot.

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 3.5.2. *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 tangent to the knot diagram in one point p (so that the whole diagram is "on one side" of this line in $\mathbb{R}^2$). Parametrise the knot in $\mathbb{R}^3$, starting over p, by a map $t \mapsto (x(t), y(t), z(t))$ (which is 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 $t \mapsto (x(t), y(t), t)$ to a vertical arc-segment connecting its endpoints $(p, 0)$ and $(p, 1)$. This knot has the same (irregular, but this is irelevant) xy-projection as K (but with different crossings) and is an unknot, as one can see by "looking along L": its projection along L onto a plane orthogonal to L has no crossings, because the z-coordinate was monotonic and the whole knot lies on one side of L. □



Corollary 3.5.3. *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 $1-t$ instead of t. This is achieved by changing exactly the crossings we *didn't* change to get K', so takes at most $c(K)/2$. □

Exercise 3.5.4. 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!)

4. The Jones polynomial

The Jones polynomial is another combinatorially-defined invariant of links. 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 realised that some of the things he had discovered could (much to everyone's surprise) be used to define a link invariant. This striking connection between two previously separate areas turned out to the tip of a very interesting iceberg, and as a result of his discovery, Jones was awarded the Fields medal in 1990.

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 few 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!

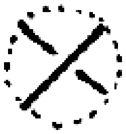
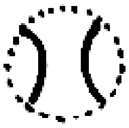
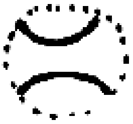
The approach we will take is not Jones' original one, which is quite different and a bit harder. We will first define the *Kauffman bracket polynomial*, which is not an invariant but isn't far off.

4.1. The Kauffman bracket.

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 rules

(0). It is invariant under planar isotopy of diagrams.

(1). It satisfies the *skein relation*

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

which is a linear relation amongst the brackets of diagrams differing only locally inside a small disc as shown.

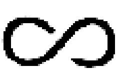
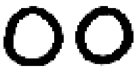
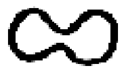
(2). It satisfies $\langle D \amalg U\rangle = (-A^2 - A^{-2})\langle D\rangle$, where U is any closed crossingless loop in the diagram (in other words, disjoint unknot diagrams may be removed at the cost of multiplication by $(-A^2 - A^{-2})$.

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

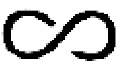
Remark 4.1.2. When applying the skein relation at a crossing, one must be careful to look at the crossing so that the overpass goes from bottom left to top right. Then the term getting the A is the "vertical" reconnection, and the term getting A^{-1} is the "horizontal" one. Alternatively, one can think of turning left from the overpass down onto the underpass to get the A term, and turning right to get the A^{-1} one.

These axioms suffice to calculate the bracket of any diagram. One can use the skein relation 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 4.1.3. By rule (1),

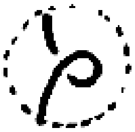
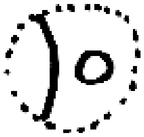
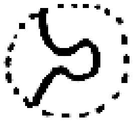
$\langle$  $\rangle = A\langle$  $\rangle + A^{-1}\langle$  $\rangle$.

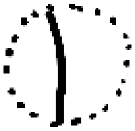
By rules (2) (and (0), which we will start to ignore) this equals $A(-A^2 - A^{-2}) + A^{-1}$ times the bracket of a single circle, which is (by rule (3)) just 1. Therefore

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

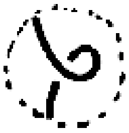
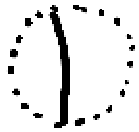
Note that one immediately sees that the Kauffman bracket is *not* an invariant of links!

Example 4.1.4. The same ideas can be applied not to entire link diagrams (as above) but to parts of them. The following identities are between brackets of diagrams differing only in the portions shown, just as in the skein relation (1). The calculation is really just the same as in the previous example.

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

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

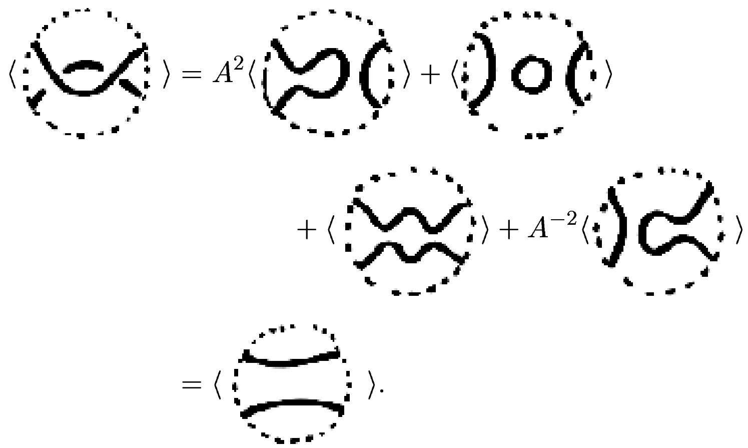
Similarly

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

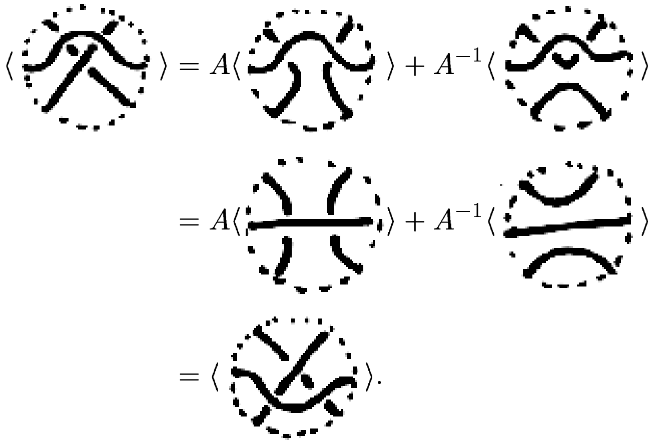
This describes the non-invariance of the bracket under the first Reidemeister move RI.

Lemma 4.1.5. *The Kauffman bracket is invariant under RII and RIII.*

Proof. Applying the skein relation twice, then removing the little circle:


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

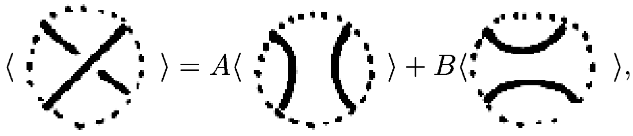
For RIII, we apply the skein relation just once, use the invariance under RII just established, and then the vertical symmetry of the picture:


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

□

Exercise 4.1.6. Suppose we defined a Kauffman-bracket-like invariant (of planar isotopy classes of diagrams) in three variables A, B, d by the following modification of the skein relations:

(1).



shown.

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

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

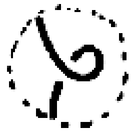
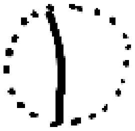
Examine how this "bracket" changes when we perform a Reidemeister move II or III on a diagram. Deduce that we have to set $B = A^{-1}$ and $d = -A^2 - A^{-2}$ in order to get invariance under these moves. Thus, these bizarre choices are essential!

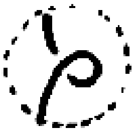
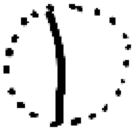
4.2. **Correcting via the writhe.** Now the Kauffman bracket is very close to being a link invariant, as it fails only RI, and even then just multiplies in a simple way by $-A^{\pm 3}$, depending on the "handedness" of the kink. If orientations are chosen everywhere then each crossing has a sign, in 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 RII, RIII but changes by* ± 1 *under RI.*

Proof. This is just another variation of the proof of theorem 3.2.3 on invariance of the linking number. For RII and RIII, it is even easier, as there is no reason to consider whether the strands involved belong to the same component or not. For RI, there is an obvious change. The slightly surprising thing is that the following identities hold *regardless of the orientation on the string* (easy check):

$w($  $) = w($  $) - 1$

$w($  $) = w($  $) + 1.$

□

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.

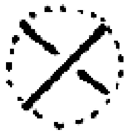
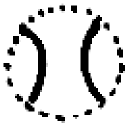
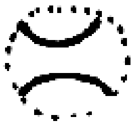
Theorem 4.2.4. *If D is an oriented link diagram, then the polynomial $f_D(A) = (-A^3)^{-w(D)}\langle D\rangle$ is invariant under all three Reidemeister moves, and hence defines an invariant of oriented links.*

Proof. Certainly it is invariant under RII, RIII since both the writhe and bracket are. All that remains is RI. 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

 with  or ,

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. This is simply what one obtains by applying the skein relation at every crossing of D and then evaluating the brackets of the crossingless diagrams that remain. In more detail: suppose one numbers the crossings of D from 1 to c. Then apply the skein relation at crossing 1: we reduce $\langle D\rangle$ 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 diagram at the crossing numbered 2. Now one has a linear combination of four diagrams, each with crossings numbered from 3 to c. Repeat... this terminates with a linear combination of brackets of 2^c crossingless diagrams, indexed by states in the obvious way, and each with a prefactor $A^{\sum s}$. Finally an n-component crossingless diagrams has bracket $(-A^2 - A^{-2})^{n-1}$, completing the proof. (Working this out on the trefoil diagram should make it completely clear.) □

Remark 4.3.5. One sees that that the Kauffman bracket really is well-defined: the above state-sum results, whatever order of application of skein relations was used.

Remark 4.3.6. The computation of the Kauffman bracket is something which can easily be *programmed* on a computer, but is less easily *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... This should be contrasted with the computation of something like the number of 3-colourings τ of a knot diagram, which is basically just Gaussian elimination on a $c \times c$ matrix, which takes of the order of c^2 operations (this is fast even when c is enormous). (Similarly, a human can compute τ of a 5-crossing diagram easily, but will go insane trying to compute its bracket.)

4.4. **The Jones polynomial and its properties.**

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 $A = t^{-1/4}$. It lives by definition in $\mathbb{Z}[t^{\pm 1/4}]$.

Remark 4.4.2. One should think of the polynomials in $\mathbb{Z}[t^{\pm 1/4}]$ as being usual integer-coefficient polynomials in a variable $t^{1/4}$ (and its inverse), for example the polynomial t is really a shorthand for $(t^{1/4})^4$. The notation $V_L(t)$ is not very sensible, given that the polynomial depends on $t^{1/4}$ not just t (for example one can evaluate it, given a value of $t^{1/4}$, but not given a value of t, because of the ambiguity of fourth roots). However, it is traditional!

Theorem 4.4.3. *The Jones polynomial satisfies*

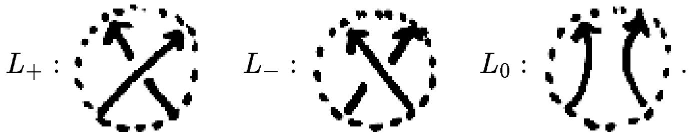
(1). *It is an invariant of oriented links lying in* $\mathbb{Z}[t^{\pm 1/2}]$.

(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. The first property is non-trivial: it asserts that the quarter-integral powers of t are not in fact needed. This is proved by looking at the state-sum definition of the Kauffman bracket: each state contributes a power of $(-A^2 - A^{-2})$ times $A^{\sum s}$, so that all powers of A occurring are even or odd according as the number of crossings of the diagram is even or odd. Therefore the whole Kauffman bracket shares this property. On multiplication by the correction factor $(-A^3)^{-w(D)}$ one ends up with only even powers of A (since the writhe and number of crossings have the same parity) and thus the Jones polynomial involves only powers of $t^{1/2}$ after all. The second property *is* trivial! For the third, 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. So it does not make sense to speak of the Jones polynomial of the link L_∞. 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 $f(D) = (-A^3)^{-w(D)}\langle D\rangle$ for each bracket, and note that the writhes of L_+, L_- are one more and one less than that of L_0. The result is

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

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

Remark 4.4.4. The three properties in this theorem in fact suffice as a *definition* of the Jones polynomial: they can be considered as *axioms* for the Jones polynomial. One can prove (without using the Kauffman bracket) that there exists a unique polynomial satisfying the three properties. This alternative construction of the Jones polynomial is harder than the one based on the bracket, but has the advantage that it generalises to produce two other polynomial invariants, the *HOMFLY* *polynomial* and *Kauffman polynomial* (not to be confused with the Kauffman bracket!), which do not have simple "bracket-style" versions. See the exercises at the end of this subsection (4.4.12-4.4.16) for details.

