

KNOTS AND SURFACES

NEIL STRICKLAND

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike license.



1. INTRODUCTION TO KNOTS

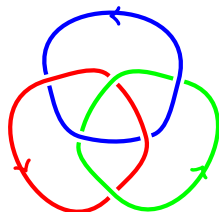
The first half of this course is about the mathematical theory of knots, and especially an invariant called the Jones polynomial.

To explain what is meant by a mathematical knot, consider the two pictures below.

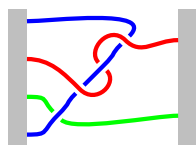


Both pictures show curves embedded in \mathbb{R}^3 . The left hand picture is not considered to be a knot, because it can be untied, but the right hand picture shows a knot.

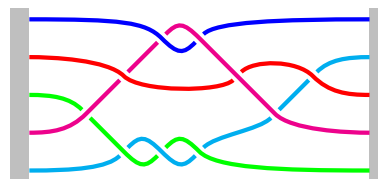
This is one of several closely related concepts, which are illustrated by the pictures below.



a link

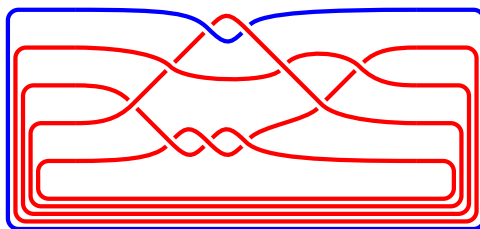


a tangle



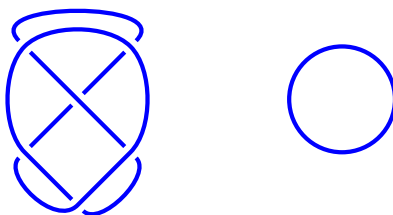
a braid

The first hand picture shows a *link*, which is like a knot but which may have more than one strand. (Knots are considered to be a special case of links.) The second picture shows a *tangle*, which is like a link, but the strands have ends, which are fixed to walls on the left or the right. The third picture shows a *braid*, which is a special kind of tangle in which all run from left to right without ever curling backwards. Braids are nice because if we have two braids with n strands then we can join them together to make a new braid with n strands, and this operation makes the set of n -stranded braids into a group. We can convert braids into links by joining the left ends to the right ends in an obvious pattern, as illustrated below:

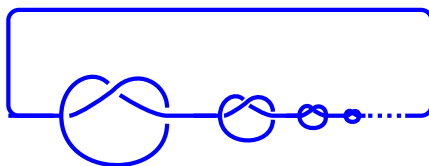


This course will focus on knots and links, with occasional comments on tangles and braids.

We will consider two links to be equivalent if one can be deformed into the other. For example, the knot shown on the left below can easily be deformed into an unknotted circle as shown on the right, so we will not distinguish between them.



We next discuss how to make our definitions more formal. There are two problems that we need to work around. The first issue is that there are things that are similar to knots but have infinite complexity, like this:



(This is supposed to show an infinite sequence of knotted loops, where the n 'th loop has size 2^{-n} .) We only want to study knots of finite complexity, so we should arrange our definitions so that the above picture does not count as a knot.

The second issue is closely related. Suppose we have a knot in a thin piece of cotton thread. If we pull it tight, the knot will disappear, becoming a small bump on the thread. Mathematical knots are considered to be made from infinitely thin thread, so if we pull them tight, there will not even be a bump. We should arrange our definitions so that knots cannot just disappear like this.

Unfortunately, a really complete and rigorous account of the definitions would take a long time, and would not shed that much light on the main questions of interest in this course. We will therefore just give an indication of the main points.

Definition 1.1. If a and b are points in \mathbb{R}^3 with $a \neq b$, we put

$$[a, b] = \text{the line segment between } a \text{ and } b = \{(1-t)a + tb \mid 0 \leq t \leq 1\} \subset \mathbb{R}^3.$$

Now consider a subset $L \subset \mathbb{R}^3$. We say that L is a *piecewise linear link* if it can be written as a finite union of line segments as above, say $L = S_1 \cup \dots \cup S_n$, in such a way that

- (a) If $i \neq j$ then either $S_i \cap S_j = \emptyset$, or there is a point x which is an endpoint of S_i and also an endpoint of S_j , such that $S_i \cap S_j = \{x\}$.
- (b) For each i , and for each endpoint x of S_i , there is precisely one other index j such that x is also an endpoint of S_j .

In other words, a piecewise linear link is a link of the type that we have illustrated previously, except that the strands can be divided into a finite number of straight sections. We would like to say that any link (of finite complexity) can be deformed into a piecewise linear one. For this, we need a formal definition of the kind of deformation that we want to consider.

- Definition 1.2.**
- (a) A *homeomorphism* of \mathbb{R}^n is a bijective map $f: \mathbb{R}^n \rightarrow \mathbb{R}^n$ such that both f and f^{-1} are continuous.
 - (b) Suppose we have a family of homeomorphisms $f_t: \mathbb{R}^n \rightarrow \mathbb{R}^n$ for $0 \leq t \leq 1$, and we define maps $h, h^*: [0, 1] \times \mathbb{R}^n \rightarrow \mathbb{R}^n$ by $h(t, x) = f_t(x)$ and $h^*(t, x) = f_t^{-1}(x)$. We say that the maps f_t form an *isotopy* of \mathbb{R}^n if h and h^* are continuous, and $f_0(x) = x$ for all x .
 - (c) Let X and Y be two subsets of \mathbb{R}^n . We say that they are *ambiently isotopic* if there is an isotopy as in (b) with $f_1(X) = Y$. (It is not hard to check that this is an equivalence relation.)

- (d) A *link* is a subset of \mathbb{R}^3 that is ambiently isotopic to a piecewise linear link. We say that two links are *equivalent* if they are ambiently isotopic.

2. WHY STUDY KNOTS?

Firstly, knots are a good introductory example of topological phenomena. There are many examples in mathematics where objects can be deformed continuously in infinitely many different ways, but there are some discrete properties that always remain the same, such as the number of strands, and certain aspects of the way they twist around each other. The theories of algebraic topology, differential topology and homotopy theory have a huge literature devoted to this kind of phenomenon. Knot theory is in some respects the simplest example that one can study. Moreover, knot theory can also be used indirectly to shed light on other kinds of topological questions. For example, there is a construction called *Dehn surgery* which allows one to create new kinds of three-dimensional spaces by taking \mathbb{R}^3 and “twisting it around a knot”.

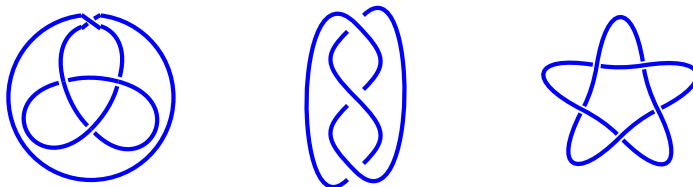
Next, there are some unexpected connections with other areas of pure mathematics. As we mentioned previously, braids with n strands form a group, called Braid_n . For various reasons it is interesting to consider homomorphisms $\rho: \text{Braid}_n \rightarrow GL_d(\mathbb{C})$ (where d is a natural number, and $GL_d(\mathbb{C})$ is the group of invertible $d \times d$ matrices over the complex numbers); these are called *linear representations* of the braid group. It turns out that the representation theory of the braid group is closely related to knot theory, and also to some questions in statistical physics and functional analysis. A significant part of this course will be devoted to studying the Jones polynomial of a knot. Historically, this was first invented by Vaughan Jones as a byproduct of his work in these areas; he was not originally studying knots at all.

Finally, there are a number of places where knotting occurs in nature. Strands of DNA sometimes become knotted, and biologists have investigated what this tells us about the means by which DNA is manipulated in cells. Magnetic field lines can be knotted, particularly in the extreme magnetic conditions that can arise in astrophysics.

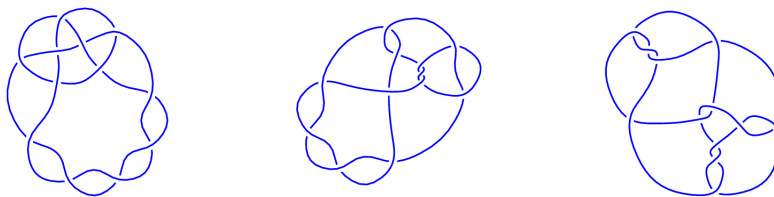
In the late nineteenth century there was actually a popular theory, proposed by the physicist Kelvin, that atoms were actually knotted structures in a hypothetical substance called aether that was thought to carry light waves. This is an attractive theory, because it would explain the discrete series of different elements in terms of a discrete series of different knot types. Although it did not turn out to be correct, it did inspire some foundational work in knot theory. Moreover, there are echoes of Kelvin’s idea in the much newer physical theory of “strings”. This theory (which may or may not be on the right track) aims to unify all the fundamental forces of nature, and has been the focus of an enormous body of work by physicists over several decades. It also rests heavily on topological ideas, starting with the theory of surfaces which forms the second half of this course (but continuing well beyond that).

3. INVARIANTS

Consider the following knots:

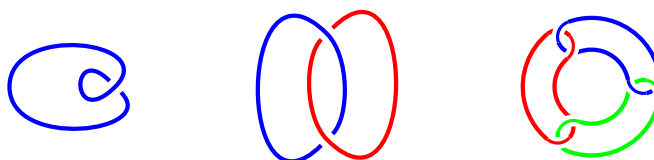


You can probably convince yourself (with the aid of string if necessary) that the first two are equivalent to each other, but not to the third one. But what about these three?



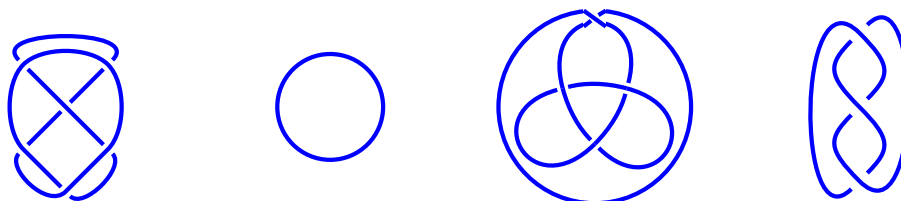
There is no obvious way to deform any of them into any of the others, but how can we tell if there might be a non-obvious way? For this, we need numerical invariants which we can calculate.

For the simplest example, note that any link can be divided into separate strands, each of which is essentially a circle. These strands are called the *components* of the link, and we write $c(L)$ for the number of components. The following pictures show a link with one component, a link with two components and a link with three components.



Suppose we have two links L and L' , and we want to decide whether they are equivalent. We can start by finding $c(L)$ and $c(L')$ (which is quite easy to do, using any planar picture). If $c(L) \neq c(L')$ then L and L' are definitely not equivalent. However, if $c(L) = c(L')$ then we have not learned very much; L and L' might or might not be equivalent. We therefore need to look for better invariants.

As an example of something that does **not** work, consider the crossing number. Given a planar picture D of a knot, we let $n(D)$ denote the number of crossings. To see what is wrong with this, consider the following pictures:



Pictures 1 and 2 show equivalent knots, but picture 1 has 6 crossings and picture 2 has no crossings. Similarly, pictures 3 and 4 show equivalent knots, but picture 3 has 4 crossings and picture 3 has 3 crossings. Thus, the crossing number is not a well-defined invariant.

We could instead define $n^*(L)$ to be the *minimal crossing number* of L , or in other words the minimum possible number of crossings in any planar picture of L . This is an invariant, but it is very hard to calculate. If someone gives us a link L , there is no obvious way to list all the possible ways to draw L in the plane, so we cannot tell how many crossings are needed.

Our aim will be to define an invariant which is quite powerful, and also quite easy to compute.

4. REIDEMEISTER MOVES

We have already drawn many planar pictures of knots and links. Before going further, we need to discuss in more detail how such pictures work. Our pictures have always had little gaps next to each crossing,

to indicate which strand lies on top. However, we sometimes need to consider pictures without this extra information. These are covered by the definition below.

Definition 4.1. A *link universe* is a planar picture, consisting of

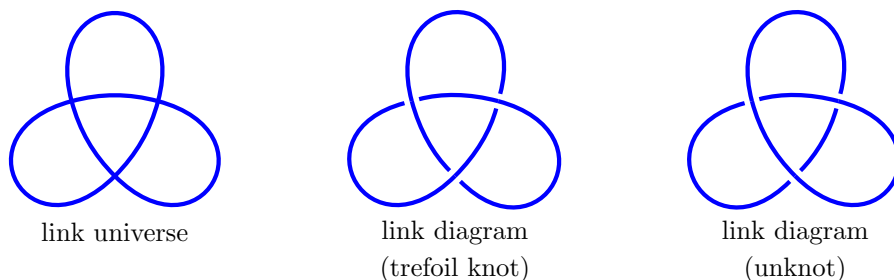
- (a) A finite set of points in the plane, called *crossings*
- (b) A set of curves in the plane, called *arcs*.

These must satisfy the following axioms:

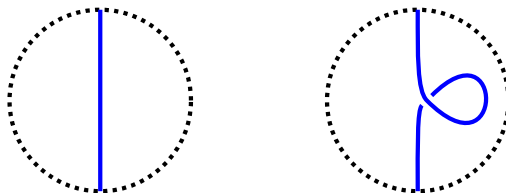
- (c) Each arc starts at a crossing, and ends at a crossing (possibly the same one).
- (d) The arcs are disjoint, except that they may intersect at the endpoints.
- (e) For each crossing, there are precisely four half-arcs with that crossing as an endpoint.
- (f) The whole picture is ambient isotopic to a piecewise linear one.

Definition 4.2. A *link diagram* is a link universe with a choice, for each crossing, of which opposite pair of strands lies on top.

Example 4.3. The left hand picture below is a link universe. It has three crossings, and at each crossing there are two ways to choose what goes on top, so there are $2 \times 2 \times 2 = 8$ possible link diagrams for this universe. Two of them are shown. The middle picture is a diagram for the trefoil knot. The right hand picture is the same as the middle picture, except that the bottom crossing has been switched over. This allows us to deform the corresponding link into an unknotted circle.

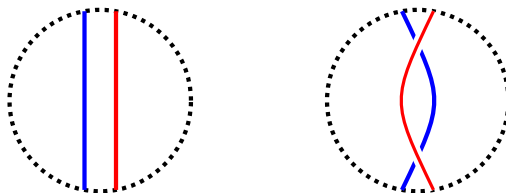


We next need to discuss the appropriate notion of equivalence for link diagrams. Suppose that we have two link diagrams that are mostly the same, except that they differ in a small disc as illustrated below.

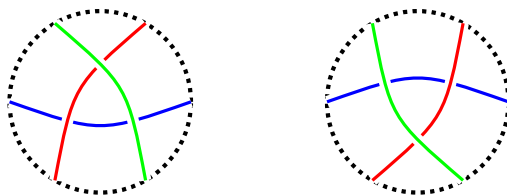


One diagram just has a strand cutting directly across the disc, and no other strand touches the disc. The other diagram is the same, except that there is an additional twisted loop in the middle of the disc. Adding or removing a twist like this is called *Reidemeister move 1*. Clearly, performing this move on the planar diagram does not change the equivalence class of the actual link.

There are two other kinds of Reidemeister move, which should be interpreted in a similar way. Move 2 pushes one strand under another, or does the reverse:



Reidemeister move 3 slides a strand under a crossing:



We can also distort a diagram by an isotopy of \mathbb{R}^2 ; this is called Reidemeister move 0. To remember the numbering, note that Move 1 involves one strand and one crossing, move 2 involves two strands and two crossings, and move 3 involves three strands and three crossings.

Definition 4.4. Two diagrams are R-equivalent if they can be converted to each other by a sequence of Reidemeister moves of types 0, 1, 2 or 3.

Theorem 4.5. Let L and L' be links, and let D and D' be planar pictures of L and L' . Then L and L' are equivalent if and only if D and D' are R-equivalent.

Proof. One half of this is straightforward. If D and D' are related by a single Reidemeister move, then it is clear that L and L' are equivalent. It follows by induction that if D and D' are R-equivalent, then L and L' must be equivalent.

The converse is harder, and we will not give a formal proof. However, the basic idea is quite simple. If you just watch the shadow of a link as it moves around and distorts in 3-dimensional space, you will usually just see Reidemeister moves happening one at a time. Occasionally something different might happen: for example, you might see shadows of six strands appearing to cross in exactly the same place. However, such strange phenomena can only appear if you watch the moving link from exactly the right angle. If you adjust your viewpoint slightly, then you will just see ordinary Reidemeister moves. \square

5. THE JONES POLYNOMIAL AND THE SKEIN RELATION

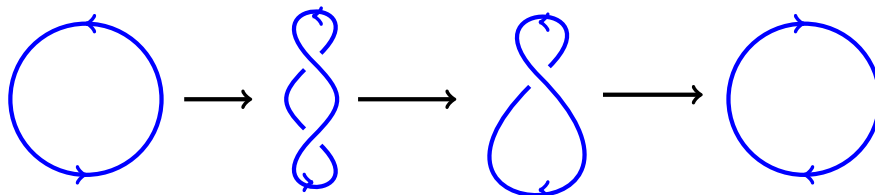
Before introducing the Jones polynomial, we need a few more preliminary ingredients.

Definition 5.1. An *orientation* for a link (or for a corresponding link diagram) is a choice of direction along each of the components of the link.

We can exhibit an orientation by drawing arrows on the arcs, as illustrated below.