Theorem 4.4.5. *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,*

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

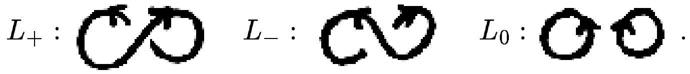
Proof. It is easy to see (either from the skein relation or the state-sum) that the effect of mirror-imaging on the Kauffman bracket is to replace A by A^{-1}. Additionally, mirror-imaging negates the writhe of any oriented diagram, since positive and negative crossings are exchanged. Hence the result. □

The immediate, and very nice consequence of this result is:

Corollary 4.4.6. *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.*

Example 4.4.7. A calculation of the Jones polynomial using only the three axioms, instead of the bracket; we will conclude that the trefoil is chiral.

(1). To compute the polynomial of the unlink, use the following trick: apply the skein relation to the three diagrams



Since the first two are both unknots, the result is

$$t^{-1} - t = (t^{1/2} - t^{-1/2})V_{L_0}(t)$$

i.e $V_{L_0}(t) = (-t^{1/2} - t^{-1/2})$.

(2). The local version of the same trick (applied to positive and negative kinks in a link diagram) shows that for any link L,

$$V_{L \amalg U} = (-t^{1/2} - t^{-1/2})V_L(t),$$

where U is an unknot.

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

$$t^{-1}V_{L_+}(t) - t(-t^{1/2} - t^{-1/2}) = (t^{1/2} - t^{-1/2}),$$

from which $V_{L_+} = -t^{5/2} - t^{1/2}$. (By mirror-imaging one also deduces that the negative Hopf link has polynomial $-t^{-5/2} - t^{-1/2}$. This shows that they are therefore inequivalent oriented links, although we already knew that by using the lining number).

(4). The right-handed trefoil (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. Thus

$$t^{-1}V_{L_+}(t) - t = (t^{1/2} - t^{-1/2})(-t^{5/2} - t^{1/2}),$$

from which $V_{L_+}(t) = -t^4 + t^3 + t$. This polynomial is not palindromic (its conjugate, which is the polynomial of the left trefoil, is $-t^{-4} + t^{-3} + t^{-1}$) and so the left and right trefoils are inequivalent knots.

Remark 4.4.8. The Jones polynomial is very powerful: in practice, it seems to distinguish most pairs of inequivalent knots, although one can construct (using some art!) pairs of inequivalent knots which have the same Jones polynomial, showing that it doesn't always work. However, nobody has ever found a non-trivial knot with polynomial 1: it is *conjectured* that the Jones polynomial *can* distinguish the unknot, i.e. that any knot with polynomial 1 must be the unknot.

Exercise 4.4.9. Compute the Jones polynomial of the figure-eight knot in two ways: first do it by its Kauffman-bracket definition, and then using the Jones skein relation. (Make sure they agree!) Check that the result is consistent with the figure-eight being amphichiral (equivalent to its mirror image)

Exercise 4.4.10. Give a formula for the Kauffman bracket of the connected sum of diagrams $\langle D_1 \# D_2 \rangle$ in terms of $\langle D_1 \rangle$ and $\langle D_2 \rangle$. Use this to derive a formula for the Jones polynomial of the connect-sum of two knots. Give a similar formula for the Jones polynomial of the disjoint union of two knots.

Exercise 4.4.11. 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. Give an example demonstrating that this independence of orientation is not generally true for links with more than one component.

The next three questions show how to work with the Jones polynomial axioms directly, instead of using the Kauffman bracket.

Exercise 4.4.12. Define the *complexity of a link 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 changes needed to make D into an unlink. *Order* these complexities by the rules

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

Let the *complexity of a link* be the minimal complexity of any diagram of it. Now let L be a given link: show that there is a diagram D with a chosen crossing C such that L is one of the three links L_+, L_-, L_0 associated to D and C, and the other two have lower complexity than L.

Exercise 4.4.13. Suppose I is a $\mathbb{Z}[t^{\pm\frac{1}{2}}]$-valued function of oriented links which (1) is an invariant, (2) satisfies the Jones skein relation, and (3) has the value $I(\text{unknot}) = 1$. Show by induction on complexity that I equals the Jones polynomial — in other words, that the Jones polynomial is characterised uniquely by these three properties.

Exercise 4.4.14. 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} appearing, and for links with an even number of components it contains only half-integral powers (i.e. $\dots, t^{-\frac{1}{2}}, t^{\frac{1}{2}}, t^{\frac{3}{2}}, \dots$). (Hint: use induction again.)

Exercise 4.4.15. The "HOMFLY" polynomial $P_L(x, z) \in \mathbb{Z}[x^{\pm 1}, z^{\pm 1}]$ of an oriented link is an invariant based on the Jones polynomial (in fact it was discovered a few months after the Jones polynomial in 1984, and its name consists of the initials of its six discoverers). It is defined rather like the Jones polynomial by (a) it is an invariant, (b) it satisfies the skein relation

$$x^{-1} P_{L_+} - x P_{L_1} = z P_{L_0}$$

(with the usual meanings of L_+, L_-, L_0), and (c) $P(\text{unknot}) = 1$. (Observe that the polynomial is uniquely defined for all links by these properties just as in exercise 4.4.13.) (Note also that many books use different names or signs for the variables.) Calculate the HOMFLY polynomial of the two-component unlink and of the left- and right-handed trefoils.

Exercise 4.4.16. Show that the HOMFLY polynomial determines the Jones polynomial of a link.

Exercise 4.4.17. 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 $z = t^{\frac{1}{2}} - t^{-\frac{1}{2}}$ 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). Show by induction that for any link, the Alexander polynomial lies in $\mathbb{Z}[z]$ (i.e. that it has no negative powers of z). Show that for a *knot K*, $\nabla_K(z)$ always has constant term 1. 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.

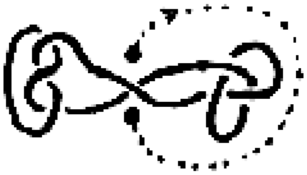
4.5. **Alternating knots and the Jones polynomial.** A natural conjecture one comes up with when playing with knot diagrams is that (as long as one avoids certain obviously reducible cases, as we will explain below) *alternating diagrams are minimal*, i.e. that any knot represented by an alternating diagram cannot be represented by any other diagram with fewer crossings. This conjecture is one of the so-called *Tait conjectures*, made about 100 years ago. No progress was made on any of these conjectures until the advent of the Jones polynomial, after which they were soon dealt with. In this section we will prove this conjecture.

Definition 4.5.1. 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). (Note: we do not attempt to define alternating links here!)

Definition 4.5.2. 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). Any diagram separates the plane into a number of *regions* (this includes the outer unbounded one).

Fact 4.5.3. For a connected diagram, all the regions are homeomorphic to discs, and the number of regions is the number of crossings plus 2. This theorem can be proved using Euler characteristic arguments from the next chapter. For the moment we will take it as a fact.

Definition 4.5.4. An *isthmus* of a knot diagram is a crossing at which there are less 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 half of the diagram, destroying an isthmus, until there are none left.



Definition 4.5.5. 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 $-A^5 + 2 + A^{-3} - 3A^{-5}$ is 10.

Remark 4.5.6. 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 $A = t^{-1/4}$).

Having set up all the relevant terminology, the theorems are as follows.

Theorem 4.5.7. *The span of the Kauffman bracket of a c-crossing reduced alternating knot diagram is exactly* $4c$.

Theorem 4.5.8. *The span of the Kauffman bracket of any c-crossing knot diagram is less than or equal to* $4c$.

The proof of these two theorems will fill the rest of the section. But let us first consider their immediate consequences. The first corollary is the positive solution of one of Tait's conjectures on alternating knots. Recall that a minimal diagram is one whose number of crossings equals the crossing number of the knot, so that the knot has no diagram with fewer crossings. Of course, there could be several other diagrams with the *same* number – we are not claiming that a minimal diagram is unique.

Corollary 4.5.9. *Any reduced alternating knot diagram is minimal.*

Proof. 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 first theorem. However, the *span* of the Kauffman bracket is a knot invariant (recall how the bracket changes under RI) and so there cannot be any diagrams of the same knot with fewer than c crossings, otherwise the second theorem would be contracdicted. □

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

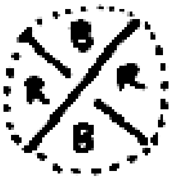
Corollary 4.5.10. *Any non-trivial reduced alternating knot diagram represents a non-trivial knot.*

Corollary 4.5.11. *All reduced alternating diagrams of the same knot have the same number of crossings.*

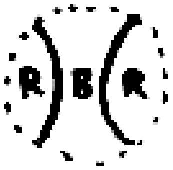
To prove the theorems, we need to identify the highest and lowest powers occurring in the bracket of a c-crossing diagram D, and for this we use the state-sum model. Let s_+, s_- be the states consisting entirely of pluses and minuses respectively. Then it seems reasonable that s_+ might contribute the highest positive power, and s_- the highest negative power, because for these states $\sum s$ is $\pm c$. What is not immediately clear is how $|sD|$ behaves as s ranges over all states, and this is what we analyse below.

Lemma 4.5.12. *For a reduced alternating knot diagram,* $|s_+D| > |s_1D|$ *for any state* s_1 *which has exactly one minus.*

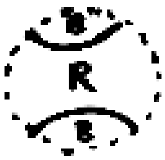
Proof. Start by colouring the corners of the regions incident at each crossing either red or blue, according to the picture



For an alternating knot diagram, every region is a polygon with crossings as vertices and arcs of the knot as edges. Alternation means that the patches of colour assigned to the corners of each polygon are the same: consequently, each region gets a well-defined colour, red or blue. Now on resolving the diagram according to s_+, one obtains a crossingless diagram with coloured complementary regions, which looks near a crossing like



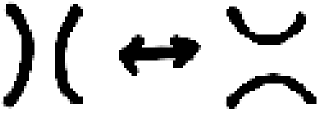
Thus, the loops of s_+D are precisely the boundaries of the red regions of the original coloured diagram. Now if s is any state with one minus then sD differs from s_+D near just one crossing according to the picture



Since the original diagram was reduced, the two red regions seen here on the left are different (otherwise this crossing would be an isthmus) and so on the right, they have been connected together. So the number of loops $|sD|$, which is the number of red regions in the right-hand figure, is one less than the number of red regions on the left, which is $|s_+D|$. □

Lemma 4.5.13. *Let D be* any *c-crossing knot diagram, and s any state with i minuses. Let $s_+ = s_0, s_1, \dots, s_i = s$ be a chain of states, each s_j having j minuses, connecting s_+ to s. Then the maximal power in $\langle D|s_{j+1}\rangle$ is less than or equal to that of $\langle D|s_j\rangle$ for each j.*

Proof. Simply examine the terms involved: $\langle D|s_j\rangle = A^{c-2j}(-A^2 - A^{-2})^{|s_jD|-1}$ and $\langle D|s_{j+1}\rangle = A^{c-2(j+1)}(-A^2 - A^{-2})^{|s_{j+1}D|-1}$. Since $s_{j+1}D$ is obtained from s_jD by a local alteration



the number of components $|s_{j+1}D|$ is either one more than or one less than $|s_jD|$ (if the two arcs belong to the same loop to start with, they don't afterwards, and vice versa). Therefore the power of $(-A^2 - A^{-2})$ can increase by at most one as one goes from s_j to s_{j+1}, but it is counteracted by the decrease of $\sum s$ by two, proving the lemma. □

Proof of theorem 4.5.7. The highest power in $\langle D|s_+\rangle$ is $c + 2(|s_+D| - 1)$, by definition. The highest power in $\langle D|s_1\rangle$, for any state s_1 with one minus, is $(c-2) + 2(|s_+D| - 2)$, which is *strictly less* (in fact four less!), by lemma 4.5.12. By lemma 4.5.13, no other states can contribute bigger powers than can s_1. So the highest degree occurring in $\langle D\rangle$ is indeed $c + 2(|s_+D| - 1)$. By the same argument starting from s_- (just exchanging the roles of plus and minus) one finds the lowest power to be $-c - 2(|s_-D| - 1)$. The span is therefore $2c + 2(|s_+D| + |s_-D|) - 4$, but since $|s_+D|, |s_-D|$ just count the numbers of red and blue regions of the original diagram, their sum is equal to $c + 2$, and the theorem is proved. □

Proof of theorem 4.5.8. If the diagram is not assumed to be alternating, two problems arise in the above proof. The first is that lemma 4.5.12 is false, so we don't get the *strict* drop in degree from s_+ to s_1. This means that other states may contribute terms that *cancel out* the $A^c + 2(|s_+D| - 1)$ contributed by s_+, leading to a possibly drastic drop in maximum degree of $\langle D \rangle$. However, lemma 4.5.13 does still apply, and shows that the biggest power of A that *might* occur in $\langle D \rangle$ is $c + 2(|s_+D| - 1)$. So we have at least proved that the span of $\langle D \rangle$, for any c-crossing diagram, is *less than or equal to* (rather than equal to, as in the previous theorem) $2c + 2(|s_+D| + |s_-D|) - 4$. The second problem is that at the end, we don't know that $|s_+D| + |s_-D|$ is just the number of regions of the original diagram. But the following lemma, applied to D and the state s_+, shows that $|s_+D| + |s_-D| \leq c + 2$, which is enough to finish the proof. □

Lemma 4.5.14 (Dual state lemma). *For any state s, let $\hat{s}$ denote its* dual *or* opposite, *given by exchanging all pluses and minuses. Then, for any connected diagram D and any state s, we have*

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

Proof. Induction on number of crossings c. Start with $c = 1$: the only diagram is a figure-of-eight, for which the two states (which are dual) result in a 1-loop and a 2-loop diagram, so the lemma is true here. Now suppose D is a c-crossing diagram with a state s, and that we have the lemma for all diagrams with $c-1$ crossings or fewer. Pick a crossing C of D, and resolve it there in both possible ways to get two diagrams with $c - 1$ crossings. At least one of these must be connected, because if neither is then the four arcs leaving the crossing must never return, which is ludicrous. Call this diagram E, and assume that it is obtained by splitting C as instructed by $s(C)$ – otherwise rename s and $\hat{s}$, which does not affect what we're trying to prove. Let t be the restriction of s to E, so that $tE = sD$. If $\hat{t}$ is the dual of t on E, then $\hat{t}E$ differs from $\hat{s}D$ near just the crossing C, because $\hat{s}D$ involves splitting *all* crossings of D according to $\hat{s}$, whereas $\hat{t}E$ involves splitting all crossings of D except C according to $\hat{s}$, and C according to s. Therefore $|\hat{t}E| = |\hat{s}D| \pm 1$, since they differ only at one crossing. Now the inductive hypothesis is that

$$|tE| + |\hat{t}E| \leq c + 1,$$

which becomes (substituting what we just worked out)

$$|sD| + (|\hat{s}D| \pm 1) \leq c + 1,$$

which implies as required that

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

□

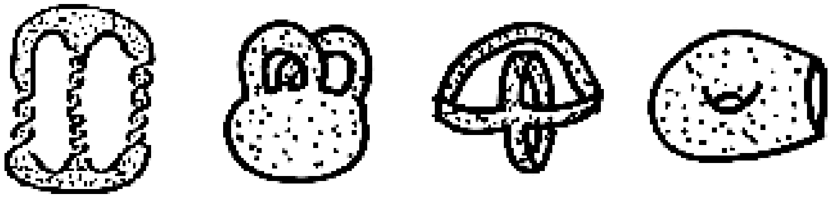
5. Surfaces

5.1. **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 of $\mathbb{R}^3$, 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.

The *intrinsic* topology of a link is not very interesting: as a topological space, it is homeomorphic to a disjoint union of circles, and so is classified up to *homeomorphism* by its number of components. Of course we are interested in the more subtle problem of understanding *equivalence classes* of knots and links.

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 relationship between knots and surfaces.

Exercise 5.1.1. These four surfaces are homeomorphic!



Definition 5.1.2. An *n-dimensional manifold* is a Hausdorff topological space M, such that every point of M has a neighbourhood homeomorphic to $\mathbb{R}^n$.

Remark 5.1.3. Since $\mathbb{R}^n$ is homeomorphic to the open unit ball, we can consider the neighbourhoods to be small open balls instead (this seems more visually appealing). One can also imagine the neighbourhoods homeomorphic to $\mathbb{R}^n$ as providing local coordinate systems $\{(x_1, x_2, \ldots, x_n)\}$ for regions of the manifold.

Remark 5.1.4. Strictly, a manifold is also required to be *second countable*, i.e. have a countable base of open sets, but we can safely ignore this technicality. Note that a manifold is not required to be compact or connected, though on the whole these will be the ones we're interested in.

Remark 5.1.5. Any subspace of $\mathbb{R}^N$ which is locally homeomorphic to $\mathbb{R}^n$ $(N \geq n)$ is a manifold in the sense above (in particular it is Hausdorff and second countable). Conversely, any *compact* n-manifold can be embedded in (mapped homeomorphically to a subspace of) $\mathbb{R}^N$, for some suitably large N. Thinking of manifolds as subspaces can help in visualising them, but it is very important to realise that the way the manifold is embedded in $\mathbb{R}^N$ is *not* an intrinsic part of its definition.

Exercise 5.1.6. Determine which of the following spaces is a manifold:

$$S^1, \quad S^1 \amalg S^1, \quad \mathbb{R}, \quad I, \quad (0,1), \quad (0,1], \quad \text{the “open letter } Y\text{”},$$

$$S^2, \quad S^1 \times S^1, \quad \text{the open unit disc}, \quad S^1 \times S^1 - (\text{point}),$$

$$S^1 \times S^1 - (\text{open disc}), \quad \text{an open subset of } \mathbb{R}^2.$$

For each space, say whether it is connected or disconnected, compact or non-compact. 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 5.1.7. 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 5.1.8. 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 5.1.9. Repeat exercise 5.1.6, identifying the manifolds-with-boundary.

Definition 5.1.10. 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 5.1.11. 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 5.1.12. 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.

Definition 5.1.13. From now on, all the manifolds we deal with will be compact, and we will redefine the terminology to make it less cumbersome: *compact manifold* will mean a compact manifold with or without boundary; *closed manifold* will mean compact manifold with empty boundary. Remember that not all manifolds are connected (though we will be interested mainly in those that are, for obvious reasons). The word *surface* means simply 2-manifold.

5.2. **Examples of surfaces.** Here are various exercises in geometric visualisation and constructing homeomorphisms. For the questions involving explicit homeomorphisms, it will be important to know how to work with the *quotient* (or *identification*) *topology* on a space.

Definition 5.2.1. If X is a topological space and $\sim$ is an equivalence relation on X, then the *quotient space* $X/\sim$ (the set of equivalence classes) inherits a *quotient topology* from X as follows. Let π be the *quotient map* $X \to X/\sim$. Then we define U to be an open set in $X/\sim$ if and only if $\pi^{-1}(U)$ is open in X. With this topology, the map π is continuous.

Remark 5.2.2. A map $g : X/\sim \to Y$ induces and is induced by a map $f : X \to Y$ taking the same values on equivalent points, according to the formula $f = g \circ \pi$.) Such a g turns out to be continuous if and only if its corresponding f is. This gives us a simple way of writing down maps from a quotient space to another space and checking their continuity.

Remark 5.2.3. Any quotient of a compact space is compact, and any quotient of a connected space is connected.

Remark 5.2.4. Often one wants to identify a quotient space $X/\sim$ as being homeomorphic to some other space Y. A particularly easy case occurs when $X/\sim$ is compact and Y is Hausdorff, because then any continuous bijection $X/\sim \to Y$ is a homeomorphism (its inverse is forced to be continuous).

Example 5.2.5. Here are some very simple examples of surfaces. We already know about the sphere S^2 and torus $S^1 \times S^1$ (closed surfaces), and the closed unit disc D (a surface with boundary). The *Möbius strip* is defined as a quotient:

$$(I \times I)/((x,0) \sim (1-x,1)),$$

as are the *Klein bottle*

$$(S^1 \times I)/((e^{i\theta},0) \sim (e^{-i\theta},1)),$$

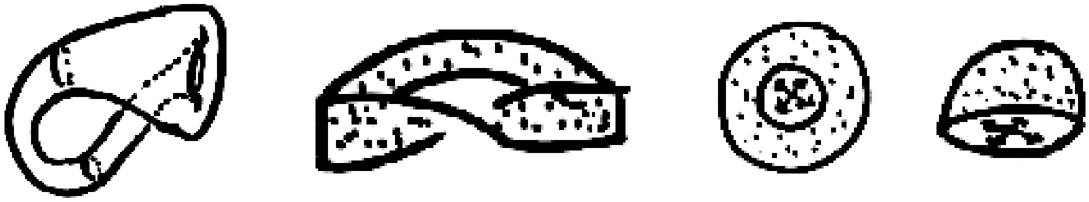
the *crosscap*

$$(S^1 \times I)/((e^{i\theta},0) \sim (-e^{i\theta},0)),$$

and the *projective plane*

$$D/(e^{i\theta} \sim -e^{i\theta}).$$

The projective plane and Klein bottle are closed, while the Möbius strip and crosscap (which are actually homeomorphic, but both visualisations are useful) have one boundary component.



Exercise 5.2.6. Show using an explicit map that the crosscap is homeomorphic to the Möbius strip.

Exercise 5.2.7. Show that the Klein bottle is homeomorphic to the union of two copies of a Möbius strip joined (by a homeomorphism) along their boundaries.

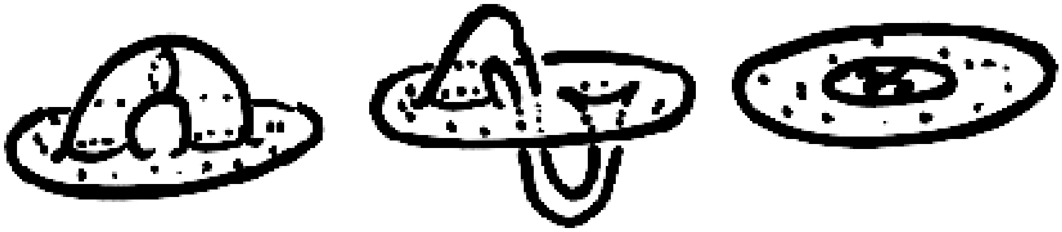
Exercise 5.2.8. Let M_1, M_2 be spaces with homeomorphisms h_1, h_2 to the standard closed unit disc D. Let A_1, A_2 be subsets of their boundaries with homeomorphisms g_1, g_2 to I. Show that the union $M_1 \cup M_2$, where $g_2^{-1}g_1 : A_1 \to A_2$ is used to glue the arcs A_i together, is also homeomorphic to a closed unit disc.

Exercise 5.2.9. Show that gluing M_1 and M_2 along their boundaries via the map $h_2^{-1}h_1 : \partial M_1 \to \partial M_2$ yields a space homeomorphic to S^2.

There are several standard ways of altering surfaces by cutting and pasting. These will be important when we come to classify surfaces.