Remark 5.2. Suppose that D is an oriented link diagram, and that D' is obtained from D by applying a Reidemeister move. Then the components of D' correspond in an obvious way to the components of D , so we can transfer the orientation of D to get an orientation of D' . The same applies (by induction) if D' is obtained from D by applying a sequence of Reidemeister moves. This gives a version of R-equivalence for oriented diagrams. For example, a circle with clockwise orientation is R-equivalent to a circle with anticlockwise orientation, by the following sequence of moves:



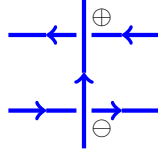
(The first move is of type 2, and the second and third moves are of type 1.).

Definition 5.3. Crossings in an oriented link diagram are classified as positive or negative by the following rule. We imagine approaching the crossing on the upper strand, in the direction indicated by the orientation, and watching the lower strand pass underneath.

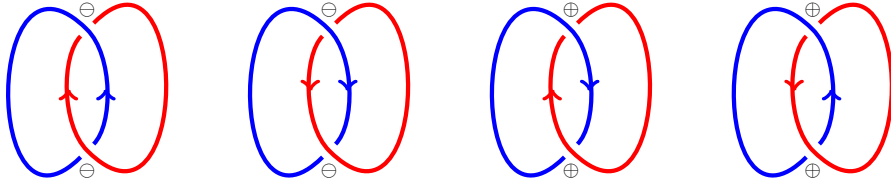
- If the lower strand passes from right to left, then the crossing is positive.
- If the lower strand passes from left to right, then the crossing is negative.

If the crossing is labelled x , then we write $\epsilon(x) = +1$ if the crossing is positive, or $\epsilon(x) = -1$ if the crossing is negative.

Example 5.4. For example, in the following picture, the top crossing is positive and the bottom crossing is negative.



Example 5.5. The Hopf link can be oriented in four different ways, as shown below.



In the first two pictures, both crossings are negative. In the last two pictures, both crossings are positive.

Definition 5.6. A *Laurent polynomial* (over \mathbb{Z}) is an expression of the form

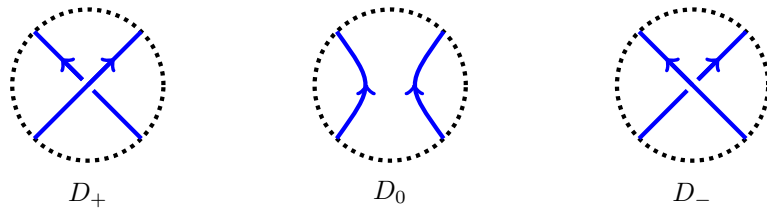
$$p = \sum_{k=-N}^N a_k A^k$$

for some natural number N and some list of coefficients $a_{-N}, \dots, a_N \in \mathbb{Z}$.

Example 5.7. A^6 , $A^4 - 5A + A^{-3}$ and $A^{999} - 99A^{-999}$ are all Laurent polynomials.

Theorem 5.8. There is a unique way to define a Laurent polynomial $f(D)$ for each oriented link diagram D , such that the following axioms are satisfied.

- If D and D' are R -equivalent, then $f(D) = f(D')$.
- If D is an unknotted circle, then $f(D) = 1$.
- Suppose that D_+ , D_- and D_0 are oriented diagrams that are essentially the same, except that there is a small disc where they differ as follows:



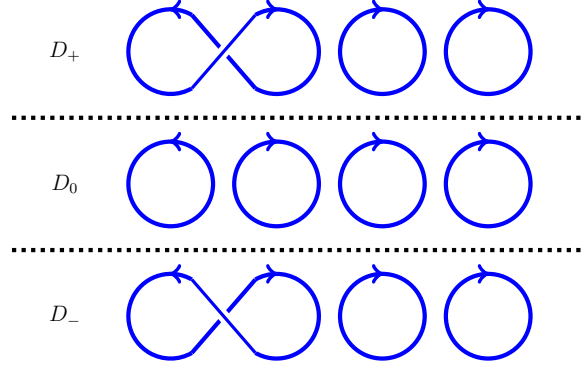
Then $A^4 f(D_+) - A^{-4} f(D_-) = (A^{-2} - A^2) f(D_0)$.

Property (a) tells us that $f(D)$ is a link invariant: it only depends on the intrinsic properties of the link L corresponding to D , so we can write $f(L)$ instead for $f(D)$. This invariant is called the *Jones polynomial* of D . Property (b) is called the *normalisation axiom*, and property (c) is the *skein relation*.

We will prove the Jones polynomial theorem in the next section. In this section, we will just assume that the theorem is true, and use it to calculate $f(L)$ for various links L .

Proposition 5.9. *Let U_n consist of n disjoint circles with no knotting or linking (where $n \geq 1$). Then $f(U_n) = -(A^2 + A^{-2})^{n-1}$.*

Proof. We argue by induction on n . The normalisation axiom says that $f(U_1) = 1$, so the claim is true for $n = 1$. We will illustrate the induction step for $n = 4$. Consider the following diagrams:



Note that we have oriented some circles clockwise and some circles anticlockwise, but this does not matter because an anticlockwise circle is equivalent to a clockwise circle, just by turning it over. The three pictures are related as in the skein relation, so we have

$$A^4 f(D_+) - A^{-4} f(D_-) = (A^{-2} - A^2) f(D_0).$$

On the other hand, the diagrams D_+ and D_- are each equivalent to U_3 , whereas D_0 is U_4 . We thus get

$$A^4 f(U_3) - A^{-4} f(U_3) = (A^{-2} - A^2) f(U_4).$$

By drawing similar pictures with more circles, we see in the same way that

$$A^4 f(U_n) - A^{-4} f(U_n) = (A^{-2} - A^2) f(U_{n+1})$$

for all $n \geq 1$. This can be rearranged to give

$$f(U_{n+1}) = \frac{A^4 - A^{-4}}{A^{-2} - A^2} f(U_n) = -(A^2 + A^{-2}) f(U_n).$$

If we already know that $f(U_n) = -(A^2 + A^{-2})^{n-1}$, we can deduce that $f(U_{n+1}) = -(A^2 + A^{-2})^n$, and this proves the original claim by induction. \square

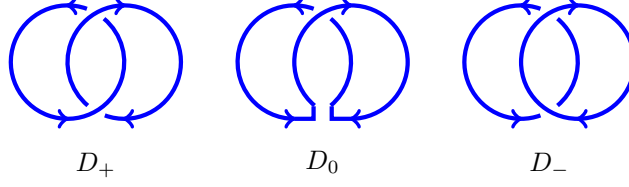
Proposition 5.10. *Let H_+ and H_- be the two versions of the Hopf link shown below, so H_+ has two positive crossings, and H_- has two negative crossings.*



Then $f(H_+) = -A^{-2}(1 + A^{-8})$ and $f(H_-) = -A^2(1 + A^8)$.

Proof. We will give the proof for H_+ , and leave the (similar) proof for H_- to the reader.

The following three diagrams are related as in the skein relation:

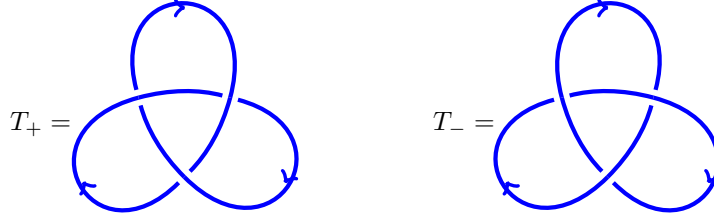


Here D_+ is H_+ , and D_0 is a single unknotted circle, so it is equivalent to U_1 . Similarly, the two circles in D_- can be separated, so D_- is equivalent to U_2 . We therefore have $f(D_0) = f(U_1) = 1$ and $f(D_-) = f(U_2) = -A^2 - A^{-2}$. The skein relation tells us that $A^4 f(H_+) - A^{-4} f(U_2) = (A^{-2} - A^2) f(U_1)$, or equivalently

$$A^4 f(H_+) - A^{-4}(-A^2 - A^{-2}) = A^{-2} - A^2.$$

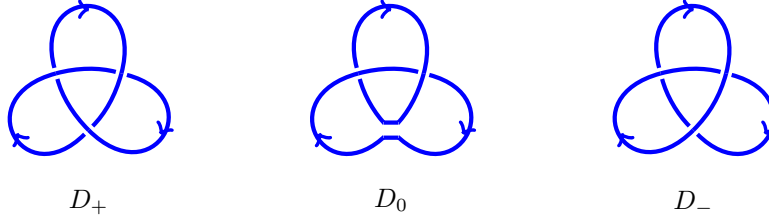
After expanding this out and rearranging we get $A^4 f(H_+) = -A^2 - A^{-6}$ and so $f(H_+) = -A^{-2} - A^{-10} = -A^{-2}(1 + A^{-8})$, as claimed. \square

Proposition 5.11. *Let T_+ and T_- be the two versions of the trefoil as shown below, so T_+ has three positive crossings, and T_- has three negative crossings.*



Then $f(T_+) = A^{-4} + A^{-12} - A^{-16}$ and $f(T_-) = A^4 + A^{12} - A^{16}$.

Proof. The following diagrams are related as in the skein relation:



Note that

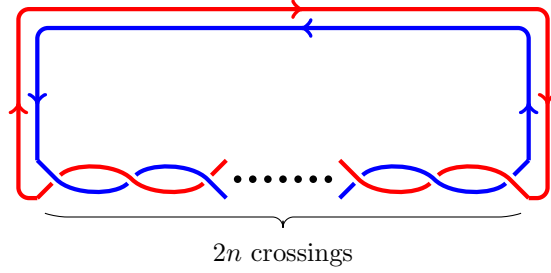
- D_+ is the same as T_+
- D_0 is a Hopf link with two positive crossings, so $f(D_0) = -A^{-2}(1 + A^{-8})$ by Proposition 5.10.
- D_- can be deformed into an unknotted circle, so $f(D_-) = 1$.

The skein relation now becomes

$$A^4 f(T_+) - A^{-4} = (A^{-2} - A^2)(-A^{-2}(1 + A^{-8})) = -A^{-4} + 1 - A^{-12} + A^{-8}.$$

This can be rearranged to give $f(T_+) = A^{-4} + A^{-12} - A^{-16}$, as claimed. The argument for $f(T_-)$ is similar. \square

Now let B_{2n} denote the following link:



You should convince yourself that this really does split into two separate strands as indicated (which would not be true if the number of crossings was odd).

Proposition 5.12. *For all $n \geq 0$ we have*

$$f(B_{2n}) = -A^2 \left(A^{8n} + \frac{A^{8n-4} + 1}{A^4 + 1} \right).$$

Proof. We define

$$p(n) = -A^2 \left(A^{8n} + \frac{A^{8n-4} + 1}{A^4 + 1} \right),$$

so the claim is that $f(B_{2n}) = p(n)$.

First note that B_0 consists of two separate unlinked circles. This was called U_2 in Proposition 5.9, and we proved there that $f(U_2) = -A^{-2} - A^2$. On the other hand, we have

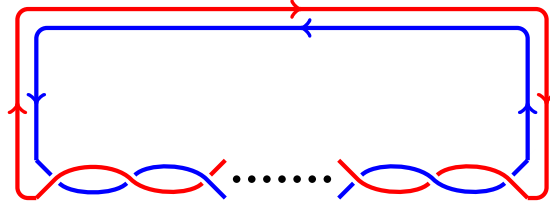
$$p(0) = -A^2 \left(1 + \frac{A^{-4} + 1}{A^4 + 1} \right) = -A^2(1 + A^{-4}) = -A^2 - A^{-2},$$

so $f(B_0) = p(0)$ as required.

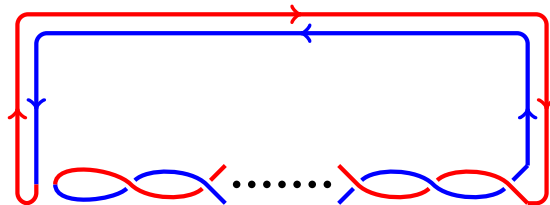
Now suppose we have shown that $f(B_{2n}) = p(n)$ for some n , and consider $f(B_{2n+2})$. We will draw the case $n = 3$ for simplicity, but it should be clear that the same pattern will work for any n . All the crossings in B_{2n+2} are negative. Let D_- be B_{2n+2} , let D_+ be the result of switching the strands in the first crossing, and let D_0 be the result of removing the first crossing, so we have a skein relation

$$A^4 f(D_+) - A^{-4} f(D_-) = (A^{-2} - A^2) f(D_0).$$

The link D_+ is like this:



The first pair of crossings can be removed by a Reidemeister move of type 2, and this just leaves a copy of B_{2n} , so we have $f(D_+) = f(B_{2n})$, and this is the same as $p(n)$ by our induction hypothesis. On the other hand, the link D_0 is like this:



The two strands have merged, and everything can be unwound to give a single unknotted circle, so $f(D_0) = 1$. The skein relation now reads

$$A^4 p(n) - A^{-4} f(B_{2n+2}) = A^{-2} - A^2,$$

so we get

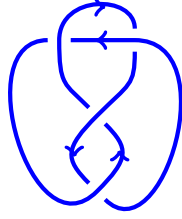
$$f(B_{2n+2}) = A^8 p(n) - A^2 + A^6.$$

After recalling the definition of $p(n)$, this becomes

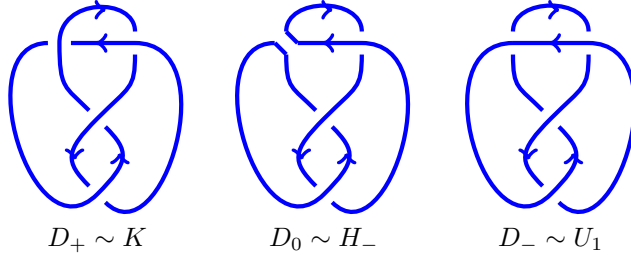
$$\begin{aligned} f(B_{2n+2}) &= -A^2 \left(A^8 \cdot A^{8n} + A^8 \cdot \frac{A^{8n-4} + 1}{A^4 + 1} \right) - A^2 + A^6 \\ &= -A^2 \left(A^{8(n+1)} + \frac{A^{8(n+1)-4} + A^8}{A^4 + 1} + 1 - A^4 \right) \\ &= -A^2 \left(A^{8(n+1)} + \frac{A^{8(n+1)-4} + A^8 + A^4 + 1 - A^8 - A^4}{A^4 + 1} \right) \\ &= -A^2 \left(A^{8(n+1)} + \frac{A^{8(n+1)-4} + 1}{A^4 + 1} \right) = p(n+1). \end{aligned}$$

This completes the required induction step, so $f(B_{2n}) = p(n)$ for all n , as claimed. \square

Example 5.13. Consider the following knot diagram K (called the figure eight). Note that the top two crossings are positive, and the bottom two are negative.



We will calculate $f(K)$ in two different ways. For the first way, we use the skein relation associated to the top left crossing. The three diagrams involved in this relation are as follows:



The first diagram, with a positive crossing at the top left, is our original diagram K . The second diagram is obtained by splitting and rejoining the top left crossing; it is easily seen to be equivalent to the Hopf link H_- discussed in Proposition 5.10. The third diagram is obtained by switching the strands at the top left crossing; it can be unwound to give the unknot U_1 . We thus have a skein relation

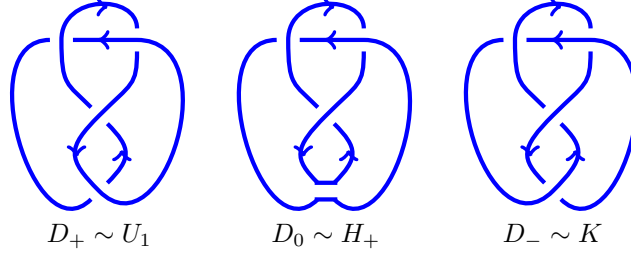
$$A^4 f(K) - A^{-4} f(U_1) = (A^{-2} - A^2) f(H_-).$$

After recalling that $f(U_1) = 1$ and $f(H_-) = -A^2 - A^{10}$ we rearrange and expand everything to get

$$\begin{aligned} f(K) &= A^{-4} (A^{-4} f(U_1) + (A^{-2} - A^2) f(H_-)) \\ &= A^{-4} (A^{-4} + (A^{-2} - A^2)(-A^2 - A^{10})) = A^{-4} (A^{-4} - 1 - A^8 + A^4 + A^{12}) \\ &= A^{-8} - A^{-4} + 1 - A^4 + A^8. \end{aligned}$$

(Note that we have written this with ascending powers of A , which keeps everything tidy and makes it easier to compare this result with other results).

For our second approach, we will instead use the skein relation for the bottom crossing. The relevant diagrams are as follows.



The third diagram, with a negative crossing at the bottom, is our original diagram K . The first diagram is equivalent to the unknot U_1 (with $f(U_1) = 1$), and the second diagram is equivalent to the positive Hopf link H_+ (with $f(H_+) = -A^{-10} - A^{-2}$). The skein relation is

$$A^4 f(U_1) - A^{-4} f(K) = (A^{-2} - A^2) f(H_+).$$

After rearranging and expanding we get

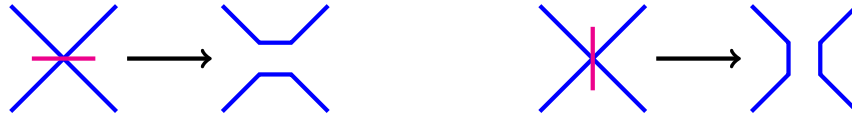
$$\begin{aligned} f(K) &= A^4 (A^4 f(U_1) - (A^{-2} - A^2) f(H_+)) \\ &= A^4 (A^4 - (A^{-2} - A^2)(-A^{-10} - A^{-2})) = A^4 (A^4 + A^{-12} + A^{-4} - A^{-8} - 1) \\ &= A^{-8} - A^{-4} + 1 - A^4 + A^8, \end{aligned}$$

which is the same answer as before.