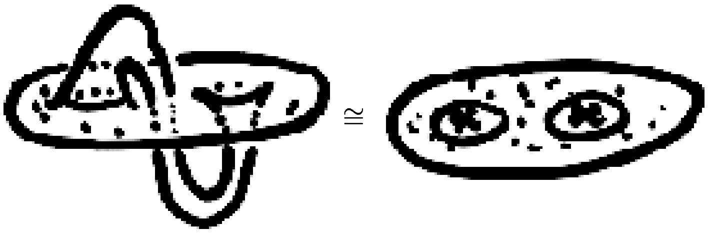
Definition 5.2.10. Consider the surfaces made by removing the interior of a closed disc from a torus or Klein bottle: they have one boundary circle, as does the crosscap. If D is a closed disc contained in a surface F then one may remove the interior of D, creating a boundary component, to which one 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.

gets replaced by one of:

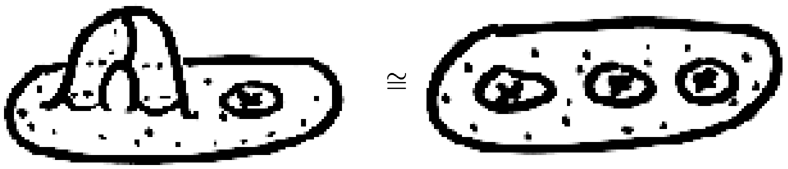


Remark 5.2.11. The resulting surface is unique up to homeomorphism: it does not matter where the disc lies in the surface.

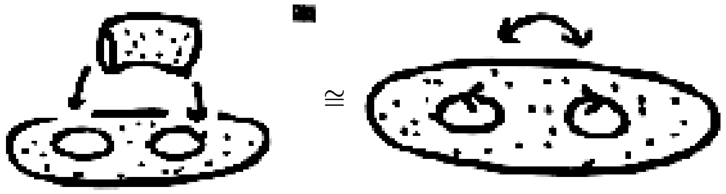
Exercise 5.2.12. Show that a disc with a twisted handle attached is homeomorphic to a disc with two crosscaps attached.



Exercise 5.2.13. Show that a disc with a crosscap and a handle attached is homeomorphic to a disc with three crosscaps attached.



Exercise 5.2.14. Let E be the disc of radius 10 in $\mathbb{C}$ minus the open unit discs centred at $z = \pm 5$. Let X be the the space $E \cup (S^1 \times I)$, where the cylinder is attached via $(e^{i\theta}, 0) \sim -5 + e^{i\theta}$ and $(e^{i\theta}, 1) \sim 5 + e^{-i\theta}$. Let Y be X with the identification $-5 + e^{i\theta} \sim 5 + e^{-i\theta}$. Prove *explicitly* that X, which is the disc with a handle added, is homeomorphic to Y. What happens when the '$e^{-i\theta}$'s are replaced with '$e^{i\theta}$'s? This gives an alternative way of looking at the addition of a handle or twisted handle.



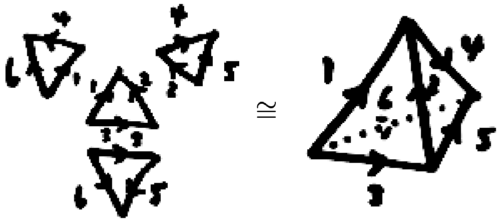
5.3. **Combinatorial surfaces.** Just as when working with knots, we will find it helpful to work with a *combinatorial* (meaning, roughly, *discrete and finite*) version of the concept. This will enable proofs by induction and many other simplifications. The appropriate concept is that of a surface built by gluing a lot of solid triangles together by identifying their edges together in pairs (and giving the resulting space the quotient topology).

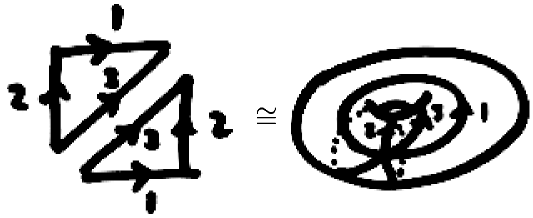
Definition 5.3.1. Let T be the standard closed triangle, the convex hull of the three standard basis vectors inside $\mathbb{R}^3$. Explicitly, this means the subset $T = \{(\lambda_1, \lambda_2, \lambda_3) : 0 \leq \lambda_i \leq 1, \lambda_1 + \lambda_2 + \lambda_3 = 1\}$. Consider T as a space in its own right (with the subspace topology from $\mathbb{R}^3$).

Definition 5.3.2. Let $\amalg T$ be a disjoint union of f copies of T, for some positive even integer f. A *gluing pattern* on $\amalg T$ is a pairing of the $3f$ edges (now you see why f must be even!), indicated by labelling the edges by symbols, each appearing twice, together with an assignment of an arrow to each edge.

A gluing pattern generates an equivalence relation on $\amalg T$. Each point on an edge of a triangle is identified with a point on the other like-labelled edge, using the unique linear homeomorphism between the two determined by their arrows. The result is a quotient space F and a quotient map $\pi : \amalg T \to F$. Clearly π 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. As a result, F can be expressed as a disjoint union of open triangles, open unit intervals and points called *faces, edges and vertices.* (We will usually think of the faces and edges as closed rather than open.) There are $3f$ faces, $3f/2$ edges and somewhere between 1 and $3f$ vertices in F (every edge is common to exactly two faces of F, but we cannot immediately say how many faces share a vertex of F).

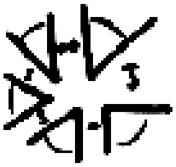
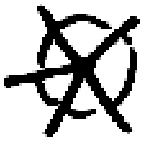
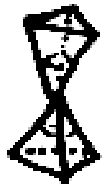
Example 5.3.3. Here are two gluing patterns, one producing a sphere and one a torus.





Lemma 5.3.4. *Any space F obtained as above is a closed manifold.*

Proof. Certainly such an F is compact, as it is the quotient of a compact space. We must prove that all points of F have a neighbourhood homeomorphic to $\mathbb{R}^n$. Since the quotient map $\pi : \amalg T \to F$ is injective on the interiors of faces, all points in the interiors of faces of F have Euclidean neighbourhoods (in fact the open faces themselves will do!). Any point in the interior of an edge of F has a neighbourhood consisting of the union of the interiors of the edge and the two incident faces, which is clearly homeomorphic to an open disc. Given a vertex v of F, consider an open ball neighbourhood of small radius about its preimage in $\amalg T$: this consists of a disjoint union of "open corners" of triangles of T, since the preimage of the vertex is just some subset of the $3f$ vertices of $\amalg T$. The image of this open set in F consists of the open corners, glued together along their edges (two open corners meeting at each edge) and with common vertex v. Since the vertices of $\amalg T$ only get glued together as a result of edge identifications, the result is a single open disc, rather than several open discs joined at their centres, as shown below. □

Near vertex:  $\cong$  Not allowed: 

Remark 5.3.5. We could similarly build 1-dimensional manifolds by pairing the vertices of a disjoint union of closed unit intervals, and the quotient space would always be a compact 1-manifold (therefore homeomorphic to a disjoint union of circles).

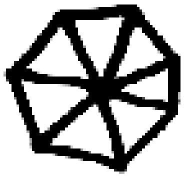
Remark 5.3.6. If we want to build surfaces with boundary as well as closed surfaces then only a small modification of the definition is needed: we redefine a gluing pattern (now allowing odd numbers of triangles!) as a pairing up of *some* of the edges of $\amalg T$. After gluing, the unpaired (unlabelled) edges will remain as "free edges" of the surface F.

Exercise 5.3.7. A space F constructed from such a generalised gluing pattern is a manifold with boundary. The boundary ∂F consists of all unpaired edges, and is a union of circles (made by gluing up unit intervals as explained above).

Remark 5.3.8. Another generalisation of the gluing procedure is that one can build n-dimensional objects by gluing together *n-simplexes* (n-dimensional analogues of tetrahedra) in pairs along their faces. But when $n \geq 3$, lemma 5.3.4 fails: the result of gluing might *not* be an n-manifold, so one has to be much more careful.

It will simplify our proofs a bit if we also restrict ourselves to surfaces satisfying an additional restriction on how their faces intersect. This condition is really just an added convenience, on top of the idea of of working with surfaces made of triangles. All the definitions and theorems below could be modified to work without it, but they are simpler with it.

Definition 5.3.9. A *cone* is a surface with boundary made by gluing d faces arranged around a common vertex, for some $d \geq 3$. Of course it is homeomorphic to the closed disc, and its boundary (a d-sided polygon) is homeomorphic to the circle. Note that a cone must have at least three sides.



Definition 5.3.10. A closed surface F made by gluing triangles is called a *closed combinatorial surface* if the union of the (closed) faces incident at any vertex of F, thought of as a subspace of F, is a cone (centred on that vertex). (On a surface with boundary, the corresponding definition is to require this condition at all vertices not in the boundary.)

Remark 5.3.11. To understand this definition, look again at the proof of lemma 5.3.4, that glued triangles always give a manifold. Disc neighbourhoods of the vertices were constructed by gluing together "open corners". If we had done a more obvious thing, gluing together all the faces incident to a vertex (rather than just their corners), we might not have ended with a disc at all. In the case of the two-triangle torus of example 5.3.3, we get the whole torus!. This is an irritation we can do without: it is caused by the fact that the two triangles are just too big, and if we chopped them up into lots of smaller ones, this problem would go away. So the definition of combinatoriality just given is in a sense an expression of the fact that the triangles making up our surface shouldn't be too big!

Exercise 5.3.12. Prove that on a combinatorial surface:

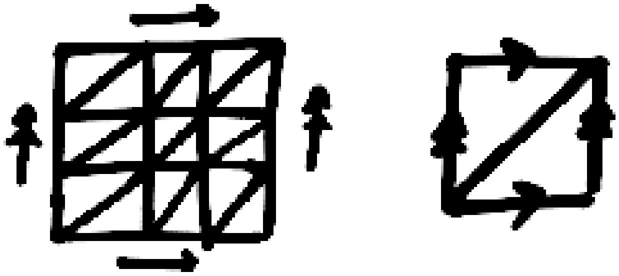
(1). The map $\amalg T \to F$, restricted to any *closed* face, is an injection ("no face is glued to itself").

(2). Any two distinct closed faces of F meet in either a single common edge, a single common vertex, or are disjoint.

Prove the converse: these two conditions imply that every interior vertex has a cone about it, so these two conditions give an equivalent definition of combinatoriality.

Remark 5.3.13. This terminology is not standard. It is equivalent to the fact that F is a *simplicial complex* (see Armstrong), but this is more complication than we need to use.

Example 5.3.14. Here are two gluing patterns forming the torus, one combinatorial and one non-combinatorial.



Remark 5.3.15. In the 1-dimensional case (remark 5.3.5) there is an analogous notion of combinatoriality. In this case a cone is simply two edges glued at a single common vertex. Thus, a circle made by gluing intervals is combinatorial if and only if it uses at least three edges. A "two-sided polygon" and a single interval with its ends glued together are ruled out.

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.

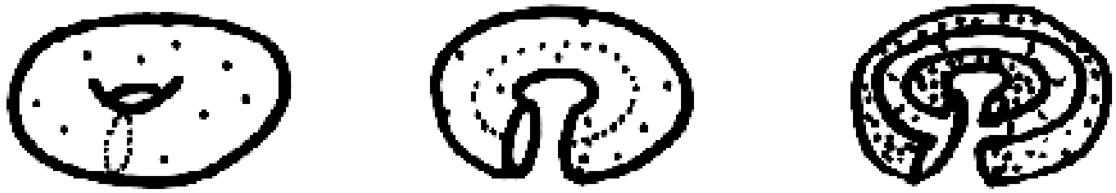
Fact 5.3.16. Any compact 2-manifold is homeomorphic to a combinatorial surface. (Idea of proof: think of dividing up the surface into regions homeomorphic to the triangle; if the triangles are "small enough" then we will satisfy the cone neighbourhood condition and get a combinatorial surface.)

5.4. **Curves in surfaces.**

Definition 5.4.1. A (simple closed combinatorial) *curve* $C \subseteq F$ is a union of edges, disjoint from ∂F, which is homeomorphic to a circle. A (proper combinatorial) *arc* in a surface with non-empty boundary is a union of edges, homeomorphic to the unit interval, and meeting ∂F only in its two endpoints.

Definition 5.4.2. If F is a surface and C a curve, then we can define a new surface F' obtained by *cutting along* C by simply removing (from the gluing pattern that builds F) the identification instructions on all edges that map to C. That is, F' is formed by identifying the same set of triangles used to build F, but without gluing across any of the edges of C. The same definition is used when cutting a surface with boundary along an arc.

Example 5.4.3. Cutting a torus along a curve then an arc.

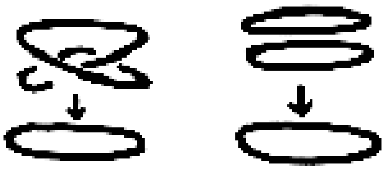


Exercise 5.4.4. The spaces obtained from cutting along curves and arcs are indeed combinatorial surfaces, according to the definitions.

Remark 5.4.5. The space F' is *not* the same as $F - C$, the complement of C (which is non-compact, while F' is compact).

Lemma 5.4.6. *There is a continuous "regluing" map $p : F' \to F$. The boundary of F' is $\partial F' = p^{-1}(C) \amalg \partial F$, and the new part $p^{-1}(C)$ consists of either one or two circles.*

Proof. The map is defined by re-identifying the edges in F' which we just "un-identified". It is a quotient map and therefore continuous. Slick proof using covering spaces: check that restricted to the boundary of F', p is a $2:1$ covering map onto C. But there are only two double covers of the circle, the connected one and the disconnected one. □



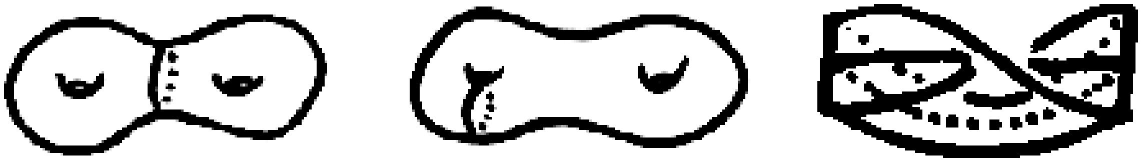
Exercise 5.4.7. Rewrite this proof in purely combinatorial language, avoiding reference to covering spaces.

Remark 5.4.8. If there are two new circles then each has as many edges as C; if there is only one new circle, it has twice as many edges as C.

Definition 5.4.9. A curve $C \subseteq F$ is called 1-*sided* or 2-*sided* according to the number of components of $p^{-1}(C)$.

Definition 5.4.10. A curve C is called *non-separating* or *separating* according to whether F' has the same number or more components than F.

Example 5.4.11. A 2-sided separating curve, 2-sided non-separating curve and 1-sided (thus non-separating) curve, the centreline of the Möbius strip. (When you cut along it you get an annulus twice as long as the original strip: it has two boundary components, compared with the Möbius strip's one.)



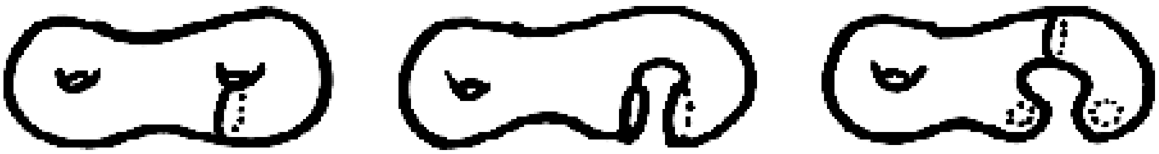
Exercise 5.4.12. Show that cutting along a separating curve increases the number of components of a surface by 1.

Exercise 5.4.13. Show that cutting along a 1-sided curve cannot separate a surface (i.e. 1-sided curves are always non-separating).

Remark 5.4.14. In order to get a better understanding of 1-sided curves, it is useful to introduce the idea of a *neighbourhood* of a subset of a surface F, meaning the image of all points in $\amalg T$ within some small distance ϵ of the preimage of the subset. The neighbourhood of a curve C, for example, consists of the union of thin strips along the sides of the triangles that map to C, and small corner segments at the vertices that map to vertices of C. The neighbourhood is a kind of thickening of the curve into a band: it is homeomorphic to $[-\epsilon, \epsilon] \times I$ glued at its thin ends, and therefore is homeomorphic either to an annulus or to a Möbius strip. If we cut the surface along C, the cut-up neighbourhood (which is either two annuli or one double-length one, accordingly) becomes a neighbourhood of the new boundary. Therefore, a neighbourhood of a 2-sided curve is an annulus and a neighbourhood of a 1-sided curve is a Möbius strip.

The proof of the classification theorem will be by a cut-and-paste process called *surgery*.

Definition 5.4.15. If C is a curve in F, then *surgery on C* is the operation of cutting F 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 F_C.

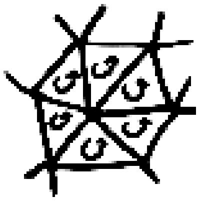


5.5. **Orientability.** There are lots of equivalent definitions of orientability, and which one to use as *the* definition is a matter of taste.

Definition 5.5.1. A surface is *orientable* if it contains no 1-sided curves.

Remark 5.5.2. A surface is orientable if and only if it does not contain any subspace homeomorphic to the Möbius strip. In one direction this follows because a neighbourhood of a 1-sided curve is a Möbius strip, but in the other one needs to assume that the existence of a Möbius strip somewhere in the surface implies the existence of one as a neighbourhood of some curve. This is a consequence of the classification of surfaces (theorem 5.7.6).

Exercise 5.5.3. Yet another alternative definition goes as follows. 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. Show directly that a surface has an orientation if and only if it is orientable in the sense that it contains no 1-sided curves.



Exercise 5.5.4. 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. (This notion is the same as the surface itself "having two sides", though this is not an intrinsic notion, which is why we restrict in this question to surfaces contained in $\mathbb{R}^3$. It will however be very useful in the next chapter, in which all our surfaces will lie inside $\mathbb{R}^3$.)

5.6. **Euler characteristic.** 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 5.6.1. 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 5.6.2. 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 5.6.3. The Euler characteristic of any combinatorial circle is 0.

Exercise 5.6.4. If F' is obtained by cutting F along C, then $\chi(F') = \chi(F)$.

Exercise 5.6.5. If F_C is obtained by doing surgery along C, then $\chi(F_C)$ is either $\chi(F) + 1$ or $\chi(F) + 2$, depending on whether C is 1-sided or 2-sided.

Definition 5.6.6. A *graph* is a space made by gluing a disjoint union of closed unit intervals together at their endpoints. This kind of gluing is more general than the kind we used (example 5.3.5) when defining a combinatorial 1-manifold, as we can identify many vertices together (rather than just gluing in pairs) and can produce multiple edges and loops in the quotient (though not isolated vertices). Graphs have an Euler characteristic $v - e$ in the obvious way. Define also the *degree* or *valence* of a vertex in a graph as the number of preimages it has under the gluing map. This is the same as the number of incident *ends* of edges, rather than of edges: a graph with one vertex and one edge attached to it has a vertex of degree 2, not 1.

Exercise 5.6.7. There are two possible definitions of connectedness for a graph: either one can think of it as a topological space (using the quotient topology from the glued intervals) and use the notion of topological connectedness, or one can ask whether any two vertices are connected by some edge-path. Prove that these are equivalent.

Definition 5.6.8. A *tree* is a connected graph containing no *cycles* (subgraphs homeomorphic to S^1).

Lemma 5.6.9. *Any connected graph G has $\chi(G) \leq 1$, with equality if and only if G is a tree.*

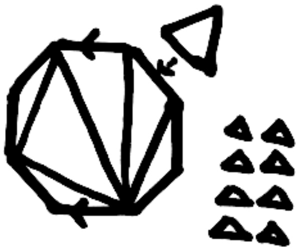
Proof. If a connected graph contains a vertex with degree 1 it may be *pruned* by removing that vertex and its incident edge (but not the vertex at the other end). The result is still connected (easy exercise), and has the *same* Euler characteristic. So apply pruning to G until one of two things happens: either all remaining vertices have degree 2 or more, or there is just one edge left (with two vertices of degree 1). (We cannot, strictly speaking, prune the last edge because an isolated vertex was not considered as a graph according our definition!) 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 cas, since the single-edge graph has Euler characteristic 1, so too did the original G. Rebuilding G by reversing the pruning sequence (budding?) one can easily check that there can be no cycles (also easy exercise). □

Exercise 5.6.10. Write down proofs of the two easy exercises just stated!

5.7. **Classification of surfaces.** The proof of the homeomorphism classification of closed connected combinatorial surfaces is actually based on a very simple idea: one simply looks for non-separating curves in a surface and does surgery on them, repeating until there are none left. A simple lemma shows that a surface with no non-separating curves is a sphere. Rebuilding the original surface by reversing the surgeries (just as we reverse the pruning in the above lemma) makes it easily identifiable. We will start with two technical lemmas and then two rudimentary classification lemmas before giving the main proof.

Lemma 5.7.1. *Any connected closed combinatorial surface F with f faces is homeomorphic to a regular polygon with $f + 2$ sides whose sides are identified in pairs (we represent this by arrows and labels as in a gluing pattern).*

Proof. Imagine the disjoint triangles $\amalg T$ out of which F is built all lying on the floor. Their edges are labelled in pairs indicating how to assemble them to make F. Pick up one starting triangle, and choose one of its edges: some distinct triangle glues on there (no face is glued to itself!), so pick this one up, attach it, and deform the result to a square. Now 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 is paired with an edge of one of the triangles still on the floor: 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. It finishes precisely when there are no free edges of the polygon left. At this stage there cannot be triangles remaining on the floor, or we could start again and end up with a completely separate component of F, 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. □



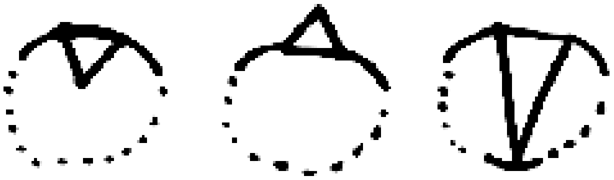
Lemma 5.7.2. *The Euler characteristic of a closed connected combinatorial surface is less than or equal to 2.*

Proof. Represent the surface F as an $(f + 2)$-gon P with identified sides, as above; call the quotient map $p : P \to F$. We will count the faces, edges and vertices of F by counting first those in $p(\partial P)$ and then the other ones, which are in one-to-one correspondence with those in the interior of P (because no identification goes on there). Since $p(\partial P)$ is a connected graph (it is a quotient of a connected polygon) it has Euler characteristic less than or equal to 1, by lemma 5.6.9. The interior of P has f faces, $f-1$ edges and no vertices , because they are all on the boundary of the polygon. Hence $\chi(F) = 0 - (f-1) + f + \chi(P) \leq 2$. □

Here is the first genuine classification lemma.

Lemma 5.7.3 (Recognising the disc). *Let F be a connected combinatorial surface with one boundary component, having the property that every arc in F separates F. Then F is homeomorphic to a disc, and $\chi(F) = 1$.*

Proof. The proof is by induction on the number of faces f. If $f = 1$ then obviously F is simply a triangle (with no self-gluing) and the result is true. In general, pick a boundary edge E of F and the unique triangle Δ incident at E. Now $\Delta \cap \partial F$ may consist of one edge, two edges or one edge and one vertex, as depicted in three configurations below.



We will only consider the first case, as the other two are very similar. Let γ be the arc in F consisting of the two edges of Δ other than E. Cutting along γ separates $F = \Delta \amalg F'$, where F' has $f-1$ faces. What we have to do is show that F' is a connected combinatorial surface with one boundary component and the separating arc property, for then it is (by inductive hypothesis) a disc with $\chi(F') = 1$, and F, which is the union of two discs along an arc, is itself a disc (by exercise 5.2.8) with $\chi(F) = 1$ (by trivial calculation), and we are done.

If F' were disconnected we could write a non-trivial disjoint decomposition $F' = F_1 \amalg F_2$. The triangle Δ would attach to this along a connected subset γ, so F would be disconnected, a contradiction.