Note that in the above example, it was not at all obvious that the two approaches would give the same answer. If we just started from the skein relation and tried to use that as the definition of the Jones polynomial, then this would be a problem: we would not know that the polynomial was well-defined, because using skein relations on different crossings might (as far as we know) give different answers. In order to prove Jones's theorem, we need to give a completely different definition of the Jones polynomial, for which the well-definedness is obvious. We will then prove that this definition obeys the skein relation.

6. PROOF OF JONES'S THEOREM

Definition 6.1. Consider a crossing in a link universe. A *splitting marker* is a short bar drawn across the crossing between the strands. Using such a marker, we can cut and rejoin the strands and eliminate the crossing. Thus, there are two possible ways to draw a splitting marker:

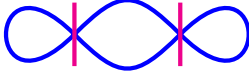
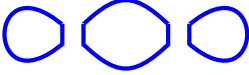

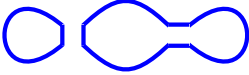

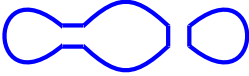

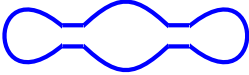


Definition 6.2. A *state* of a universe U is a choice of splitting marker for each crossing in U . If there are n crossings, then there are 2^n possible states. If S is a state, then we can split and rejoin the strands in accordance with S , to obtain a new diagram which we call D/S . This just consists of disjoint circles with no crossings. The number of different circles is called the *disconnectedness* of S , and is written $|S|$.

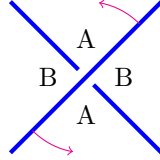
Example 6.3. Consider the following link universe:



This has four possible states, which are tabulated below:

S	D/S	$ S $
		3
		2
		2
		1

Definition 6.4. Now suppose we have a link diagram, so every crossing has an upper strand and a lower strand. We label the four sectors next to each crossing by the following rule: the sectors on the anticlockwise side of the upper strand are labelled A , and the sectors on the clockwise side of the upper strand are labelled B .



Note that there are two ways to draw a splitting marker across this crossing. One possibility is to draw the marker so that it joins the two sectors marked A ; this is a *type A splitting marker*. The other possibility is to draw the marker so that it joins the two sectors marked B ; this is a *type B splitting marker*.

Definition 6.5. Now suppose we have a link diagram D and a state S of D , so S specifies a splitting marker at each crossing. Because D is a link diagram (and not just a link universe), we know which strand is the upper strand at each crossing, and we can use this to label all the sectors with A or B . Let p be the number of crossing markers of type A , and let q be the number of crossing markers of type B . We then write

$$\langle D|S \rangle = A^p B^q.$$

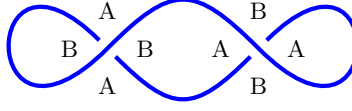
(This should be regarded as an abstract polynomial in variables A and B .) We then introduce a third variable C , and put

$$\langle\langle D \rangle\rangle = \sum_S \langle D|S \rangle C^{|S|-1}.$$

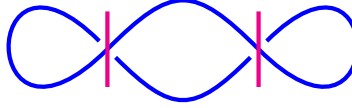
(If D has n crossings, then $\langle\langle D \rangle\rangle$ will be a sum of 2^n terms, one for each possible state S .) This is called the *unnormalised bracket* of D . The *normalised bracket* (or *Kaufmann bracket*) is the expression that we get by substituting $B = A^{-1}$ and $C = -A^2 - A^{-2}$ in $\langle\langle D \rangle\rangle$.

It will turn out that $\langle D \rangle$ is almost the same as $f(D)$, but there is one more ingredient that we will discuss later.

Example 6.6. Consider the following link diagram D , in which we have labelled the sectors with A or B , as discussed above.



There are four possible states; the following picture shows one of them, which we call S .



By comparing with the previous diagram, we see that the left hand crossing has type A, but the right hand crossing has type B, so $\langle D|S \rangle = AB$. Also, if we split the diagram using the crossing markers, then the resulting diagram D/S consists of three separate circles. We therefore have $|S| = 3$ and $C^{|S|-1} = C^2$, which means that the corresponding term $\langle D|S \rangle C^{|S|-1}$ in $\langle\langle D \rangle\rangle$ is just ABC^2 . The corresponding term in $\langle D \rangle$ is $A \times A^{-1} \times (-A^2 - A^{-2}) = -A^2 - A^{-2}$.

If we repeat this analysis for the other three states, we obtain the following table:

S	$ S $	types	$\langle D S \rangle$	term in $\langle\langle D \rangle\rangle$	term in $\langle D \rangle$
	3	A, B	AB	ABC^2	$(-A^2 - A^{-2})^2$
	2	A, A	A^2	A^2C	$A^2(-A^2 - A^{-2})$
	2	B, B	B^2	B^2C	$A^{-2}(-A^2 - A^{-2})$
	1	B, A	AB	AB	1

By adding up the terms in column 5, we get

$$\langle\langle D \rangle\rangle = ABC^2 + A^2C + B^2C + AB.$$

We can also calculate $\langle D \rangle$ by adding up the entries in column 6. However, it is not hard to see that the term for the first state cancels the sum of the terms for the second and third states, so we just end up with $\langle D \rangle = 1$.

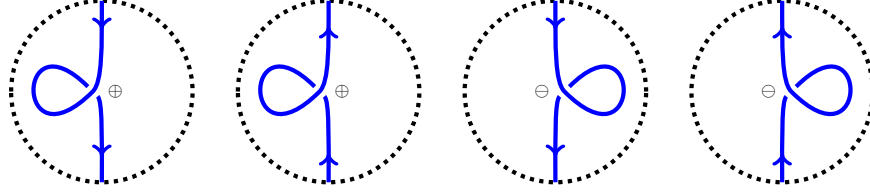
Definition 6.7. If D is an oriented knot diagram, then the *writhe* of D is the number of positive crossings minus the number of negative crossings. We write $w(D)$ for this number.

Remark 6.8. Recall (from Definition 5.3) that we write $\epsilon(x) = 1$ if x is a positive crossing, and $\epsilon(x) = -1$ if x is a negative crossing. With this notation, we have $w(D) = \sum_x \epsilon(x)$.

We are finally ready to define the Jones polynomial:

Definition 6.9. For any oriented link diagram D , we put $f(D) = (-A^{-3})^{w(D)}\langle D \rangle$.

We next want to prove that f is invariant under Reidemeister moves of type 1. For this, we first need to be more precise about a point that previously glossed over: there are two slightly different versions of Reidemeister move 1. Consider the following pictures:



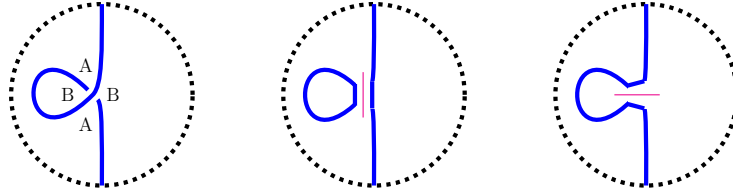
In the first two pictures, the region enclosed by the loop is on the clockwise side of the top strand at the crossing. We can orient the strand either upwards or downwards, but either way it works out that the crossing is positive. We call this kind of loop a *positive loop*. In the third and fourth pictures, the region enclosed by the loop is instead on the anticlockwise side of the top strand. This makes the crossing negative, for both possible choices of orientation. We call this kind of loop a *negative loop*.

Proposition 6.10. Let D be an oriented link diagram.

- Suppose that D' is obtained from D by adding a positive loop. Then $\langle\langle D' \rangle\rangle = (AC + B)\langle\langle D \rangle\rangle$ and $\langle D' \rangle = -A^3\langle D \rangle$ and $f(D') = f(D)$.
- Suppose that D' is obtained from D by adding a negative loop. Then $\langle\langle D' \rangle\rangle = (A + BC)\langle\langle D \rangle\rangle$ and $\langle D' \rangle = -A^{-3}\langle D \rangle$ and $f(D') = f(D)$.

In both cases we have $f(D') = f(D)$, so f is invariant under Reidemeister moves of type 1.

Proof. First consider the case of a positive loop. The following picture shows the A and B regions, and the effect of adding a splitting marker of type A or type B.



Suppose we have a state S for D . We let S_A denote the state for D' obtained by adding a type A marker at the extra crossing, and we let S_B denote the state obtained by adding a type B marker. Every state S' for D' is of the form S_A or S_B for some S , so

$$\langle\langle D' \rangle\rangle = \sum_{S'} \langle D' | S' \rangle C^{|S'| - 1} = \sum_S \left(\langle D' | S_A \rangle C^{|S_A| - 1} + \langle D' | S_B \rangle C^{|S_B| - 1} \right).$$

From the definitions it is clear that $\langle D' | S_A \rangle = A \langle D | S \rangle$ and $\langle D' | S_B \rangle = B \langle D | S \rangle$. If we split D' using S_A then the result is the same as splitting D using S , and then adding an extra circle, so $|S_A| = |S| + 1$. On the other hand, the result of splitting D' using S_B is exactly the same as the result of splitting D using S , so $|S_B| = |S|$. This gives

$$\begin{aligned} \langle\langle D' \rangle\rangle &= \sum_S \left(\langle D' | S_A \rangle C^{|S_A| - 1} + \langle D' | S_B \rangle C^{|S_B| - 1} \right) \\ &= \sum_S \left(A \langle D | S \rangle C^{|S|} + B \langle D | S \rangle C^{|S| - 1} \right) = (AC + B) \sum_S \langle D | S \rangle C^{|S| - 1} \\ &= (AC + B) \langle\langle D \rangle\rangle. \end{aligned}$$

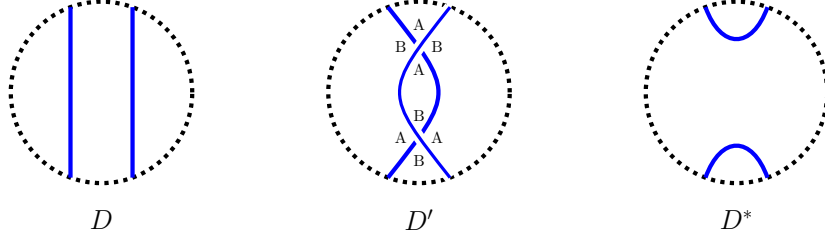
Substituting $B = A^{-1}$ and $C = -A^{-2} - A^2$ gives $AC + B = -A^{-1} - A^3 + A^{-1} = -A^3$ so $\langle D' \rangle = -A^3 \langle D \rangle$. Moreover, D' has one extra positive crossing compared to D , so $w(D') = w(D) + 1$, so

$$f(D') = (-A^{-3})^{w(D')} \langle D' \rangle = (-A^{-3})^{w(D) + 1} (-A^3) \langle D \rangle = (-A^{-3})^{w(D)} \langle D \rangle = f(D),$$

as claimed. This completes the argument for a positive loop. The case of a negative loop is similar. The first difference is that the inside of the loop is marked A instead of B, so the so it is the type B splitting marker that creates an extra circle, not the type A marker. This gives $\langle\langle D' \rangle\rangle = (A + BC)\langle\langle D \rangle\rangle$ (instead of $(AC + B)\langle\langle D \rangle\rangle$, as in the positive case). After substituting $B = A^{-1}$ and $C = -A^2 - A^{-2}$ we get $\langle D' \rangle = -A^{-3}\langle D \rangle$. However, we have added a negative crossing rather than a positive crossing, so we now have $w(D') = w(D) - 1$ (rather than $w(D') = w(D) + 1$, as in the positive case). We again find that everything cancels out. \square

Proposition 6.11. *Let D be an oriented link diagram, and let D' be obtained by adding two extra crossings via a Reidemeister move of type 2. Then $\langle D' \rangle = \langle D \rangle$ and $w(D') = w(D)$ and $f(D') = f(D)$.*

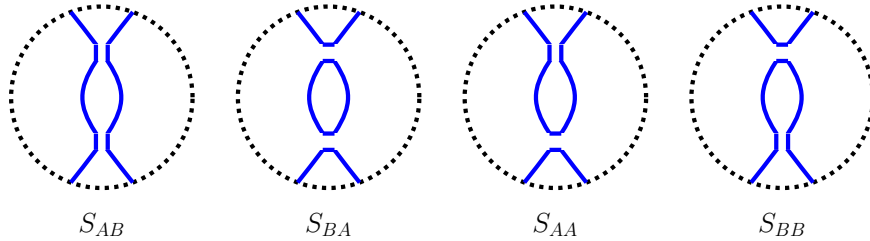
Proof. By assumption, D and D' are the same except that they differ in a small disc as shown in the first two pictures below. We will also consider the diagram D^* as shown in the third picture.



Consider a state S for D . As D^* has the same crossings as D , there is a corresponding state S^* for D^* with $\langle D^* | S^* \rangle = \langle D | S \rangle$. (There is no obvious relationship between $|S|$ and $|S^*|$, but it turns out that that does not matter.) On the other hand, there are four different ways to turn S into a state for D' , by adding markers for the two extra crossings. We write S_{AB} for the state obtained by adding a type A marker to the top crossing and a type B marker to the bottom crossing. We also define S_{BA} , S_{AA} and S_{BB} in the analogous way. It is clear that

$$\langle D' | S_{AB} \rangle = \langle D' | S_{BA} \rangle = AB \langle D | S \rangle \quad \langle D' | S_{AA} \rangle = A^2 \langle D | S \rangle \quad \langle D' | S_{BB} \rangle = B^2 \langle D | S \rangle.$$

Next, the effect of splitting along the various states is as follows.



From the first picture, we see that $D'/S_{AB} = D/S$ and so $|S_{AB}| = |S|$. From the third and fourth pictures, we see that $D'/S_{AA} = D'/S_{BB} = D^*/S^*$, so $|S_{AA}| = |S_{BB}| = |S^*|$. On the other hand, D'/S_{BA} is the same as D^*/S^* but with an extra circle, so $|S_{BA}| = |S^*| + 1$.

Now consider $\langle\langle D' \rangle\rangle$. This has a term for each of the states S_{AB} , S_{BA} , S_{AA} and S_{BB} . The term for S_{AB}

$$\langle D' | S_{AB} \rangle C^{|S_{AB}|-1} = AB \langle D | S \rangle C^{|S|-1},$$

which is AB times the term for S in $\langle\langle D \rangle\rangle$. For the other three terms, we have

$$\langle D' | S_{AB} \rangle C^{|S_{AB}|-1} = AB \langle D | S \rangle C^{|S^*|} = ABC \langle D^* | S^* \rangle C^{|S^*|-1}$$

$$\langle D' | S_{AA} \rangle C^{|S_{AA}|-1} = A^2 \langle D | S \rangle C^{|S^*|-1} = A^2 \langle D^* | S^* \rangle C^{|S^*|-1}$$

$$\langle D' | S_{BB} \rangle C^{|S_{BB}|-1} = B^2 \langle D | S \rangle C^{|S^*|-1} = B^2 \langle D^* | S^* \rangle C^{|S^*|-1}.$$

Together, these three terms give $ABC + A^2 + B^2$ times the term for S^* in $\langle\langle D^* \rangle\rangle$. Putting all this together, we find that

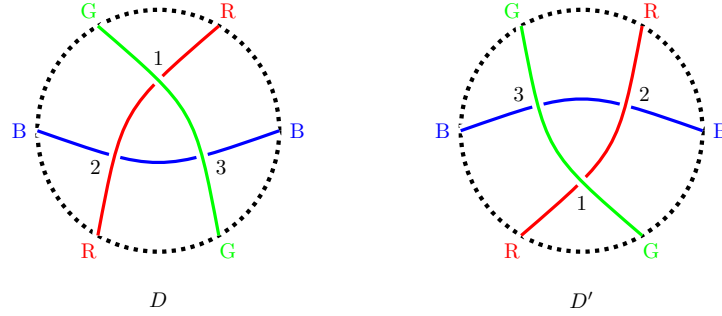
$$\langle\langle D' \rangle\rangle = AB \langle\langle D \rangle\rangle + (ABC + A^2 + B^2) \langle\langle D^* \rangle\rangle.$$

Now substitute $B = A^{-1}$ and $C = -A^2 - A^{-2}$, so AB becomes 1 and $ABC + A^2 + B^2$ becomes 0. The above equation then becomes $\langle D' \rangle = \langle D \rangle$. Finally, there are four different ways to orient the strands in D' ,

but it is not hard to see that in each case one crossing becomes positive and the other becomes negative. It follows that $w(D') = w(D)$, and so $f(D') = f(D)$. \square

Proposition 6.12. *Let D and D' be oriented link diagrams that are related by a Reidemeister move of type 3. Then $\langle D \rangle = \langle D' \rangle$ and $w(D) = w(D')$ and $f(D) = f(D')$.*

Proof. We will label the strands and crossings as follows.



For those reading in black and white: the strands labelled R, G and B are red, green and blue respectively. Note that in both diagrams

- The crossing of the red and green strands is labelled 1.
- The crossing of the red and blue strands is labelled 2.
- The crossing of the green and blue strands is labelled 3.