If A is an arc in F' then it is also an arc in F, so cutting along it separates $F = F_1 \amalg F_2$. Then F' cut along A is obtained from this by removing a connected subset Δ, which must come WLOG from F_1. So F' is still disconnected unless $F_1 = \Delta$, which cannot happen unless $A = \gamma$, which is not a proper arc in F'. □

Lemma 5.7.4 (Recognising the sphere). *Let F be a connected closed combinatorial surface, having the property that every curve separates F. Then F is homeomorphic to a sphere, and $\chi(F) = 2$.*

Proof. Remove a single face Δ from F. Then what remains is a connected surface F' with one boundary component, and all we need to do is show that every arc in F' separates F' to conclude that it F' is a disc with $\chi = 1$, and therefore (adding back Δ) that F is a sphere (see example 5.2.9) and has $\chi = 2$. To do this, let A be an arc in F'; its endpoints must be two of the three boundary vertices of F', and so they span a unique edge e in $\partial F'$. Adding e to A gives a curve in F, which separates it (by assumption) non-trivially into $F_1 \amalg F_2$, only one of which can contain the removed triangle Δ, since this is connected. Suppose it is F_1; then $F_1 \neq \Delta$ because $A \not\subseteq \partial\Delta$, so $F' = (F_1 - \Delta) \amalg F_2$ is a non-trivial splitting of F', as required. □

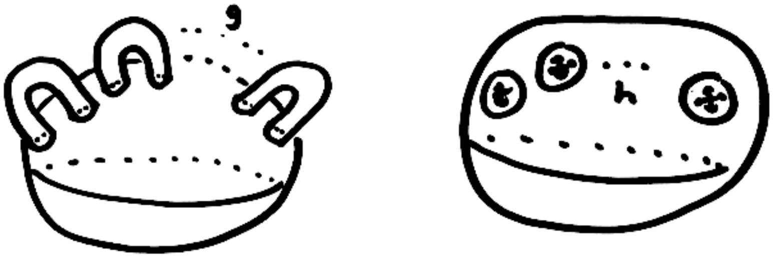
Corollary 5.7.5 (Characterisation of the sphere). *If F is a closed connected combinatorial surface then the three properties* (1) *every curve separates F* (2) *F is a sphere* (3) *$\chi(F) = 2$ are equivalent.*

Proof. Lemma 5.7.4 shows that (1) $\implies$ (2), (3). But (3) $\implies$ (1) because if there were a non-separating curve, we could do surgery on it and produce a connected surface with Euler characteristic 3 or 4, which contradicts the bound of lemma 5.7.2. And (2) $\implies$ (1) by the polygonal Jordan curve theorem, exercise 2.1.8. □

Theorem 5.7.6 (Classification of surfaces)**.** (1). *Any closed connected combinatorial surface F is homeomorphic to exactly one of the surfaces M_g ($g = 0, 1, 2, \dots$, a "sphere with g handles") or N_h ($h = 1, 2, 3, \dots$, a "sphere with h crosscaps") shown below.*

(2). *The Euler characteristic is an invariant of closed connected combinatorial surfaces – in other words, homeomorphic surfaces have the same Euler characteristic. A surface F homeomorphic to M_g has $\chi(F) = 2 - 2g$, and one homeomorphic to N_h has $\chi(F) = 2 - h$*

(3). *The Euler characteristic and orientability of a closed connected surface suffice to determine it up to homeomorphism – they form a "complete set of invariants" for such surfaces.*



Proof. The reduction part of the proof is best stated as an algorithm. We will construct a finite sequence of closed connected surfaces $F = F_0, F_1, \dots, F_k = S^2$, where each F_{i+1} is obtained from its predecessor F_i by surgery. Reversing direction, we will rebuild F starting from the sphere, and obtain the result.

To construct F_{i+1} from F_i, look at $\chi(F_i)$, which must be less than or equal to 2, by lemma 5.7.2. If $\chi(F_i) = 2$ then F_i is a sphere (and has no non-separating curves) by corollary 5.7.5, so we are finished (with $k = i$). If instead $\chi(F_i) < 2$; then F_i is not a sphere, so it must contain a non-separating curve C_i. Do surgery on C_i to produce a connected closed surface F_{i+1}, with $\chi(F_{i+1})$ greater than $\chi(F)$ by 1 or 2, depending on whether C_i is 1- or 2-sided. Because of the overall bound on Euler characteristic, the procedure must terminate in finitely-many steps.

To rebuild F 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 F is homeomorphic 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 visualisation exercises), such a surface is homeomorphic either to M_a (if $b, c = 0$) or to $N_{2a+2b+c}$ (if $b + 2c \geq 1$).

To show that the surfaces M_g, N_h ($g \geq 0, h \geq 1$) are pairwise distinct (so that the list of surfaces has no redundancy) it is easiest to use their fundamental groups. (Unfortunately these will not be properly defined and computed until the final chapter.) Homeomorphic spaces have isomorphic fundamental groups. So proving that the groups are pairwise non-isomorphic is enough to show that the spaces are pairwise non-homeomorphic. The fundamental groups themselves are described in exercises 7.3.6, 7.3.9 and the proof that no two are isomorphic is exercise 7.3.14. This finishes part (1).

For part (2): the above process gives such an explicit way of reconstructing F from a combinatorial sphere (whose Euler characteristic we know to be 2) that we can reconstruct its Euler characteristic too. 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 $\chi(F) = 2 - 2a - 2b - c$. But if $F \cong M_g$ then $c = 0, a = g$ and hence $\chi(F) = 2 - 2g$, whilst if $F \cong N_h$ then $h = 2a + 2b + c$ so that $\chi(F) = 2 - h$.

Part (3) is then just the observation that from the orientability of a surface we can determine whether it is an 'M' or an 'N', and then having established that, the Euler characteristic tells us what is the value of g or h. □

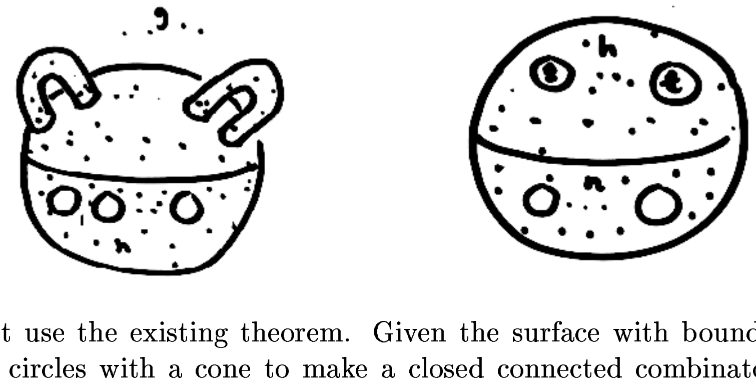
Remark 5.7.7. Surfaces with odd Euler characteristic must be non-orientable, since $2 - 2g$ is always even. In this case, the Euler characteristic on its own is enough to identify the surface.

Remark 5.7.8. The *genus* g of a closed surface F is defined by $g(F) = 1 - \frac{1}{2}\chi(F)$ for an orientable surface and $g(F) = 2 - \chi(F)$ 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 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 5.7.9 (Classification of surfaces with boundary)**.** (1). *A connected combinatorial surface with $n \geq 1$ boundary components is homeomorphic to exactly one of the surfaces M_g^n ($g = 0, 1, 2, \dots$) or N_h^n ($h = 1, 2, 3, \dots$) shown below. (The number g or h is called the* genus.*)*

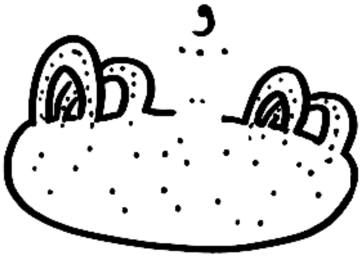
(2). *The Euler characteristic is an invariant for surfaces with boundary, and $\chi(M_g^n) = 2 - 2g - n$, $\chi(N_h^n) = 2 - h - n$ (and conversely, $g = 1 - \frac{1}{2}(\chi + n)$ and $h = 2 - (\chi + n)$).*

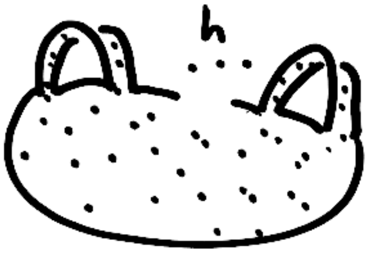
(3). *The number of boundary components, Euler characteristic and orientability form a complete set of invariants for connected combinatorial surfaces.*



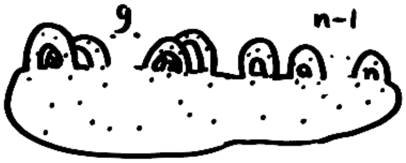
Proof. We can just use the existing theorem. Given the surface with boundary F, cap off each of its n boundary circles with a cone to make a closed connected combinatorial surface $\hat{F}$ with $\chi(\hat{F}) = \chi(F) + n$. This $\hat{F}$ must be homeomorphic to one of the M_g or N_h, with $\chi(\hat{F}) = 2 - 2g, 2 - h$ accordingly. Therefore F is one of these surfaces with n open discs removed, and has the asserted Euler characteristic. Obviously these surfaces are pairwise non-homeomorphic, since the number of boundary components and the homeomorphism type of the closed-up surface are homeomorphism invariants. The final part is then obvious. □

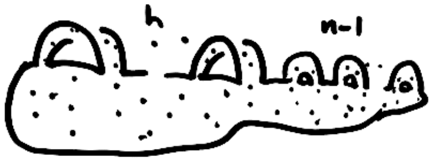
Exercise 5.7.10. Show that any compact connected orientable surface with one boundary component is homeomorphic to one of the following surfaces.



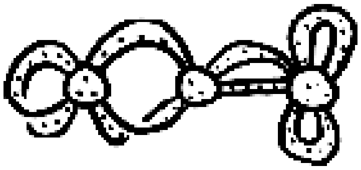


Exercise 5.7.11. Show that any compact connected surface with boundary is homeomorphic to one of the following surfaces.

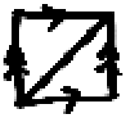


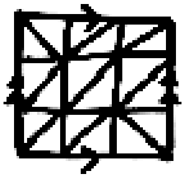


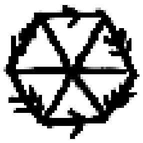
Exercise 5.7.12. Suppose that a connected surface F is made by starting with v closed discs and attaching e bands to them, as in the example. Prove that $\chi(F) = v - e$. What does the formula suggest to you?

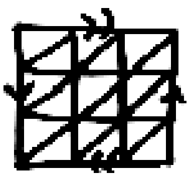


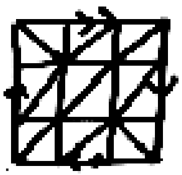
Exercise 5.7.13. Which of the following figures represents a combinatorial surface, and why? Use the classification theorem to identify those that are. (Each picture represents a gluing pattern of triangles, where most of the gluing has been performed already, and only the edges remain to be identified. In the square pictures, the *whole* sides are to be glued according to the arrows.)





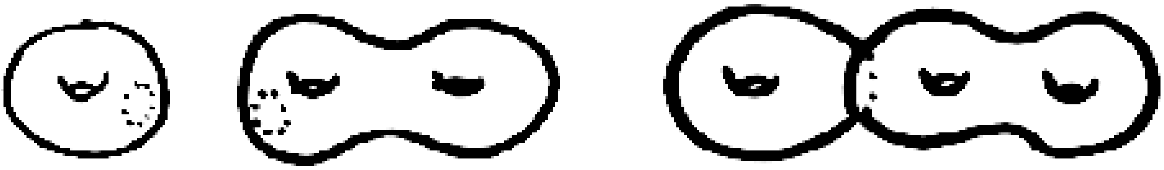






Exercise 5.7.14. Identify the following surfaces.

Exercise 5.7.15. Define the *connected sum* $F_1 \# F_2$ of connected combinatorial surfaces F_1, F_2 to be the surface made by removing an open face from each and gluing the resulting boundary triangles together. Show that $\chi(F_1 \# F_2) = \chi(F_1) + \chi(F_2) - 2$ 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 5.7.16. Show (using Euler characteristic and the classification theorem) that cutting a sphere along a curve always results in two discs.