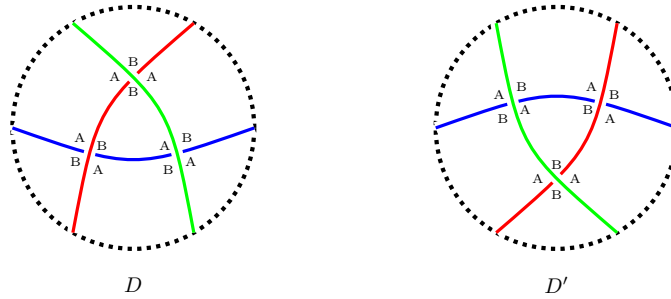
Recall that we are given orientations on D and D' , which are assumed to correspond in the obvious way: if the blue strand runs from left to right in D , then it must also run from left to right in D' , and similarly for the other strands. It is not hard to see that the sign of each crossing in D will be the same as the sign of the corresponding crossing in D' , so $w(D) = w(D')$. For the rest of the proof we will not need to worry about orientations.

Now let S be a state for all the crossings in D outside the dotted circle. There are eight ways to assign splitting markers to the remaining crossings and thus get a state for D . For example, we write S_{ABA} for the state that adds a type A marker at crossing 1, a type B marker at crossing 2, and a type A marker at crossing 3. Each of these eight states contributes a term to $\langle\langle D \rangle\rangle$, and we write T for the sum of these terms. This is a polynomial in A , B and C , so we can set $B = A^{-1}$ and $C = -A^2 - A^{-2}$ to get a Laurent polynomial $T_* \in \mathbb{Z}[A, A^{-1}]$, which contributes to $\langle D \rangle$.

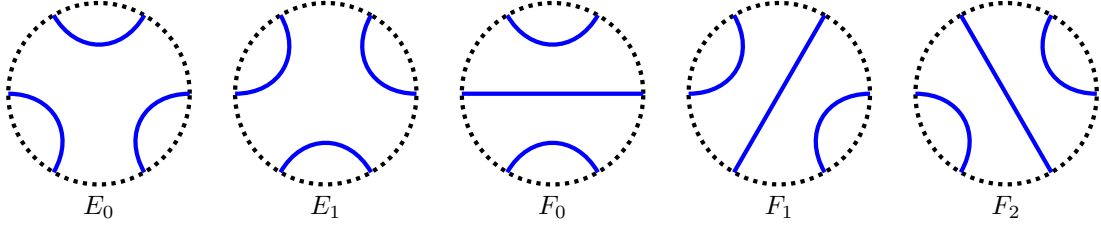
Similarly, the same states contribute eight terms to $\langle\langle D' \rangle\rangle$, and we write T' for the sum of these terms. We also write T'_* for the corresponding Laurent polynomial, obtained by substituting $B = A^{-1}$ and $C = -A^2 - A^{-2}$ again. We will need to show that $T'_* = T_*$.

First let α and β denote the numbers of type A and type B markers in S . Each term in T or T' will then be $A^\alpha B^\beta$ multiplied by some extra factors depending on what happens inside the dotted circle.

Next, we record the A and B sectors for D and D' :



There are a number of different patterns that we can get by cutting and rejoining D or D' ; these will be labelled as follows.



We also write E_i^+ for the pattern consisting of E_i together with an extra circle. The eight terms in T can be tabulated as follows.

AAA E_0	AAB E_0^+	ABA F_0	ABB E_0
BAA F_2	BAB E_0	BBA E_1	BBB F_1

The label in the top left of each box shows the splitting markers, and the label in the bottom right shows the corresponding pattern of connections. For the top right box, for example, the corresponding term in T is $A^\alpha B^\beta$ (for the crossings outside the dotted circle), multiplied by AB^2 (for the crossings inside the dotted circle), multiplied by $C^{|E_0|-1}$. After collecting the terms together, we get

$$T = A^\alpha B^\beta C^{-1} \left(C^{|E_0|} (A^3 + A^2 BC + 2AB^2) + C^{|E_1|} AB^2 + C^{|F_0|} A^2 B + C^{|F_1|} B^3 + C^{|F_2|} A^2 B \right).$$

We can now analyse T' in a similar way. The terms are as follows:

AAA E_1	AAB E_1^+	ABA F_0	ABB E_1
BAA F_2	BAB E_1	BBA E_0	BBB F_1

From this we get

$$T' = A^\alpha B^\beta C^{-1} \left(C^{|E_0|} AB^2 + C^{|E_1|} (A^3 + A^2 BC + 2AB^2) + C^{|F_0|} A^2 B + C^{|F_1|} B^3 + C^{|F_2|} A^2 B \right).$$

We can now subtract this from our earlier expression for T . The terms involving F_0 , F_1 and F_2 are exactly the same, so they cancel out. The term involving E_0 has a factor

$$(A^3 + A^2BC + 2AB^2) - AB^2 = A(A^2 + ABC + B^2).$$

The term involving E_1 has essentially the same factor, but with an extra minus sign. We therefore get

$$T - T' = A^\alpha B^\beta C^{-1} (C^{|E_0|} - C^{|E_1|}) A(A^2 + ABC + B^2).$$

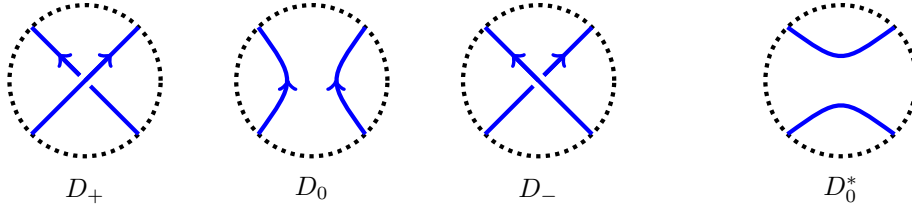
Finally, to find $T_* - T'_*$ we must substitute $B = A^{-1}$ and $C = -A^{-2} - A^2$, but this gives

$$A^2 + ABC + B^2 = A^2 + AA^{-1}(-A^{-2} - A^2) + A^{-2} = 0,$$

so $T_* - T'_* = 0$, or in other words $T_* = T'_*$. We can now take the sum over all possible states S to get $\langle T \rangle = \langle T' \rangle$. As we explained earlier, we also have $w(D) = w(D')$, and it follows that $f(D) = f(D')$. \square

Proposition 6.13. *The function f (as defined in this section) satisfies the skein relation.*

Proof. We consider three oriented link diagrams D_+ , D_0 and D_- , related in the usual way that we have seen before. We also consider one more link diagram D_0^* , as shown below. (There is no natural way to choose an orientation for D_0^* , but we will not need one.)



Note that $w(D_+) = w(D_0) + 1$ and $w(D_-) = w(D_0) - 1$. Next, as D_0 and D_0^* have the same crossings, any state S for D_0 has a corresponding state S^* for D_0^* . These satisfy $\langle D_0 | S \rangle = \langle D_0^* | S^* \rangle$, but there is no obvious relation between $|S|$ and $|S^*|$. Next, let S_V be the state for D_+ or D_- obtained by adding a vertical splitting marker, and let S_H be obtained by adding a horizontal splitting marker. Note that S_V has type A for D_+ and type B for D_- , whereas S_H has type B for D_+ and type A for D_- . This gives

$$\begin{aligned} \langle\langle D_+ \rangle\rangle &= \sum_S \left(\langle D_+ | S_V \rangle C^{|S|-1} + \langle D_+ | S_H \rangle C^{|S^*|-1} \right) \\ &= \sum_S \langle D_0 | S \rangle C^{-1} \left(AC^{|S|} + BC^{|S^*|} \right) \\ \langle\langle D_- \rangle\rangle &= \sum_S \left(\langle D_- | S_V \rangle C^{|S|-1} + \langle D_- | S_H \rangle C^{|S^*|-1} \right) \\ &= \sum_S \langle D_0 | S \rangle C^{-1} \left(BC^{|S|} + AC^{|S^*|} \right). \end{aligned}$$

From this we get

$$A\langle\langle D_+ \rangle\rangle - B\langle\langle D_- \rangle\rangle = \sum_S \langle D_0 | S \rangle C^{-1} (A^2 - B^2) C^{|S|} = (A^2 - B^2) \langle\langle D_0 \rangle\rangle.$$

We can now put $B = A^{-1}$ and $C = -A^{-2} - A^2$ to get

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

This in turn gives

$$\begin{aligned}
A^4 f(D_+) - A^{-4} f(D_-) &= A^4 (-A^{-3})^{w(D_+)} \langle D_+ \rangle - A^{-4} (-A^{-3})^{w(D_+)} \langle D_- \rangle \\
&= A^4 (-A^{-3})^{w(D_0)+1} \langle D_+ \rangle - A^{-4} (-A^{-3})^{w(D_0)-1} \langle D_- \rangle \\
&= (-A^{-3})^{w(D_0)} (A^4 \cdot (-A^{-3}) \langle D_+ \rangle - A^{-4} \cdot (-A^{-3})^{-1} \langle D_- \rangle) \\
&= -(-A^{-3})^{w(D_0)} (A \langle D_+ \rangle - A^{-1} \langle D_- \rangle) \\
&= -(-A^{-3})^{w(D_0)} (A^2 - A^{-2}) \langle D_0 \rangle \\
&= (A^{-2} - A^2) (-A^{-3})^{w(D_0)} \langle D_0 \rangle = (A^{-2} - A^2) f(D_0).
\end{aligned}$$

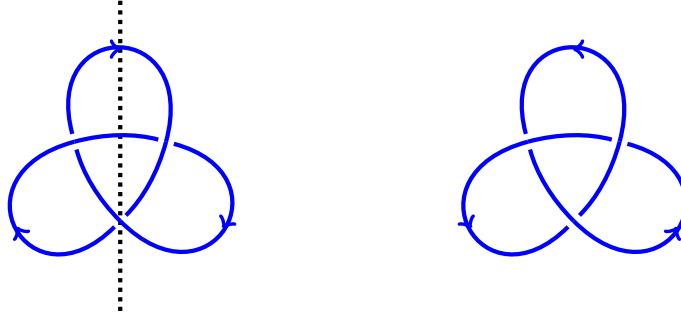
This is the expected skein relation. □

7. ADDITIONAL PROPERTIES OF THE JONES POLYNOMIAL

Definition 7.1. Let D be an oriented link diagram. The *reverse* of D is the oriented diagram D^* obtained by reversing all the orientations. We say that D is *reversible* if D is ambient isotopic to D^* .

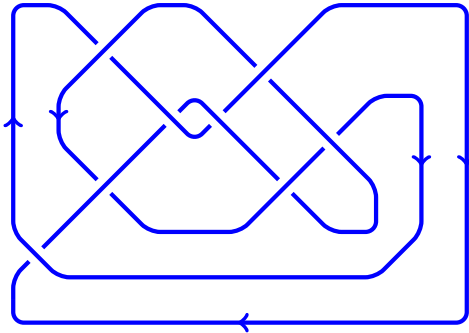
Similarly, given an oriented link L , we write L^* for the same link with the orientations reversed, and we say that L is reversible if it is ambient isotopic to L^* . This is essentially the same as the previous definition for link diagrams, by Reidemeister's Theorem.

Example 7.2. The following picture shows a trefoil and its reverse.



The first picture can be converted to the second one by rotating through π around the dotted line, and that rotation is an ambient isotopy, so the trefoil is reversible.

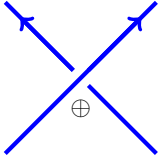
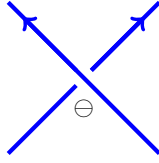
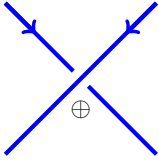
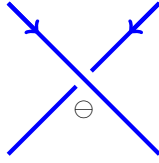
Example 7.3. It is hard to find non-reversible knots, and even harder to prove that they are non-reversible. One example is the knot below, which is 8_{17} in the Rolfsen table.



The Jones polynomial does not help us to decide whether links are reversible, because of the following fact.

Proposition 7.4. $f(D^*) = f(D)$

Proof. Most of the ingredients in the Jones polynomial do not depend on the orientations, so they are obviously the same for D and D^* . The only exception is the writhe, so we just need to check that $w(D) = w(D^*)$. For this, we just need to check that the signs of the crossings are the same in D and \overline{D} . This is clear from the following pictures:

D		
D^*		

□

Remark 7.5. Suppose we have a link with several components. The proposition tells us that if we reverse the orientation of *all* the components, then the Jones polynomial will be unchanged. However, if we reverse some components but not others, then we will usually get a different Jones polynomial. For example, the negative Hopf link H_- can be obtained from H_+ by reversing one of the two strands, and $f(H_+) \neq f(H_-)$.

Definition 7.6. Let D be an oriented link diagram. The *mirror image* of D is the oriented diagram \bar{D} obtained by changing all the under crossings to over crossings and *vice versa*, while keeping the orientations the same. (This corresponds to applying the operation $(x, y, z) \mapsto (x, y, -z)$ to a link in \mathbb{R}^3 .)

Proposition 7.7. If $f(D) = p(A)$, then $f(\bar{D}) = p(A^{-1})$.

This will follow immediately from the following more detailed statement:

Lemma 7.8. If

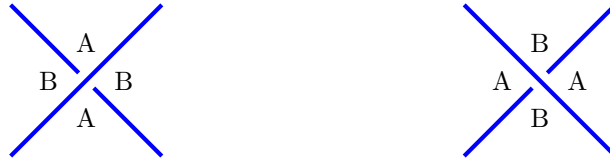
$$\langle\langle D \rangle\rangle = r(A, B, C) \quad \langle D \rangle = q(A) \quad w(D) = m \quad f(D) = p(A),$$

then

$$\langle\langle \bar{D} \rangle\rangle = r(B, A, C) \quad \langle \bar{D} \rangle = q(A^{-1}) \quad w(\bar{D}) = -m \quad f(\bar{D}) = p(A^{-1}).$$

Proof. As D and \bar{D} have the same link universe, they have the same states. Undercrossings and overcrossings do not enter into the definition of D/S , so $D/S = \bar{D}/S$ and $|S|$ is the same for D and \bar{D} .

Switching undercrossings and overcrossings also switches the A and B sectors:



Thus, if $\langle D|S \rangle C^{|S|-1} = A^i B^j C^k$ then $\langle \bar{D}|S \rangle C^{|S|-1} = B^i A^j C^k$. We can now take the sum over all states S : we see that if $\langle\langle D \rangle\rangle = r(A, B, C)$ then $\langle\langle \bar{D} \rangle\rangle = r(B, A, C)$ as claimed. Now put

$$q(A) = r(A, A^{-1}, -A^2 - A^{-2}) = \langle D \rangle.$$

Recall that $\langle \bar{D} \rangle$ is the result of putting $B = A^{-1}$ and $C = -A^2 - A^{-2}$ in $\langle\langle \bar{D} \rangle\rangle = r(B, A, C)$, so $\langle \bar{D} \rangle = r(A^{-1}, A, -A^2 - A^{-2})$, and this is the same as $q(A^{-1})$.

Next, note that switching undercrossings and overcrossings also converts positive crossings to negative crossings and *vice versa*:



It follows that if $w(D) = m$, then $w(\overline{D}) = -m$. Thus, if we put $p(A) = f(D) = (-A^{-3})^m q(A)$, we get

$$f(\overline{D}) = (-A^{-3})^{w(\overline{D})} \langle \overline{D} \rangle = (-A^{-3})^{-m} q(A^{-1}) = (-A^3)^m q(A^{-1}) = p(A^{-1}),$$

as required. \square

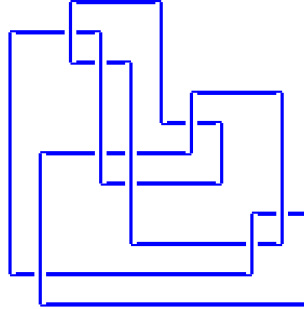
Definition 7.9. A link L is *amphicheiral* if it is ambient isotopic to its mirror image. A link diagram D is amphicheiral if it is R-equivalent to its mirror image. (These concepts are essentially the same, by Reidemeister's Theorem.)

Proposition 7.10. Let D be a link diagram, with $f(D) = p(A)$ say.

- (a) If D is amphicheiral, then $p(A) = p(A^{-1})$.
- (b) If $p(A) \neq p(A^{-1})$, then D is not amphicheiral.
- (c) If $p(A) = p(A^{-1})$, then D may or may not be amphicheiral.

Proof.

- (a) If D is amphicheiral then $\overline{D} \sim D$, so $f(\overline{D}) = f(D) = p(A)$. However, Proposition 7.7 tells us that $f(\overline{D}) = p(A^{-1})$, so we must have $p(A^{-1}) = p(A)$.
- (b) This statement is just the contrapositive of (a), and so is logically equivalent to (a).
- (c) Consider for example the following knot diagram D



This is 9_{42} in the Rolfsen table. It is known that

$$f(D) = A^{-12} - A^{-8} + A^{-4} - 1 + A^4 - A^8 + A^{12}.$$

This is unchanged if we replace A by A^{-1} , so we also have

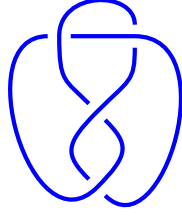
$$f(\overline{D}) = A^{-12} - A^{-8} + A^{-4} - 1 + A^4 - A^8 + A^{12}.$$

This suggests that D and \overline{D} might be R-equivalent, so D might be amphicheiral. However, using more subtle invariants, one can show that D is actually *not* amphicheiral. Examples like this are relatively rare, but they do exist. \square

Remark 7.11. For the sake of variety, we drew the above counterexample as an *arc presentation*. There is one vertical segment with $x = 0$, one vertical segment with $x = 1$, and so on up to $x = 10$. There is also one horizontal segment with $y = 0$, one horizontal segment with $y = 1$ and so on up to $y = 10$. At each crossing, a vertical segment passes over the top of a horizontal segment.

Example 7.12. We saw in Proposition 5.11 that the negative trefoil T_- has $f(T_-) = A^4 + A^{12} - A^{16}$, so $f(\overline{T_-}) = A^{-4} + A^{-12} - A^{-16}$. This proves that T_- cannot be amphicheiral.

Example 7.13. In Example 5.13 we considered the following knot K , called the figure eight:



We showed that

$$f(K) = A^{-8} - A^{-4} + 1 - A^4 + A^8.$$

This is unchanged if we replace A by A^{-1} , so $f(\overline{K}) = f(K)$. This suggests, but does not prove, that K might be R-equivalent to \overline{K} , so K might be amphicheiral. In fact, in this case it is true that K is amphicheiral. (It is easiest to see this using a movie, but I have not prepared one yet.)

Proposition 7.14. *Let D be an oriented link diagram that consists of two separate diagrams D_0 and D_1 , with no common arcs or crossings. Then*

$$f(D) = -(A^2 + A^{-2})f(D_0)f(D_1).$$

Proof. Any state S for D consists of a state S_0 for D_0 and a state S_1 for D_1 . Then D/S is a copy of D_0/S_0 together with a disjoint copy of D_1/S_1 , so $|S| = |S_0| + |S_1|$. Suppose that S_0 has p_0 markers of type A and q_0 markers of type B, and similarly for S_1 . Then the number of type A markers in S is $p_0 + p_1$, and the number of type B markers is $q_0 + q_1$, so

$$\langle D_0 | S_0 \rangle = A^{p_0} B^{q_0} \quad \langle D_1 | S_1 \rangle = A^{p_1} B^{q_1} \quad \langle D | S \rangle = A^{p_0+p_1} B^{q_0+q_1} = \langle D_0 | S_0 \rangle \langle D_1 | S_1 \rangle.$$

From this we get

$$\begin{aligned} \langle\langle D \rangle\rangle &= \sum_S \langle D | S \rangle C^{|S|-1} = \sum_{S_0, S_1} \langle D_0 | S_0 \rangle \langle D_1 | S_1 \rangle C^{|S_0|+|S_1|-1} \\ &= C \left(\sum_{S_0} \langle D_0 | S_0 \rangle C^{|S_0|-1} \right) \left(\sum_{S_1} \langle D_1 | S_1 \rangle C^{|S_1|-1} \right) \\ &= C \langle\langle D_0 \rangle\rangle \langle\langle D_1 \rangle\rangle. \end{aligned}$$

Putting $B = A^{-1}$ and $C = -(A^2 + A^{-2})$ we get $\langle D \rangle = -(A^2 + A^{-2})\langle D_0 \rangle \langle D_1 \rangle$. It is also clear that $w(D) = w(D_0) + w(D_1)$, so

$$f(D) = (-A^{-3})^{w(D)} \langle D \rangle = -(A^2 + A^{-2})(-A^{-3})^{w(D_0)+w(D_1)} \langle D_0 \rangle \langle D_1 \rangle = -(A^2 + A^{-2})f(D_0)f(D_1).$$

□

Proposition 7.15. *Suppose we have oriented link diagrams D_0 and D_1 as shown on the left below, and we combine them to form a diagram D as shown on the right. Then $f(D) = f(D_0)f(D_1)$.*

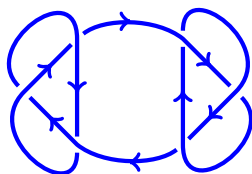


(In this situation we say that D is a connected sum of D_0 and D_1 .)

Proof. This is very similar to the previous proposition. The crossings in D consist of the crossings in D_0 and the crossings in D_1 , so every state S for D consists of a state S_0 for D_0 and a state S_1 for D_1 . We again have $\langle D | S \rangle = \langle D_0 | S_0 \rangle \langle D_1 | S_1 \rangle$ and $w(D) = w(D_0) + w(D_1)$. The only thing that changes is the analysis of D/S . The diagram for D/S is almost a disjoint union of the diagrams for D_0/S_0 and D_1/S_1 , except that one of the components of D_0/S_0 is stitched to one of the components of D_1/S_1 by the two arcs that bridge between D_0 and D_1 . Because of this, we lose one component and we have $|S| = |S_0| + |S_1| - 1$, so

$(|S| - 1) = (|S_0| - 1) + (|S_1| - 1)$. This removes the extra factor of C that we had in the previous proposition, leaving $\langle\langle D \rangle\rangle = \langle\langle D_0 \rangle\rangle \langle\langle D_1 \rangle\rangle$ and $\langle D \rangle = \langle D_0 \rangle \langle D_1 \rangle$ and $f(D) = f(D_0)f(D_1)$. \square

Example 7.16. The following picture D is a connected sum of two positive trefoils.



We therefore have

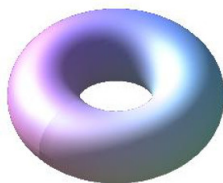
$$f(D) = f(T_+)^2 = (A^{-4} + A^{-12} - A^{-16})^2.$$

8. INTRODUCTION TO SURFACES

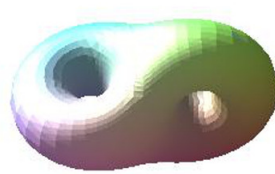
Here are some pictures which we might or might not describe as surfaces. We will discuss some distinctions between them in a moment. After discussing these distinctions we will give some formal definitions. The course website has versions of these pictures which you can rotate and zoom with your mouse.



Sphere



Torus



Double torus



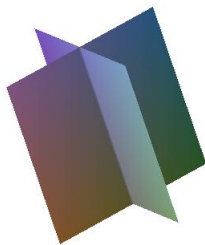
Cube with holes



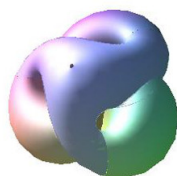
Cylinder



Möbius strip



Two squares



Boy's space



Icosahedron

The first key feature of these sets is that they have the following property.

Definition 8.1. A subset $X \subseteq \mathbb{R}^n$ is *compact* if

- (a) There is a finite radius r such that for all $x \in X$ we have $\|x\| \leq r$.
- (b) For any convergent sequence $x_k \rightarrow a$ in \mathbb{R}^n , if all the terms x_k lie in X , then the limit point a also lies in X .

Example 8.2. The set $(0, 5) \subset \mathbb{R}$ is not compact. It has property (a) (with $r = 5$) but not property (b), because there is a convergent sequence $1/n \rightarrow 0$ where all the terms $1/n$ lie in $(0, 5)$ but the limit point does not lie in $(0, 5)$. However, the set $[0, 5]$ is compact.

Example 8.3. The subset $\mathbb{Z}^n \subseteq \mathbb{R}^n$ is not compact; it satisfies property (b) but not property (a).

Remark 8.4. As many of you will know, Definition 8.1 is not the official definition of compactness; instead, it is a nontrivial theorem that Definition 8.1 is equivalent to the official definition of compactness. However, this distinction will be harmless for us.

Remark 8.5. It is standard to say that a set X is *bounded* if it has property (a), and *closed* if it has property (b), and you will need to be aware of this if you read other books. However, we will not emphasise this terminology, as it can create confusion with the terms *boundary* and *closed surface* which we will introduce shortly.

In the first half of this course we noted that there are things that are like knots but which have infinitely many loops, and we decided to exclude them from consideration. For similar reasons, we will study only compact surfaces.

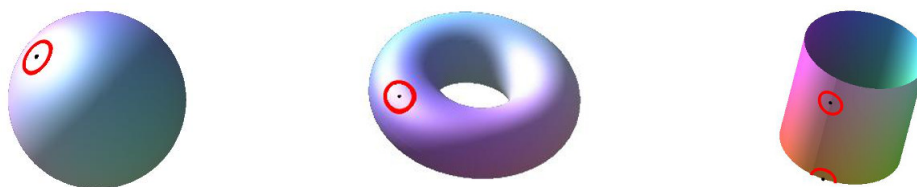
This resolves an ambiguity about our picture of the cylinder: are the two end circles part of the picture or not? In other words, which of the following two sets are we looking at?

$$C = \{(\cos(u), \sin(u), v) \mid u, v \in \mathbb{R}, -1 < v < 1\}$$

$$\overline{C} = \{(\cos(u), \sin(u), v) \mid u, v \in \mathbb{R}, -1 \leq v \leq 1\},$$

It is not hard to see that \overline{C} is compact, but C is not. Thus, we will define the cylinder to be \overline{C} rather than C , so the ends are included. Similarly, the edge of the Möbius strip will be regarded as part of the strip, to ensure that the strip is a compact set.

The second key feature of the sphere and the torus is as follows: near any point, we can cut out a small piece which can be identified with a disc. The double torus and the cube with holes have the same property.



Predefinition 8.6. A *closed surface* is a compact subset $X \subseteq \mathbb{R}^n$ for some n with the above property: any point has a neighbourhood which is a disc.

Later we will give a more careful version.

Note that the cylinder is *not* a closed surface. Instead, it satisfies the following (as should be clear from the above picture):

Predefinition 8.7. A *surface with boundary* is a compact subset $X \subseteq \mathbb{R}^n$ for some n such that any point $x \in X$ has a neighbourhood which is either a disc or a half-disc. We say that x is a *boundary point* if it does not have a disc neighbourhood, but only a half-disc neighbourhood. We write ∂X for the set of boundary points.

The Möbius strip is also a surface with boundary. On the other hand, if we take the union of two squares as in our earlier picture, then points on the intersection line will not have a disc neighbourhood or a half-disc neighbourhood, so the union is not a closed surface or a surface with boundary. Similarly, there are points in Boy's space where a small neighbourhood looks like a pair of intersecting squares, so Boy's space is also not a closed surface. (For this reason we have avoided calling it Boy's surface, which is the more traditional name.) However, this space is closely related to a closed surface called $\mathbb{R}P^2$ (the real projective plane), as we now explain.

We will use the following definition for $\mathbb{R}P^2$:

$$\mathbb{R}P^2 = \{A \in M_3(\mathbb{R}) \mid \text{trace}(A) = 1, A^2 = A^T = A\}.$$

A 3×3 matrix has 9 entries, so we can identify $M_3(\mathbb{R})$ with \mathbb{R}^9 , and so identify $\mathbb{R}P^2$ with a subspace of \mathbb{R}^9 . It turns out that this is a closed surface. We will only sketch the first step in the proof of this fact. Recall that the unit sphere can be described as

$$S^2 = \{(x = (x_1, x_2, x_3) \in \mathbb{R}^3 \mid \sum_i x_i^2 = 1\}.$$

Lemma 8.8. *There is a continuous surjective map $q: S^2 \rightarrow \mathbb{R}P^2$ such that $q(x) = q(y)$ iff $x = \pm y$. In other words, $\mathbb{R}P^2$ can be obtained from S^2 by identifying x with $-x$ for all x .*

Proof. Given any $x \in S^2$, we define a matrix $q(x) \in M_3(\mathbb{R})$ by $q(x)_{ij} = x_i x_j$. We then have

$$\begin{aligned} \text{trace}(q(x)) &= \sum_i q(x)_{ii} = \sum_i x_i^2 = 1 \\ (q(x)^T)_{ij} &= q(x)_{ji} = x_j x_i = q(x)_{ij} \\ (q(x)^2)_{ik} &= \sum_j q(x)_{ij} q(x)_{jk} = x_i x_k \sum_j x_j^2 = x_i x_k = q(x)_{ik}, \end{aligned}$$

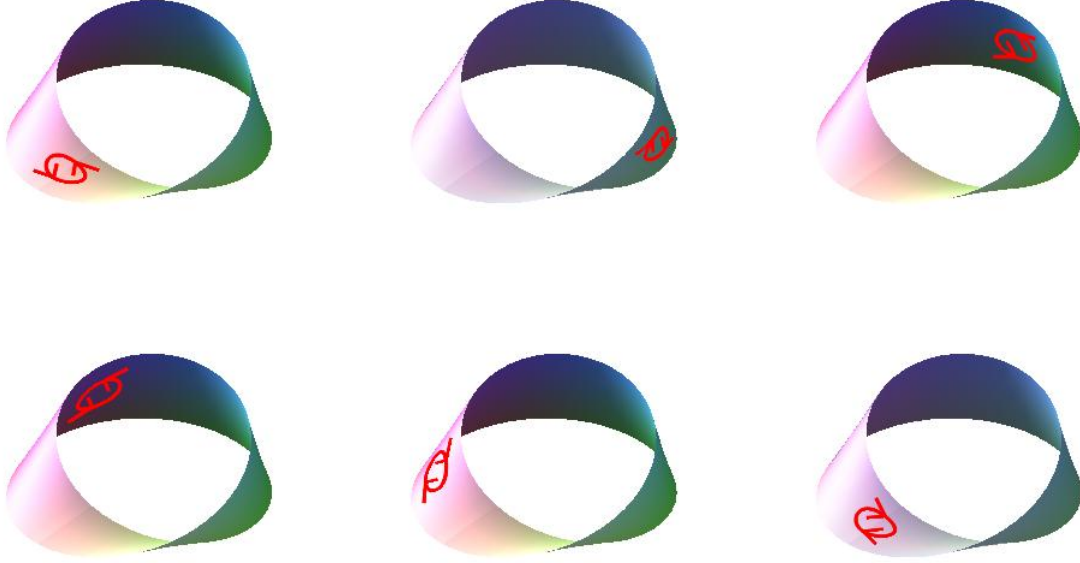
so $q(x) \in \mathbb{R}P^2$. In the opposite direction, suppose that $A \in \mathbb{R}P^2$, and put $L = \{x \mid Ax = x\}$ and $Z = \{x \mid Ax = 0\}$. Using $A^2 = A$ one can check that $\mathbb{R}^3 = L \oplus Z$. One can then choose a basis for L and a basis for Z , and by writing A in terms of these bases we find that $\text{trace}(A) = \dim(L)$. By assumption we have $\text{trace}(A) = 1$, so L is a line, so it contains precisely two unit vectors, say x and $-x$. With a little more linear algebra (and using the additional property $A^T = A$), one can check that $A = q(x) = q(-x)$, and that x and $-x$ are the only vectors with this property. Thus, we see that $q: S^2 \rightarrow \mathbb{R}P^2$ is surjective, and we have $q(x) = q(y)$ iff $x = \pm y$. \square

Now if U is a small disc in S^2 , then $-U$ will be another small disc that does not overlap with $-U$, and $q(U)$ will be a small disc in $\mathbb{R}P^2$. By analysing this situation a little more carefully, one can check that $\mathbb{R}P^2$ is indeed a closed surface.

However, it turns out that it is not possible to make a copy of the space $\mathbb{R}P^2 \subseteq \mathbb{R}^9$ in \mathbb{R}^3 . One can make a copy in \mathbb{R}^4 , but if you try to do it in \mathbb{R}^3 , then the resulting surface will intersect itself. There is a map $f: \mathbb{R}P^2 \rightarrow \mathbb{R}^3$ such that $f(\mathbb{R}P^2)$ is Boy's space, so Boy's space is almost a copy of $\mathbb{R}P^2$, but this is not quite right, because f fails to be injective in some places.

This is analogous to the fact that the trefoil (or any other interesting knot) cannot be represented in \mathbb{R}^2 without self-intersections.

The fact that $\mathbb{R}P^2$ cannot be embedded in \mathbb{R}^3 is closely related to the fact that $\mathbb{R}P^2$ is not orientable. We will explain what this means in the simpler case of the Möbius strip.



We can take a circle marked with arrows circulating anticlockwise and slide it around the strip. When we have slid it around once, the arrows end up circulating clockwise. If we go around a second time, then they end up anticlockwise again. This is not possible with the sphere or the torus or any other closed surface embedded in \mathbb{R}^3 ; for those surfaces, there is a consistent definition of clockwise and anticlockwise, and we cannot convert a clockwise circle to an anticlockwise circle by sliding it around.

We will eventually be able to classify both orientable and nonorientable surfaces, but many things work differently in the two cases.

9. CONNECTED SUMS

Let X and Y be connected surfaces, possibly with boundary. We can make a new surface $X \# Y$ (called the *connected sum* of X and Y) as follows. First, we let $2D$ denote the disc of radius 2 centred at the origin in \mathbb{R}^2 , and choose a continuous embedding $i: 2D \rightarrow X$. Recall that D' is the open disc of radius one, and remove $i(D')$ from X to get a new space $X^* = X \setminus i(D')$. Note that we still have a circle $i(S^1) \subset X^*$, which forms part (or maybe all) of the boundary of X^* . Similarly, we choose an embedding $j: 2D \rightarrow Y$, and put $Y^* = Y \setminus j(D')$. We again have a circle $j(S^1) \subseteq Y^*$. We let $X \# Y$ denote the space formed from X^* and Y^* by attaching $i(x, y)$ to $j(x, -y)$ for all $(x, y) \in S^1$.