Remark 5.7.17. For closed connected 2-manifolds F we have shown that every closed curve separates F if and only if F is homeomorphic 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 homeomorphic to the n-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 trivial". Poincaré's manifold actually has a fundamental group with 120 elements called the *binary icosahedral group* whose abelianisation is trivial.

Consequently Poincaré reformulated his conjecture with a stronger hypothesis by just dropping the word "abelianisation":

Every closed connected 3*-manifold with trivial fundamental group is homeomorphic to the* 3*-sphere.*

Amazingly, the truth of this assertion is still unknown: the Poincaré conjecture is one of the great unsolved problems in mathematics (though for various reasons, most topologists seem to believe it is true).

6. Surfaces and knots

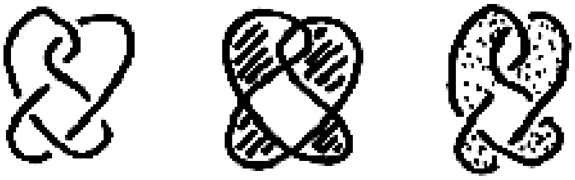
We are now going to use surfaces to study knots, so from now on they will tend to be embedded in $\mathbb{R}^3$. This certainly helps to visualise them, but remember that the way a surface is tangled inside $\mathbb{R}^3$ does not affect its homeomorphism type. All surfaces will be assumed to be combinatorial, despite being drawn "smoothly".

6.1. **Seifert surfaces.**

Definition 6.1.1. 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

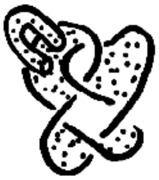
Lemma 6.1.2. *Any knot* K *bounds some surface* F.

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 6.1.3. Why is it possible to chessboard-colour a knot projection in two colours, as we did above?

Remark 6.1.4. 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.

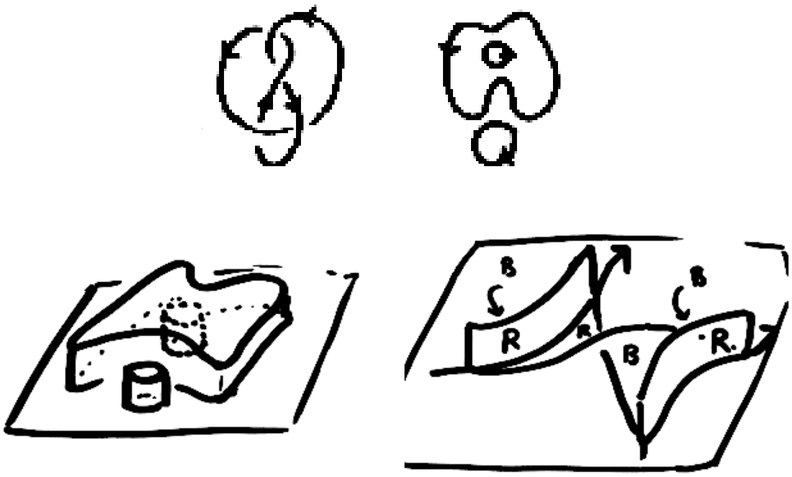


One problem with this construction is that the resulting surface may be non-orientable, which makes it harder to work with. Fortunately we can do a different construction which always produces an orientable surface.

Definition 6.1.5. A *Seifert surface* for K is just a connected orientable surface in $\mathbb{R}^3$ bounded by K.

Lemma 6.1.6. *Any knot has a Seifert surface.*

Proof. (Seifert's algorithm.) Pick a diagram of the knot and choose an orientation on it. Smooth all the crossings in the standard orientation-respecting way to obtain a disjoint union of oriented circles, called *Seifert circles*, in the plane. The idea is that if we make each of these circles bound a disc, and connect them with half-twisted bands at the crossings (just like the previous construction joined up the chessboard regions) then the result will be orientable. 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 the Seifert circles using their orientation, colouring each cylinder red on the right-hand side and blue on the left-hand side, extending this colouring onto the top disc. (This makes the upper side of the disc red if the circle is anticlockwise, blue if clockwise.) Then clearly at each crossing the half-twisted band connects like-coloured sides of the surface. □

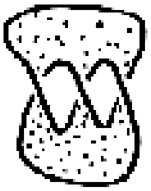


Because of this theorem we can immediately define a useful new invariant of knots.

Definition 6.1.7. The *genus* $g(K)$ of a knot K is the minimal genus of any Seifert surface for K.

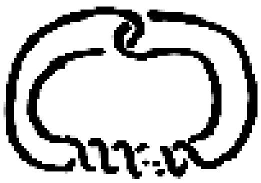
Example 6.1.8. A knot has genus 0 if and only if it is the unknot. This is because having genus 0 is equivalent to bounding a disc in $\mathbb{R}^3$. If a knot bounds a disc, the triangles making up the disc give a sequence of Δ-moves that deform the knot down to a single triangle.

Example 6.1.9. 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.



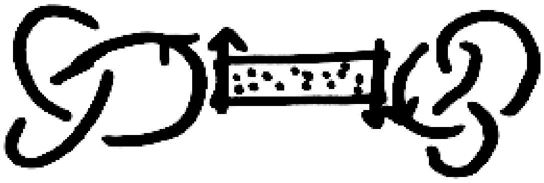
Exercise 6.1.10. By viewing the Seifert surface constructed from Seifert's algorithm as a disc-and-band surface (exercise 5.7.12), show that the genus of any knot is bounded in terms of its crossing number by the formula $g(K) \leq c(K)/2$.

Exercise 6.1.11. Show that all the knots in the family of twisted doubles of the unknot shown below have genus 1.



6.2. **Additivity of the genus.**

Definition 6.2.1. 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 6.2.2. 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 6.2.3. 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 6.2.4. *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 F_1, F_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 $F_1 \amalg F_2$ 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

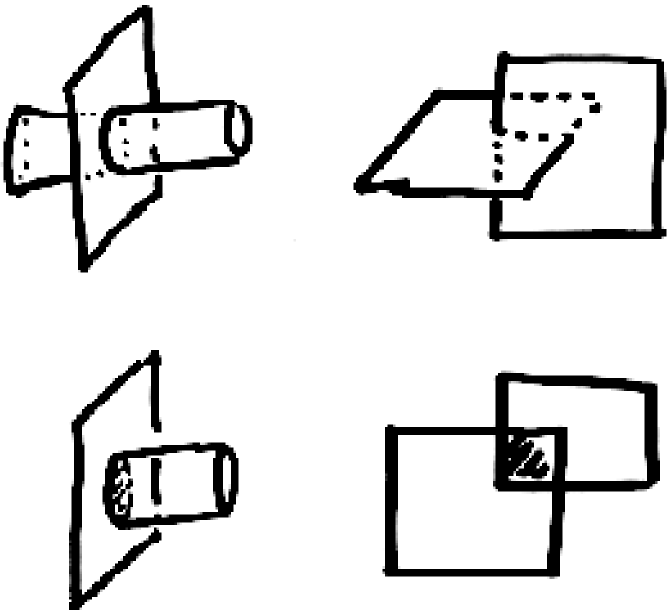
$$\chi(F) = \chi(F_1) + \chi(F_2) + 1 - 1 - 1,$$

because the Euler characteristic of the band is 1, and its intersection with $F_1 \amalg F_2$ 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 $g(F) = g(F_1) + g(F_2)$, and hence

$$g(K_1 \# K_2) \leq g(F) = g(F_1) + g(F_2) = g(K_1) + g(K_2).$$

($\geq$). 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 6.2.5. 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 F_1 and F_2 are two surfaces contained in $\mathbb{R}^3$, then by an arbitrarily small perturbation of F_2 (say) we can arrange that F_1, F_2 *meet transversely*: that $F_1 \cap F_2$ consists of a union of circles disjoint from the boundaries of both surfaces, and arcs whose interiors are disjoint from ther 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.

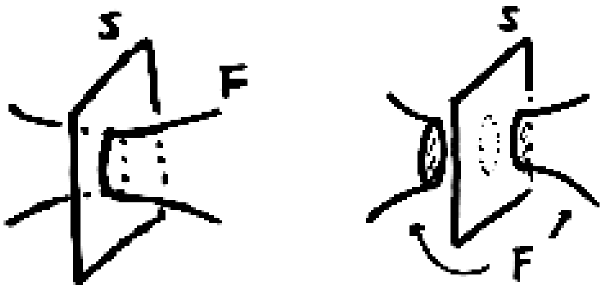


Recall then that F is a *minimal genus* Seifert surface for $K = K_1 \# K_2$, and let S be a separating sphere (remark 6.2.3). Let us make S and F transverse, as explained above. Then their intersection is a union of circles and a single arc, which runs between the two points of $K \cap S$. (All arcs 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 5.7.16). 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, beacuse 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

$$\chi(F) + 2 = \chi(F') = \chi(F'') + \chi(X).$$

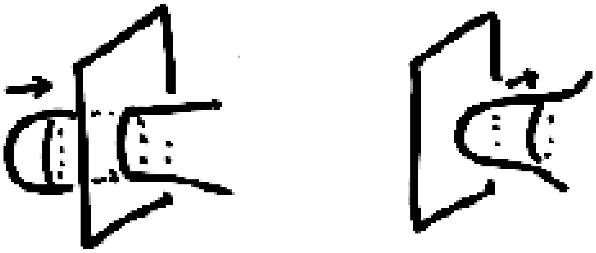
But $\chi(X) \leq 2$ (by lemma 5.7.2) and again by minimality of F, $\chi(F'')$ cannot be bigger than $\chi(F)$. Hence $\chi(X) = 2$, X is a sphere, $\chi(F'') = \chi(F)$, and so F and F'' have the same genus. Note that $F'' \cap S$ is some proper subset of $F \cap S$; at least one (possibly more, when we throw out X) circles of intersection have been eliminated.

Repeat this procedure 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 $F_1 \amalg F_2$ (one part inside the sphere S and one part outside), where F_1, F_2 are connected orientable surfaces with boundaries 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:

$$g(K_1) + g(K_2) \leq g(F_1) + g(F_2) = g(G) = g(F) = g(K_1 \# K_2).$$

□

Remark 6.2.6. 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 6.2.7. Use genus to show that there are infinitely-many distinct knots.

Definition 6.2.8. 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 6.2.9. Show that any genus-1 knot is prime.

Corollary 6.2.10. *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 6.2.11. *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 6.2.12. 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!).

7. 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.

7.1. Presentations of groups.

Definition 7.1.1. 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 7.1.2. 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 7.1.3. *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 7.1.4. *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 7.1.5. *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 7.1.6. 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 7.1.7. 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:

$$1; \quad a, b, a^{-1}, b^{-1}; \quad ab, ab^{-1}, a^2, ba, ba^{-1}, b^2, a^{-1}b, a^{-1}b^{-1}, a^{-2}, b^{-1}a, b^{-1}a^{-1}, b^{-2}; \quad \ldots$$

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 7.1.8. Show that the alternating group A_4 has a presentation

$$\langle a, b : a^2 = 1, b^2 = 1, (ab)^3 = 1\rangle.$$

(Define a map from the group with this presentation to A_4 and check that it's an isomorphism.)

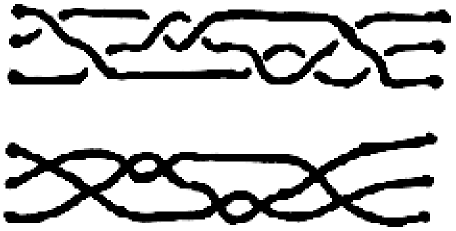
Exercise 7.1.9. Show that the symmetric group S_3 has a presentation

$$\langle a, b : a^2 = 1, b^2 = 1, aba = bab\rangle.$$

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 picture is relevant!



Remark 7.1.10. 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).)

7.2. **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.