Proposition 9.1. *The space $X \# Y$ is again a surface (possibly with boundary).*

Proof. Inside $X \# Y$ we have a circle C where X^* and Y^* are joined together. Consider a point u that does not lie on C . If u lies in $X^* \setminus i(S^1)$, then we can choose a disc or half-disc neighbourhood U of u in X , and provided we take U to be small enough, it will be contained in $X^* \setminus i(S^1)$. Thus, this will give a disc or half-disc neighbourhood of u in $X \# Y$. The same argument works if u lies in $Y^* \setminus j(S^1)$. This just leaves the case of a point $u \in C$. Put

$$A = 2D' \setminus D' = \{(x, y) \in \mathbb{R}^2 \mid 1 \leq \|(x, y)\| < 2\},$$

so we have an subset $i(A)$ in X^* and an open subset $j(A)$ in Y^* . Together, these give an open subset of $X\#Y$, which consists of two copies of A glued together along a circle. To see the effect of this, put

$$\begin{aligned} A' &= \{(x, y) \in \mathbb{R}^2 \mid 1/2 < \|(x, y)\| \leq 1\} \\ B &= A \cup A' = \{(x, y) \in \mathbb{R}^2 \mid 1/2 < \|(x, y)\| < 2\}. \end{aligned}$$

There is a homeomorphism $A \rightarrow A'$ given by

$$(x, y) \mapsto (x/(x^2 + y^2), -y/(x^2 + y^2)).$$

(If we identify \mathbb{R}^2 with \mathbb{C} , this is just $z \mapsto 1/\bar{z}$.) On S^1 , this is just $(x, y) \mapsto (x, -y)$, which is the rule we used when identifying $i(S^1)$ with $j(S^1)$. Thus, the set $U = i(A) \cup j(A)$ can be identified with $A \cup A' = B$. From this it follows that every point in U has a full disc neighbourhood in U , which completes the proof that $X\#Y$ is a surface. \square

Remark 9.2. The notation $X\#Y$ suggests that we get a well-defined result, independent of the choice of embeddings i and j . This is in fact true, but we will not prove it.

The most basic case is as follows:

Proposition 9.3. *For any surface X , we have $X\#S^2 \simeq X$.*

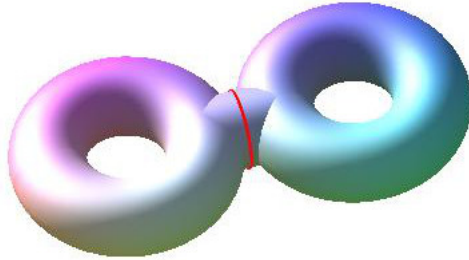
Proof. Note that $(S^2)^*$ is obtained from S^2 by removing an open disc, which we can choose to be the upper hemisphere. That means that $(S^2)^*$ is the (closed) lower hemisphere, which is again a disc. We form X^* by removing a disc from X , but we just replace that disc when we attach $(S^2)^*$ to form $X\#S^2$, so $X\#S^2 = X$. \square

Another basic case is as follows:

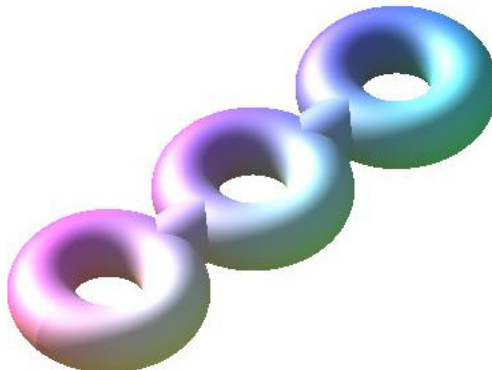
Proposition 9.4. *For any surface X , the surface $X\#D$ is homeomorphic to X^* (the space obtained by removing a disc from X).*

Proof. Note that D^* is just $D \setminus D'$, where D' is a smaller open disc nested in D . To form $X\#D$ we first remove a disc from X , then we attach $D \setminus D'$. The overall effect is the same as just removing a copy of D' from X , which is again X^* . \square

If we let T denote the torus, then $T\#T$ is as follows:



The left half is one copy of T^* , the right half is the other copy, and they are joined along the circle shown in red. Similarly, the surface $T^{\#3} = T \# T \# T$ looks like this:



We will show later that every closed, connected orientable surface is homeomorphic to $T^{\#n}$ for some $n \geq 0$.

10. FORMAL DEFINITIONS

We next give a more rigorous version of the definitions in Section 8. Most of this should be revision for most students, so we will be brief.

Surfaces are modelled on discs, defined as follows.

Definition 10.1. We put

$$\begin{aligned} D &= \{(x, y) \in \mathbb{R}^2 \mid x^2 + y^2 \leq 1\} = \text{the closed unit disc} \\ D' &= \{(x, y) \in \mathbb{R}^2 \mid x^2 + y^2 < 1\} = \text{the open unit disc} \\ D'_+ &= \{(x, y) \in D' \mid y \geq 0\} = \text{the standard open half-disc.} \end{aligned}$$

In a closed surface, every point should have a neighbourhood which can be identified with D' . We next give a formal definition for the kind of identification that we need to consider.

Definition 10.2. We recall that the standard definition of continuity works perfectly well for maps between subsets of \mathbb{R}^d : if $X \subseteq \mathbb{R}^n$ and $Y \subseteq \mathbb{R}^m$ and $f: X \rightarrow Y$, we say that f is *continuous* if for every convergent sequence $x_k \rightarrow a$ with all x_k and a in X , the sequence $f(x_k)$ converges to $f(a)$. We say that f is a *homeomorphism* if it is a bijection, and both $f: X \rightarrow Y$ and $f^{-1}: Y \rightarrow X$ are continuous. We say that X and Y are *homeomorphic* if there exists a homeomorphism from X to Y .

We next need a definition of “neighbourhood”.

Definition 10.3. Let x be a point in \mathbb{R}^n , and let ϵ be a positive real number. We then put

$$B_\epsilon(x) = \{y \in \mathbb{R}^n \mid \|x - y\| < \epsilon\}.$$

This is called the *open ball of radius ϵ around x* .

Definition 10.4. Let X be a subset of \mathbb{R}^n , and consider a set $U \subseteq X$. We say that U is *open in X* if for each $x \in X$ there exists $\epsilon > 0$ such that $X \cap B_\epsilon(x) \subseteq U$.

An *open neighbourhood* of a point $a \in X$ is a set $U \subseteq X$ such that U is open in X and $a \in U$.

Definition 10.5.

- (a) A *closed surface* is a compact subset $X \subset \mathbb{R}^n$ (for some n) such that every point $a \in X$ has an open neighbourhood that is homeomorphic to D' .
- (b) A *surface with boundary* is a compact subset $X \subset \mathbb{R}^n$ (for some n) such that every point $a \in X$ has an open neighbourhood that is either homeomorphic to D' or homeomorphic to D'_+ .

Definition 10.6. Let X be a closed surface. A *disconnection* of X is a pair of subsets $Y, Z \subseteq X$ such that

- Y is itself a closed surface, and so is Z

- $Y \cup Z = X$ and $Y \cap Z = \emptyset$
- Neither Y nor Z is empty.

We say that X is *connected* if no such disconnection exists.

We will mostly be interested in connected surfaces.

11. PIECEWISE LINEAR VERSIONS

In practice, we will mostly work with piecewise linear surfaces, which we now introduce.

Definition 11.1. Let a_1, a_2 and a_3 be vectors in \mathbb{R}^n . We say that they are in *general position* if the vectors $a_2 - a_1$ and $a_3 - a_1$ are linearly independent, so there is no line that contains all three points. If so, we put

$$\Delta(a_1, a_2, a_3) = \{t_1 a_1 + t_2 a_2 + t_3 a_3 \mid t_1, t_2, t_3 \geq 0, t_1 + t_2 + t_3 = 1\}.$$

This is a triangle in \mathbb{R}^n with vertices a_i . The three edges are

$$\begin{aligned} E_1 &= \{t_2 a_2 + t_3 a_3 \mid t_2, t_3 \geq 0, t_2 + t_3 = 1\} \\ E_2 &= \{t_1 a_1 + t_3 a_3 \mid t_1, t_3 \geq 0, t_1 + t_3 = 1\} \\ E_3 &= \{t_1 a_1 + t_2 a_2 \mid t_1, t_2 \geq 0, t_1 + t_2 = 1\}. \end{aligned}$$

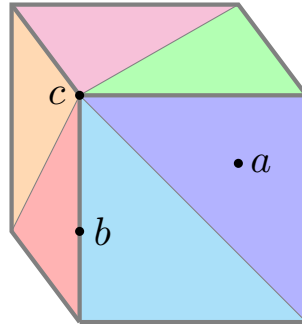
Definition 11.2. Let X be a subset of \mathbb{R}^n . A *linear surface triangulation* of X is a list of triangles T_1, \dots, T_N as above, such that

- For each $i \neq j$, the intersection $T_i \cap T_j$ is either empty; or a single point which is a vertex of T_i and also a vertex of T_j ; or a line segment which is an edge of T_i and also an edge of T_j .
- For each triangle T_i and each edge $E \subset T_i$, there is at most one other triangle T_j such that E is also an edge of T_j . (We say that E is an *interior edge* if there exists a triangle T_j as above; otherwise we say that E is a *boundary edge*.)
- For any vertex V in X , the set of triangles containing V can be listed as T_{i_1}, \dots, T_{i_m} in such a way that T_{i_k} shares an edge with $T_{i_{k+1}}$, and T_{i_m} may or may not share an edge with T_{i_1} , and there are no other common edges. (We say that V is an *interior vertex* if T_{i_m} shares an edge with T_{i_1} ; otherwise we say that V is a *boundary vertex*.)

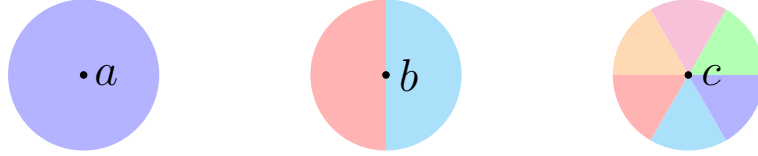
We say that a linear surface triangulation is *closed* if there are no boundary edges (and therefore no boundary vertices.)

A *piecewise-linear surface with boundary* is a subset $X \subseteq \mathbb{R}^n$ for which there exists a linear surface triangulation. A *piecewise-linear closed surface* is a subset for which there exists a closed linear surface triangulation.

Example 11.3. Let X be the surface of a cube. This consists of six squares. We can draw a diagonal line across each square to divide it into two triangles. The resulting twelve triangles give a linear surface triangulation of X .



Each of the points a , b and c has a disc neighbourhood:



It should be clear that the above construction works for any PL closed surface. Axiom (a) says that the triangles do not interfere with each other except on the boundary, so for any point in a triangle that is not on an edge, a small piece of the triangle can be used as a disc neighbourhood. For a point on an edge that is not a vertex, Axiom (b) says that there will be precisely two triangles that contain the relevant edge, and we can take a half-disc from each of these triangles to get a full disc neighbourhood of the point. Finally, Axiom (c) is precisely what we need to get a disc neighbourhood for each vertex. Thus, any PL closed surface is a closed surface, as the name suggests. Similarly, a PL surface with boundary is a surface with boundary.

Theorem 11.4. *Let X be any closed surface. Then there is a PL closed surface Y such that Y is homeomorphic to X .*

The proof of this theorem is quite hard, and we will not discuss it. However, it is reasonably easy to see that it is true in specific cases. In fact, if you look closely at the computer pictures of smoothly curved surfaces, you will see that the computer has actually used triangles, and has drawn a PL surface which is a close approximation to the surface we first considered.

For the rest of this course, we will work with PL closed surfaces.

12. SURFACE WORDS

Our main aim is to classify surfaces up to homeomorphism. The first step is to introduce surface words, which are a kind of combinatorial recipe for building surfaces by starting with a disc and gluing together certain parts of the boundary. We will show that every connected surface can be obtained in this way, if we choose a suitable surface word. Our first choice of surface word may be very complicated, but we will also show that the word can be simplified in various ways, without changing the corresponding surface. (This is loosely analogous to simplifying knot diagrams by Reidemeister moves.)

Definition 12.1. A *surface word* is a sequence of barred or unbarred letters, in which each letter (ignoring bars) occurs at most twice. Such a word is *closed* if every letter occurs precisely twice. It is *nonorientable* if there is a letter which occurs twice without a bar or twice with a bar; otherwise it is *orientable*.

Remark 12.2. In various places we will refer to \bar{x} , where x might already be a barred letter. This should be interpreted by the rule $\bar{\bar{a}} = a$.

Example 12.3.

- The *sphere word* S is the empty word, with no letters.
- The *torus word* is $T = xy\bar{x}\bar{y}$.
- The *projective plane word* is $P = xx$.
- The *Klein bottle word* is $K = xy\bar{x}y$.
- The *disc word* is $D = x$.
- The *cylinder word* is $C = xy\bar{x}z$.
- The *Möbius word* is $M = xyxz$.

Note that

- S , T , P and K are closed, but the others are not.
- S , T , C and D are orientable, but the others are not.

Example 12.4. The sequence $xy\bar{x}\bar{y}x$ is not a surface word, because x occurs three times (once barred and twice unbarred).

Definition 12.5. Given a surface word W , we construct a space $\Sigma(W)$ as follows. Let n be the number of letters in W . If $n = 0$ (so W is the empty word) we define $\Sigma(W)$ to be the sphere S^2 . Otherwise, we take the disc

$$D = \{(x, y) \in \mathbb{R}^2 \mid x^2 + y^2 \leq 1\},$$

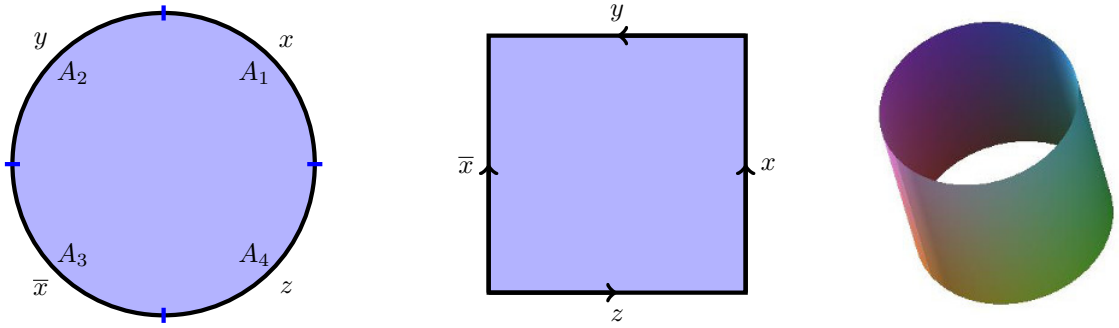
and we divide the boundary circle into n equal arcs

$$A_k = \{(\cos(\theta), \sin(\theta)) \mid k-1 \leq \frac{n\theta}{2\pi} \leq k\}$$

for $1 \leq k \leq n$. We then define $\Sigma(W)$ to be the space obtained by identifying A_j with A_k whenever the j 'th and k 'th letters in W are the same. More precisely, if the j 'th and k 'th letters are the same and both are barred or both are unbarred, then we identify A_j with A_k in the same direction, but if one of the letters has a bar and the other does not, then we identify A_j with A_k in the opposite direction.

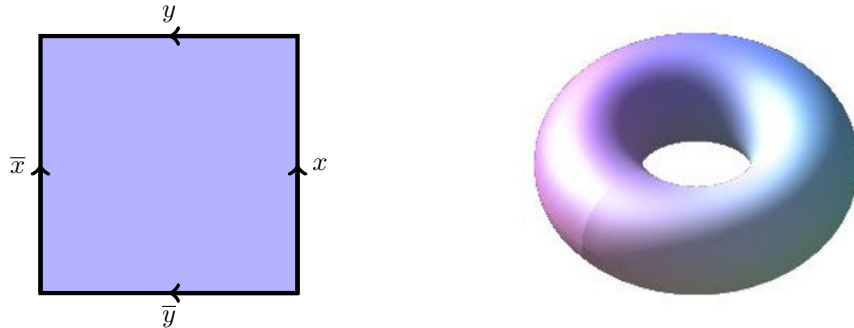
Remark 12.6. It is often more convenient to draw a polygon with n straight sides, instead of a disc with n curved arcs. This is fine as long as $n \geq 3$.

Example 12.7. Consider the cylinder word $C = xy\bar{x}, z$. We can draw $\Sigma(C)$ as shown on the left below.



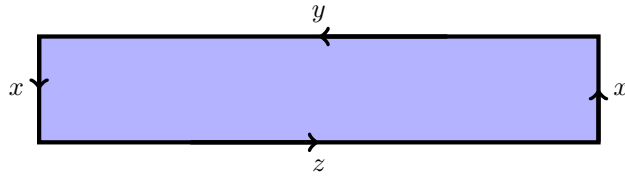
The arcs A_1 and A_3 are labelled x and \bar{x} , so they are glued with reversed directions. It is easier to see the effect of this if we redraw the picture as shown in the middle. There, we have rotated the picture and straightened the sides to make a square. Each arc with an unbarred letter has been given an anticlockwise arrow, and each arc with a barred letter has been given a clockwise arrow. The rule is now that sides with matching labels get glued in the direction indicated by the arrows. Attaching the two x 's together gives a cylinder. Thus, $\Sigma(C)$ is a cylinder, which justifies the name “cylinder word” for C .

Example 12.8. For the torus word $T = xy\bar{x}\bar{y}$, we have the following picture.



Attaching the two x 's together gives a cylinder. The two edges marked y become the top and bottom circles of the cylinder, and when we glue them together we get a torus. Thus, $\Sigma(T)$ is a torus, which justifies the name “torus word” for T .

Example 12.9. For the Möbius word $M = xyxz$, it is convenient to stretch the picture horizontally:



Here we need to twist the strip before connecting the ends together, in order to make the arrows match up. This gives a Möbius band.