Definition 7.2.1. 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 7.2.2. 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 7.2.3. (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 7.2.4. (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 7.2.5. 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 7.2.6. *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. □

7.3. **Van Kampen's theorem.** The statement of this theorem is rather long-winded, but it's easier than it sounds:

Theorem 7.3.1. *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 7.3.2. 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.

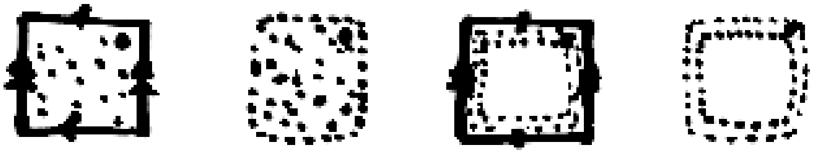
Example 7.3.3. Let X be the join of two circles. Let U (V) be the left (right) circle union a small open neighbourhood of the vertex. Each of U, V is homotopy-equivalent to its circle (shrink the extra bits). Then W is a small open cross shape, which is contractible. We can take the presentations $\langle a \rangle, \langle b \rangle$ 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 7.3.4. 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 7.3.5. 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 7.3.6. 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)

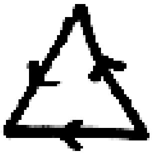
$$a_1 b_1 a_1^{-1} b_1^{-1} a_2 b_2 a_2^{-1} b_2^{-1} \cdots a_g b_g a_g^{-1} b_g^{-1}.$$

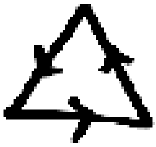
(Using this gluing scheme certainly gives an orientable surface, by the "circulation" argument of exercise 5.5.3. It also makes all vertices of the polygon equivalent, and therefore the Euler characteristic of the resulting closed surface, which consists of the disjoint union of an open disc, a vertex and $2g$ edges is $1 - 2g + 1 = 2 - 2g$, proving that this surface is M_g.) Applying exactly the same method as above gives

$$\pi_1(M_g) \cong \langle a_1, b_1, a_2, b_2, \ldots, a_g, b_g : \prod_{i=1}^{g} [a_i, b_i] = 1\rangle,$$

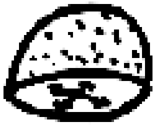
where $[a, b]$ denotes the *commutator* $aba^{-1}b^{-1}$.

Exercise 7.3.7. 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.





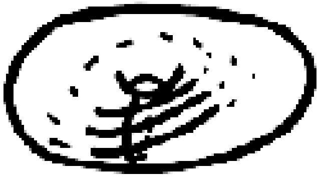
Exercise 7.3.8. Compute the fundamental group of the projective plane (shown below as a hemisphere with antipodal boundary points identified) by applying van Kampen's theorem.



Exercise 7.3.9. Show that the non-orientable surface N_h has a fundamental group

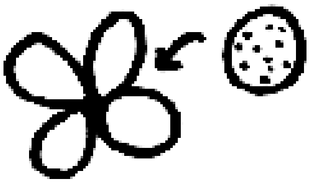
$$\pi_1(N_h) \cong \langle a_1, a_2, \dots , a_h : \prod_{i=1}^{h} a_i^2 = 1\rangle,$$

Exercise 7.3.10. 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 7.3.11. 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 7.3.12. 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 $\langle a_1, \dots , a_g : w_\gamma =\rangle >$.



Exercise 7.3.13. 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 7.3.14. 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 7.3.15. 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 7.3.16. 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 $\pi_1(\Sigma) \to \mathbb{Z}$.

7.4. **The knot group.**

Definition 7.4.1. 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 7.4.2. 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 7.4.3. (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.

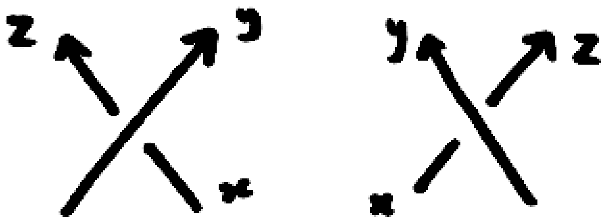
(2). "The knot group is an invariant of knots", because equivalent knots have homeomorphic complements which therefore have isomorphic fundamental groups (it is the isomorphism class of the group which is really considered as the invariant here.)

(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 7.4.8 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 7.4.10 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.

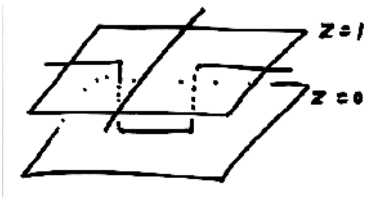
Theorem 7.4.4 (The Wirtinger presentation)**.** *A presentation of $\pi(K)$ may be obtained as follows. Take a diagram D of the knot and orient it. Label the arcs $a_1, a_2, \ldots a_k$, and let $S = \{a_1, a_2, \ldots a_k\}$. At each (signed) crossing one sees three incident labels x, y, z as shown below.*



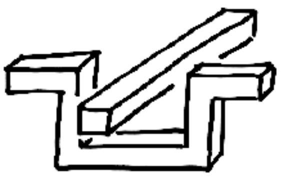
To each positive crossing associate the relation $x^y = z$ and to each negative crossing $x^{\bar{y}} = z$ to obtain a set of relations R. Then $\pi(K) \cong \langle S : R \rangle$.

Proof. The proof is basically just van Kampen's theorem, although I will not appeal directly to it below. Consider the knot to lie in the plane $z = 1$ inside $\mathbb{R}^3$, except in a small neighbourhood of the crossings, where one arc makes a small rectangular detour downwards into the plane $z = 0$. Let us consider "dragging a small open cube" along the knot, to make an open solid torus neighbourhood N of K (with a square cross-section), and consider the knot complement X as being the closed subset $\mathbb{R}^3 - N$. Decompose X into the parts $X_+ = X \cap \{z \geq 0\}, X_- = X \cap \{z \leq 0\}$. Their intersection W is a plane minus some small open squares, two per crossing of the knot. The part X_- is a half-space minus some little rectangular trenches, one per crossing, whilst X_+ is a halfspace minus an open solid cylinder, one per arc of the diagram. Now X_+ is homotopy-equivalent to a bouquet of k circles. In fact, pick a basepoint high up on the z-axis and drop a loop from it to hook under each borehole in X_+, adding an orientation so that it goes under from right to left (in terms of the orientation of the arc that made the hole). Name the homotopy classes of these loops after their arcs, so that we have an isomorphism $\pi_1(X_+) \cong \langle a_1, a_2, \ldots, a_k \rangle$. Now X_- is homotopy equivalent to the punctured plane W union the faces of the trenches, via a more-or-less vertical retraction. Thus, attaching X_- to X_+ is in homotopy terms the same as attaching k discs to a bouquet of spheres (see exercise 7.3.13). Each such attachment adds a relation, which says that the homotopy class of the attaching loop in X_+ becomes trivial. All we need to do is identify this loop in terms of the generators $\langle a_1, a_2, \ldots, a_k \rangle$ to finish. The final picture below shows how the conjugation relation arises.

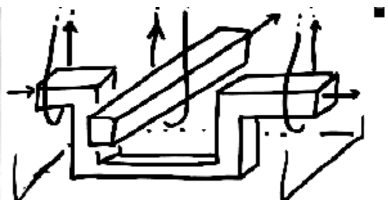
A crossing:



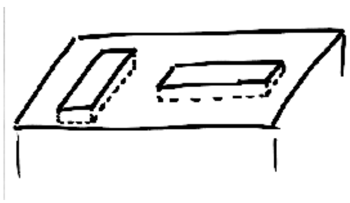
Neighbourhood N:



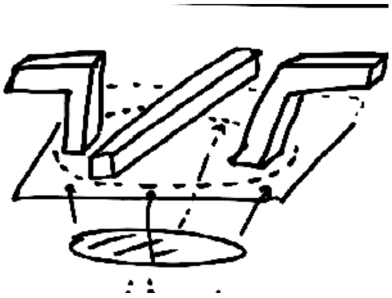
X_+ :



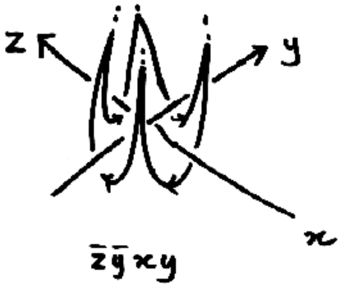
X_- :



Attaching loop:



Relator:

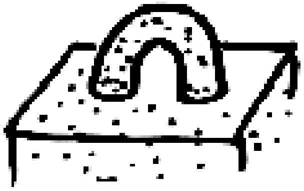


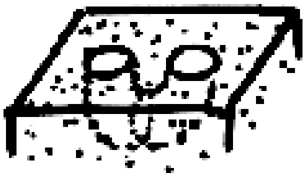
□

Example 7.4.5. Applying the theorem to the standard picture of the right trefoil (the one whose writhe is +3) gives the presentation

$$\langle x, y, z : x^y = z, y^z = x, z^x = y\rangle.$$

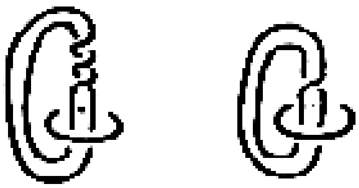
Exercise 7.4.6. Let X be the closed upper half-space with g handle-shaped holes removed from it, and let Y be the same space with g solid handle-shaped protrusions added to it. Show that these spaces are homeomorphic, and further that they are both homotopy-equivalent to a bouquet of g circles.





Exercise 7.4.7. 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 7.4.8. 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 7.4.9. Show that the knot groups of any knot and its mirror-image are isomorphic (explaining the problem with Whitten's result).

Exercise 7.4.10. (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 7.3.14 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 7.4.11. *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 7.1.5, 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 7.4.12. If G is any finite group then we can define an *invariant* of finitely-presented groups $\lambda(-, G)$ by $\lambda(\pi, G) = |\,\mathrm{Hom}(\pi, G)|$; 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 7.4.13. 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 7.4.14. 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 7.4.15. 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 7.4.4. 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 7.4.16. 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 7.4.17. *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 7.4.18. Show that $\lambda(K, S_3, C) = 3 + \tau(K)$.

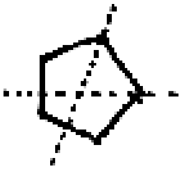
Exercise 7.4.19. Suppose A is a finite abelian group. Show that the number of labellings $\lambda(K, A)$ equals the order of A, regardless of the knot K.

Exercise 7.4.20. How many conjugacy classes are there in the symmetric group S_5, and how many elements are there in each?

Exercise 7.4.21. 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

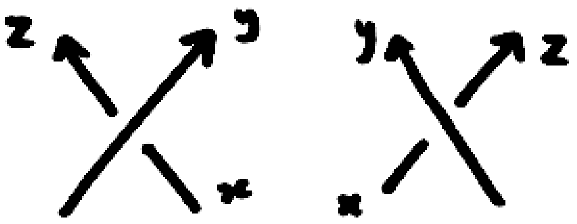
$$R_\theta^{-1} R_\phi R_\theta = R_{2\theta-\phi}.$$

(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 7.4.22. 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 7.4.23. Show that $\lambda(K, G)$ does not depend on the orientation of the knot.

Exercise 7.4.24. Compute the number of labellings of the trefoil knot by 3-cycles from the symmetric group S_4.

Exercise 7.4.25. Show that the abelianisation of any knot group π (see exercise 7.3.15) is isomorphic to $\mathbb{Z}$.

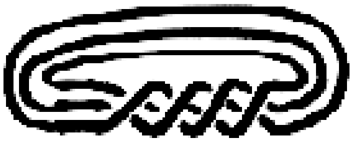
Exercise 7.4.26. Using the notation $x^y = y^{-1}xy$ for conjugation, show that

$$(x^y)^z = x^{yz} \quad \text{and} \quad (x^y)^{(z^y)} = x^{zy}.$$

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?



Department of Mathematics and Statistics, Edinburgh University, EH3 9JZ, Scotland

E-mail address: justin@maths.ed.ac.uk