Example 12.10. For the sphere word S , the space $\Sigma(S)$ is a sphere by definition. For the disc word $D = x$, the space $\Sigma(D)$ consists of a disc with a single boundary arc A_1 and no gluing (because there are no repeated letters in the word). In other words, $\Sigma(D)$ is a disc.

Example 12.11. Now consider the projective plane $\mathbb{R}P^2$. If $a \in \mathbb{R}P^2$, we can choose a point $u \in S^2 \subset \mathbb{R}^3$ with $q(u) = q(-u) = a$. If the z -coordinate of u is negative then we put $v = -u$, otherwise we put $v = u$. We now have a point v in the upper hemisphere S^2_+ of S^2 such that $q(v) = a$. If v does not lie on the equator, then v is the *only* point in S^2_+ such that $q(v) = a$. However, if v lies on the equator, then $-v$ also lies on the equator, and $q(-v) = q(v)$. This means that $\mathbb{R}P^2$ can be obtained from S^2_+ by identifying v with $-v$ whenever v lies on the equator.

Note also that S^2_+ can be identified with the closed disc D by the map $(x, y) \mapsto (x, y, \sqrt{1 - x^2 - y^2})$. The equator corresponds to the boundary circle $S^1 \subset D$, consisting of points $v = (\cos(\theta), \sin(\theta))$. In this picture, we have

$$-v = (-\cos(\theta), -\sin(\theta)) = (\cos(\pi + \theta), \sin(\pi + \theta)).$$

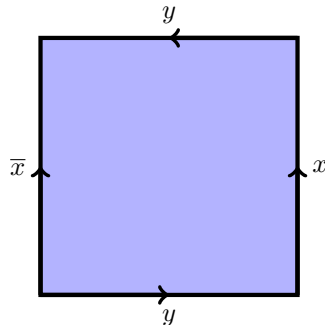
Thus, if we divide S^1 into two arcs

$$A_1 = \{(\cos(\theta), \sin(\theta)) \mid 0 \leq \theta \leq \pi\}$$

$$A_2 = \{(\cos(\theta), \sin(\theta)) \mid \pi \leq \theta \leq 2\pi\},$$

then A_1 gets identified with A_2 , with no change of direction. This means that $\mathbb{R}P^2 \simeq \Sigma(xx) = \Sigma(P)$.

Example 12.12. The picture on the left shows the Klein bottle word $K = xy\bar{x}y$. After identifying the two edges marked x , we have a cylinder. We then need to identify the two end circles, but without reversing the orientation. This cannot be done in \mathbb{R}^3 without self-intersections. We need to pass one end of the cylinder through the wall of the cylinder, then we can connect the two ends together while preserving the direction of the arrows.

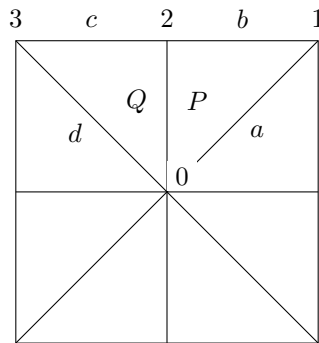


This gives a self-intersecting surface in \mathbb{R}^3 , as shown on the right. In higher dimensions we can perform the gluing without creating self-intersections; this gives a surface called the *Klein bottle*.

Lemma 12.13. For any surface word W , the space $\Sigma(W)$ can be realised as a simplicial complex in \mathbb{R}^N for some N .

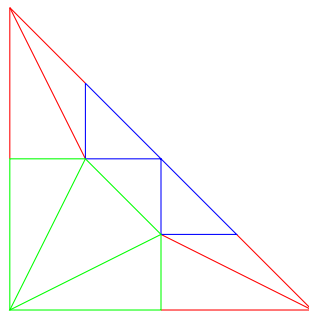
Proof. The basic idea is to divide the disc up into triangles, and then embed each triangle linearly in \mathbb{R}^N . If two arcs in the disc have the same label, then we should place the corresponding triangles in \mathbb{R}^N in such a way that the relevant edges match up. This will give us an embedding of $\Sigma(W)$ in \mathbb{R}^N .

However, the above process will not work if we use too few triangles. To see this, consider the following subdivision of the square.

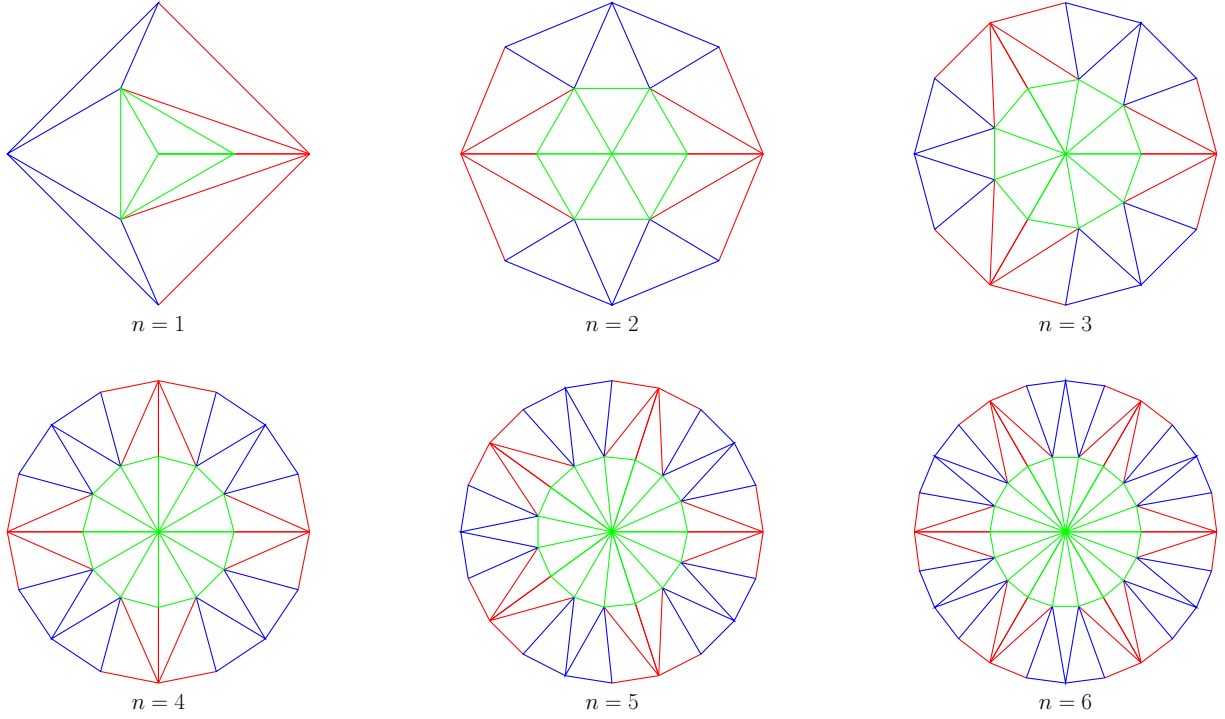


If we wrap this up to make a torus, then point 1 gets identified with point 3, so arcs a and d have the same endpoints, and arcs b and c have the same endpoints, so regions P and Q have the same corners. This means that we cannot embed the torus in \mathbb{R}^N in such a way that regions P and Q become flat triangles, because any two flat triangles with the same corners are necessarily the same.

To avoid this kind of problem, we need to divide the disc more finely. If W has n letters, then we first divide the disc into n wedges, then we divide each wedge into ten triangles according to the following scheme:



The effect for small values of n is as follows:



One can check that this subdivision has the following property: if e_0 and e_1 are two different edges, and the endpoints of e_0 get identified in $\Sigma(W)$ with the endpoints of e_1 , then the whole edge e_0 gets identified with e_1 . Moreover, if t_0 and t_1 are two different triangles, then it never happens that all the vertices of t_0 get identified with the vertices of t_1 . Thus, we have a subdivision of $\Sigma(W)$ into triangles with no strange anomalies: every edge has two distinct endpoints, every triangle has three distinct corners, there are never two different edges with the same endpoints, and there are never two different triangles with the same corners.

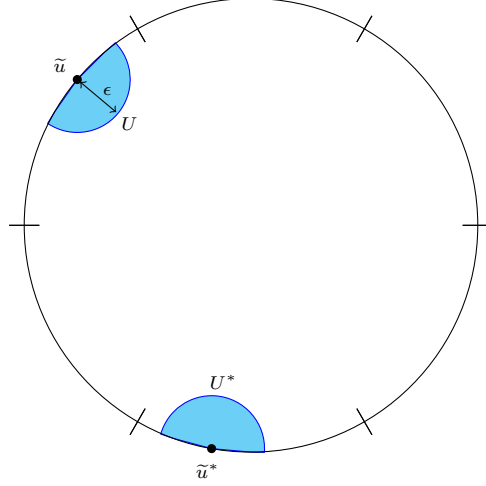
Let w_1, \dots, w_M be the vertices in the above subdivision of the disc. Some of these will be identified together in $\Sigma(W)$, so $\Sigma(W)$ will have a smaller number of vertices, say v_1, \dots, v_N . Let e_1, \dots, e_N be the standard basis of \mathbb{R}^N . Now consider a point $u \in \Sigma(W)$. Choose a corresponding point \tilde{u} in the disc. (There may only be one possible choice, or there may be several possibilities.) This point will lie in one of our triangles, with vertices w_i, w_j and w_k say. This means that $u = a_i w_i + a_j w_j + a_k w_k$ for some coefficients $a_i, a_j, a_k \geq 0$ with $a_i + a_j + a_k = 1$. There is an index i' such that w_i becomes $v_{i'}$ in $\Sigma(W)$, and similarly for j' and k' . We define $f(u) = a_i e_{i'} + a_j e_{j'} + a_k e_{k'} \in \mathbb{R}^N$. (We need to check that this only depends on u and not on the choice of \tilde{u} , but that is precisely the point of making a fine subdivision, as discussed above.) This gives an embedding $f: \Sigma(W) \rightarrow \mathbb{R}^N$, whose image is a simplicial complex, as required. \square

Proposition 12.14. *For any closed surface word W , the space $\Sigma(W)$ is a closed surface. For any non-closed surface word W , the space $\Sigma(W)$ is a surface with boundary.*

Proof. The space $\Sigma(W)$ is constructed from the disc D by identifying certain points on the boundary. This means that there is a surjective map $q: D \rightarrow \Sigma(W)$. Consider a point $u \in \Sigma(W)$, so we can choose a point $\tilde{u} \in D^2$ with $q(\tilde{u}) = u$.

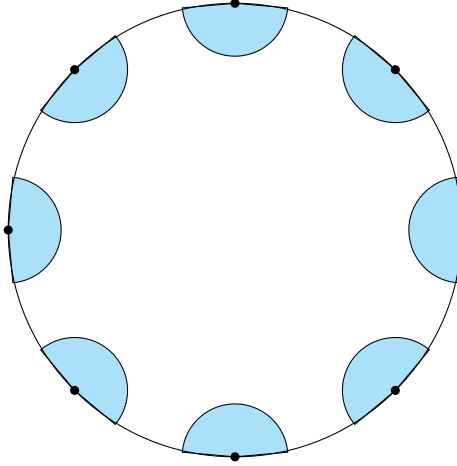
The simplest case is when \tilde{u} lies in the open disc D' . As there are no identifications between points in D' , we see that $q: D' \rightarrow q(D')$ is a homeomorphism, and $q(D')$ is a disc neighbourhood of u , as required.

Next, suppose that u lies on one of the boundary arcs A_j , but not it is not one of the endpoints of A_j . Let ϵ be less than the minimum distance from \tilde{u} to the endpoints of A_j , and put $U = B_\epsilon(\tilde{u}) \cap D$. As we see in the picture below, U is homeomorphic to the half-disc D'_+ .



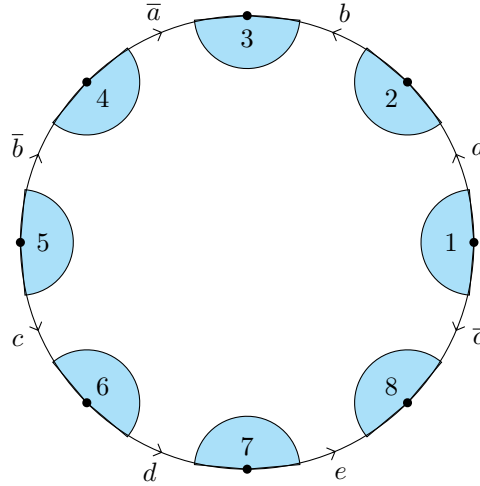
If the arc A_j has no other arc with the same letter, then $q(U)$ will be a half-disc neighbourhood of u in $\Sigma(W)$, as required for a surface with boundary. On the other hand, if A_j has the same letter as A_k , then there will be another point \tilde{u}^* in A_k that gets identified with \tilde{u} , and we can define $U^* = B_\epsilon(\tilde{u}^*) \cap D$, which is a half-disc neighbourhood of \tilde{u}^* . The half-discs U and U^* get glued together in $\Sigma(W)$ to form a disc neighbourhood of u , as required for a closed surface.

Finally, we need to consider the case where \tilde{u} is one of the endpoints of one of the arcs A_j . To understand this case, let P be the set consisting of a small half-disc around each each of these endpoints, as illustrated below.

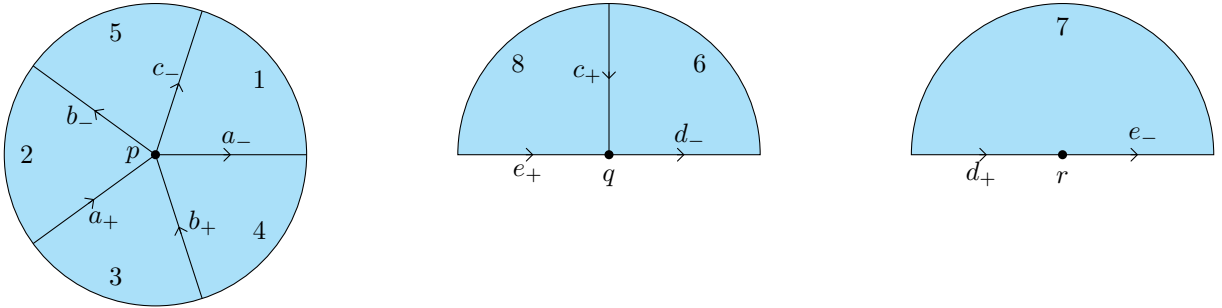


We then focus on the effect of the gluing rules on P (noting that no point in P ever gets identified with any point outside P). Each half-disc has two short straight edges and one longer curved edge. When forming $\Sigma(W)$, we glue some of the short edges together in pairs. As each letter occurs at most two times in W , we never glue three different edges together. Moreover, the curved edges are never glued. It is not hard to see that there are only two possible patterns for the outcome: each component is either a full pizza or a half pizza, divided into a number of slices (possibly only one slice). Each vertex appears as the centre of one of these pizzas, so it has a disc neighbourhood (for a full pizza) or a half-disc neighbourhood (for a half-pizza). Moreover, if all edges occur in pairs then we can always keep gluing extra slices until we get a complete pizza, so in this context every vertex has a full disc neighbourhood, and we have a closed surface. \square

Example 12.15. Consider the word $W = ab\bar{a}\bar{b}cde\bar{e}$. This can be drawn as follows:



The set P is the union of the shaded discs, which we have labelled 1 to 8. We will also write a_- for the first third of the edge a , and a_+ for the last third, in the direction indicated by the arrow. The two short edges of region 1 are a_- and c_- . The edge a_- also appears as one of the short edges of region 4, so region 4 gets glued to region 1. The other short edge of region 4 is b_+ , which also appears in region 3, so region 3 gets glued to region 4. In the same way, region 2 is glued to region 3 along a_+ , and then region 2 is glued to region 5 along b_- , and region 5 is glued to the remaining short edge of region 1 along c_- . This gives a complete pizza with a single vertex in the middle, which we call p . The remaining regions form two half-pizzas with centres q and r say. We now see that p has a full disc neighbourhood, and q and r have half-disc neighbourhoods.



Proposition 12.16. *Let X be any connected PL surface (closed or with boundary). Then there is a surface word W such that X is homeomorphic to $\Sigma(W)$.*

It will be more convenient to prove this after we have discussed some additional techniques, so we will defer the proof.

13. WORD MOVES

The next issue is to understand some cases where we have different words V and W , but $\Sigma(V)$ is homeomorphic to $\Sigma(W)$.

Definition 13.1. For a surface word W , we define \overline{W} to be the word obtained by reversing the order, adding bars to the unbarred letters, and removing bars from the barred letters.

Example 13.2. If $W = abc\bar{a}\bar{c}$ then $\overline{W} = ca\bar{c}\bar{b}\bar{a}$.

Definition 13.3. The surface word moves are as follows:

- (a) Replace UV by VU (rotation)
- (b) Replace $Ux\bar{x}V$ by UV (cancellation)
- (c) Replace $TxUV\bar{x}W$ by $TxVU\bar{x}W$ (inner rotation)
- (d) Replace $TxUxV$ by $Txx\overline{UV}$ (crosscap move)

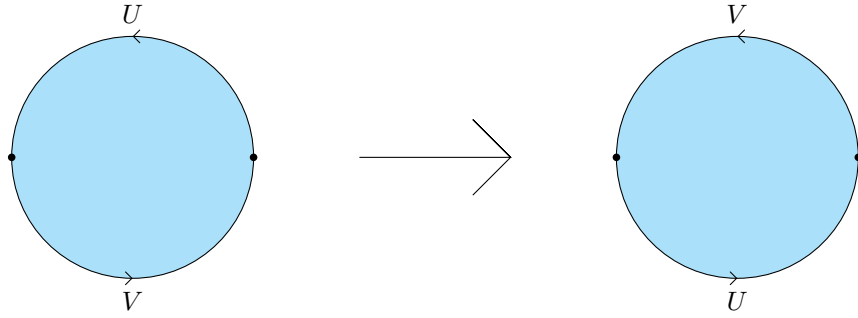
- (e) If x and y occur only once, replace $UxyV$ with UxV (merger).
- (f) Replace all instances of x by y , and all instances of \bar{x} by \bar{y} (relabelling)

Here T, U, V and W are assumed to be surface words. In (b), (c), (d) and (e) we assume that x is a barred or unbarred letter which does not occur (in barred or unbarred form) in T, U, V or W . In move (f), we assume that x is a letter which occurs in the relevant word W , and y is another letter. Usually we insist that y does not already occur in W , except that we allow the case where $y = \bar{x}$, so the move replaces x by \bar{x} and \bar{x} by x .

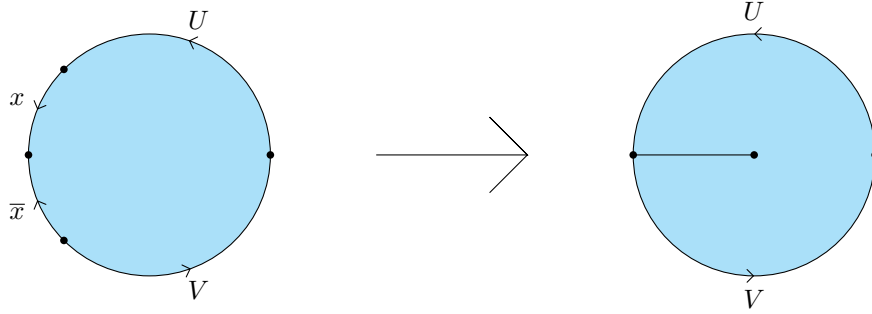
Remark 13.4. Inner rotations are also called *handle moves*.

Proposition 13.5. *Let A and B be surface words. If A and B are W -equivalent, then $\Sigma(A)$ is homeomorphic to $\Sigma(B)$.*

Proof. It will be enough to prove this when B is obtained from A by a single surface word move. For a rotation move (type (a)), this is clear, because we can just rotate the disc:

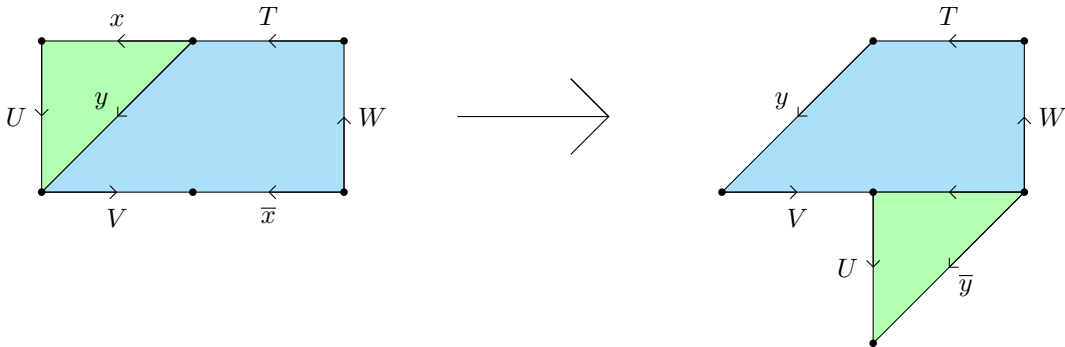


For a cancellation move (type (b)), we have the picture shown on the left below.



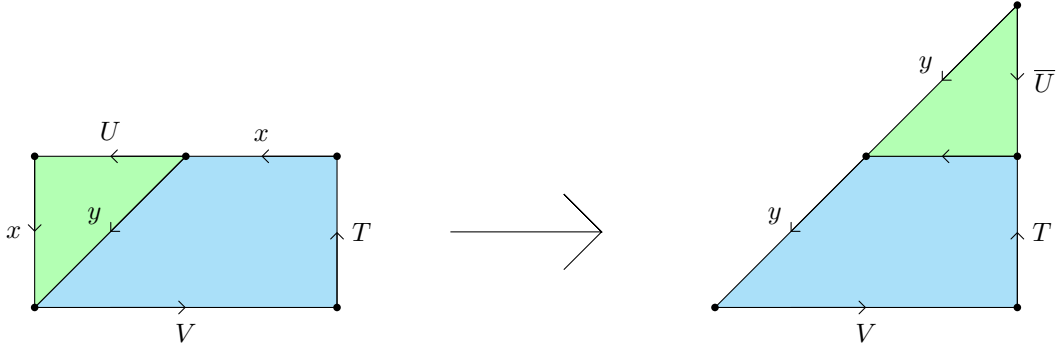
We can first glue together the two copies of x to give the picture shown on the right, then perform any additional identifications encoded by UV ; this makes it clear that $\Sigma(Ux\bar{x}V) \simeq \Sigma(UV)$.

Now consider an inner rotation $TxUV\bar{x}W \mapsto TxVU\bar{x}W$. The left hand picture below shows $TxUV\bar{x}W$, but with an extra edge across the middle of the disc, labelled y .



We can cut along y and glue together the two copies of x to get the right hand picture. As the two x 's have been glued, we no longer need to label them. Reading around the edge of the right hand picture, we have the word $TyVU\bar{y}W$. If identify the edges according to this word, we reverse the effect of cutting along y , and then perform exactly the same identifications as for the original word. We thus have a homeomorphism $\Sigma(TxUV\bar{x}W) \simeq \Sigma(TyVU\bar{y}W)$. Moreover, the letter x no longer appears anywhere, so we can replace y by x without creating any clashes, and we see that $\Sigma(TxUV\bar{x}W) \simeq \Sigma(TxVU\bar{x}W)$ as claimed.

The argument for a crosscap move is essentially the same as for an inner rotation, and is illustrated by the diagram below. The only difference is that we need to turn the triangle over before we can reattach it, and this converts U to \bar{U} .



For a merger move $UxyV \mapsto UxV$, we just note that two edges can be joined end to end to give a single edge. If we were gluing these edges to other edges, then we might disrupt the gluing pattern by joining the edges together. However, in a merger move it is assumed that x and y only occur once, so they are not glued to anything else, and no problems can arise.

Finally, as the letters serve only to indicate which edges will be glued to each other, it is clear that relabelling does not change the surface. \square

Remark 13.6. When explaining sequences of surface word moves, we will use brackets and lines to indicate which move we want to do next. In more detail:

- (a) We write $U|V$ to indicate that we are about to do a rotation and obtain VU .
- (b) We write $U(x\bar{x})V$ to indicate that we are about to do a cancellation and obtain UV . For the reverse, we write $U()V$.
- (c) We write $Tx(U|V)\bar{x}W$ to indicate that we are about to do an inner rotation and obtain $TxVU\bar{x}W$.
- (d) We write $Tx[U]xV$ to indicate that we are about to do a crosscap move and obtain $Txx\bar{U}V$. For the reverse, we write $Txx[\bar{U}]V$.
- (e) We write $U\{xy\}V$ to indicate that we are about to do a merger and obtain UxV . For the reverse, we write $U\{x\}V$.
- (f) We write $U \cong V$ if V is obtained from U by relabelling.

Definition 13.7. Let A and B be surface words. We say that A and B are W -equivalent if there is a sequence of surface words C_0, \dots, C_m such that $C_0 = A$ and $C_m = B$ and C_{i+1} is obtained from C_i by applying a surface word move, or *vice versa*. We write $A \sim B$ to indicate that A and B are W -equivalent.

We can generalise the internal rotation slightly by combining it with ordinary rotation:

Lemma 13.8. *There are W -equivalences $xT\bar{x}UV \sim xT\bar{x}VU$ and $UVxT\bar{x} \sim VUxT\bar{x}$.*

Proof.

$$\begin{aligned} x|T\bar{x}UV &\sim T\bar{x}(U|V)x \sim T\bar{x}VU|x \sim xT\bar{x}VU \\ UVxT|\bar{x} &\sim \bar{x}(U|V)xT \sim \bar{x}|VUxT \sim VUxT\bar{x}. \end{aligned}$$

\square

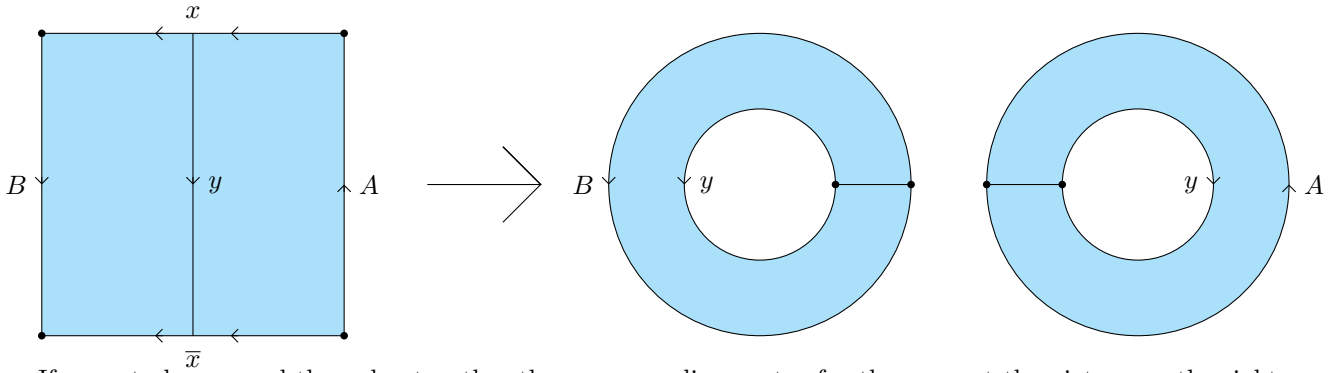
We will write $xT\bar{x}(U|V)$ and $(U|V)xT\bar{x}$ for the above moves, as a slight generalisation of Remark 13.6.

Definition 13.9. Let A and B be surface words that have no letters in common. We then define $A\#B = AxB\bar{x}$, where x is any letter not occurring in A or B . We call this the *connected sum* of A and B . If A and B do have letters in common, we define $A\#B$ to be the word obtained by renaming letters to avoid the clash, and then performing the above construction.

Remark 13.10. Some sources define the connected sum to be AB instead of $AxB\bar{x}$. It turns out that when A and B are closed surface words then $AxB\bar{x}$ is W -equivalent to AB . However, this is not easy to prove, and it is false when A and B are not closed. Moreover, some properties of the connected sum are easy to prove from our definition, but much harder from the alternative definition.

Proposition 13.11. $\Sigma(A\#B)$ is homeomorphic to $\Sigma(A)\#\Sigma(B)$.

Proof. The picture on the left shows $A\#B = AxB\bar{x}$, together with an extra edge marked y across the middle of the disc.



If we cut along y and then glue together the corresponding parts of x then we get the picture on the right. If we perform the identifications encoded by B , then the first ring becomes $\Sigma(B)$ with a hole removed, or in other words $\Sigma(B)^*$. Similarly, if we perform the identifications encoded by A then the right hand ring becomes $\Sigma(A)^*$. Thus, when we glue the two copies of y back together, we get $\Sigma(A)\#\Sigma(B)$. \square

Proposition 13.12. Suppose that A, B, A' and B' are surface words with $A \sim A'$ and $B \sim B'$; then $A\#B \sim A'\#B'$.

Proof. We first prove that $A\#B \sim A\#B'$. It will be enough to do this when B' is obtained from B by a single move. We may also assume that everything has been relabelled if necessary so that there are no accidental coincidences of letters.

- Suppose that B' is obtained from B by rotation, say $B = UV$ and $B' = VU$. Then $A\#B = AxUV\bar{x}$, and we can perform internal rotation to get $AxUV\bar{x} \sim AxVU\bar{x} = A\#B'$. In the notation of Remark 13.6, we have

$$A\#B = Ax(U|V)\bar{x} \sim AxVU\bar{x} = A\#B'$$

- For a cancellation $B = U(y\bar{y})V \sim UV = B'$, we have

$$A\#B = AxU(y\bar{y})V\bar{x} \sim AxUV\bar{x} = A\#B'.$$

- For an internal rotation $B = Ty(U|V)\bar{y}W \sim TyVU\bar{y}W = B'$, we have

$$A\#B = AxTy(U|V)\bar{y}\bar{x} \sim AxTyVU\bar{y}\bar{x} = A\#B'.$$

- For a crosscap move $B = Ty[U]yV \sim Ty\bar{y}\bar{U}V = B'$, we have

$$A\#B = AxTy[U]yV\bar{x} \sim AxTy\bar{y}\bar{U}V\bar{x} = A\#B'.$$

- For a merger $B = U\{yz\}V \sim UyV$ we have

$$A\#B = AxU\{yz\}V\bar{x} \sim AxUyV\bar{x} = A\#B'.$$

- It is also clear that the connected sum is compatible with relabelling.

This proves that $A\#B \sim A\#B'$, and essentially the same argument shows that $A\#B' \sim A'\#B'$, so $A\#B \sim A'\#B'$, as claimed. \square

Remark 13.13. It is easy to see that

- $U\#V$ is closed iff U and V are both closed.
- $U\#V$ is orientable iff U and V are both orientable.

Proposition 13.14. *The connected sum operation is commutative, associative and unital up to W -equivalence. In more detail:*

- (a) *For any surface word U we have $S\#U \sim U \sim U\#S$.*
- (b) *For any surface words U and V we have $U\#V \sim V\#U$.*
- (c) *For any surface words U, V and W we have $U\#(V\#W) \sim (U\#V)\#W$.*

Proof.

- (a) $S\#U = x|U\bar{x} \sim U(\bar{x}x) \sim U$, and $U\#S = U(x\bar{x}) \sim U$.
- (b) $U\#V = Ux|V\bar{x} \sim V\bar{x}Ux \cong VxU\bar{x} = V\#U$.
(Recall here that \cong indicates relabelling; in this case we have replaced x by \bar{x} .)
- (c) Put $M = pU\bar{p}qV\bar{q}rW\bar{r}$. We will prove that M is equivalent to both $U\#(V\#W)$ and $(U\#V)\#W$.

$$\begin{aligned} M &= p|U\bar{p}qV\bar{q}rW\bar{r} \sim U\bar{p}(q|V\bar{q}rW\bar{r})p \\ &\sim U\bar{p}V\bar{q}(rW|\bar{r})qp \sim U\bar{p}V\bar{q}(\bar{r}r)Wqp \\ &\sim U\bar{p}V\bar{q}Wqp \cong UxVyW\bar{y}\bar{x} \\ &= U\#(V\#W) \end{aligned}$$

$$\begin{aligned} M &= pU\bar{p}qV\bar{q}r|W\bar{r} \sim W\bar{r}(p|U\bar{p}qV\bar{q})r \\ &\sim W\bar{r}U\bar{p}(qV|\bar{q})pr \sim W\bar{r}U\bar{p}(\bar{q}q)Vpr \\ &\sim W\bar{r}|U\bar{p}Vpr \sim U\bar{p}VprW\bar{r} \cong WxU\bar{x}yW\bar{y} \\ &= (U\#V)\#W. \end{aligned}$$

□

Definition 13.15. We write $U^{\#n}$ for $U\#U\#\dots\#U$ (with n terms). This should be interpreted as the empty word S when $n = 0$.

Proposition 13.16. *Let $T = xy\bar{x}\bar{y}$ and $P = xx$ as before, and let U be any surface word. Then $TU \sim T\#U$ and $PU \sim P\#U$.*

Proof.

$$\begin{aligned} P\#U &= xxyU|\bar{y} \sim \bar{y}xx[y]U \sim \bar{y}[x]\bar{y}xU \sim \bar{y}\bar{y}(\bar{x}x)U \\ &\sim \bar{y}\bar{y}U \cong xxU = PU \\ T\#U &= xy|\bar{x}\bar{y}zU\bar{z} \sim \bar{x}(\bar{y}zU|\bar{z})xy \\ &\sim (\bar{x}|\bar{z})\bar{y}zUxy \sim \bar{z}\bar{x}\bar{y}(zU|x)y \\ &\sim \bar{z}(\bar{x}\bar{y}|x)zUy \sim \bar{z}(x\bar{x})\bar{y}zUy \\ &\sim \bar{z}\bar{y}zU|y \sim y\bar{z}\bar{y}zU \cong xy\bar{x}\bar{y}U \\ &= TU \end{aligned}$$

□

Proposition 13.17. *If U is a non-orientable surface word, then $U \sim P\#V$ for some word V which is shorter than U .*

Proof. As U is non-orientable, we can write it as $AxBxC$ for some A, B, C and x . Put $V = \bar{B}CA$, which is two letters shorter than U . We have

$$U = A|xBxC \sim x[B]xC A \sim x\bar{x}\bar{B}CA = PV \sim P\#V,$$

as required. □

Corollary 13.18. *For any surface word U , there is a natural number $n \geq 0$ and an orientable surface word V such that $U \sim P^{\#n} \# V$.*

Proof. If U itself is orientable, we just take $n = 0$ and $V = U$. Otherwise, the proposition tells us that $U \sim P \# U'$ for some U' which is shorter than U . By induction on the length, we can assume that $U' \sim P^{\#m} \# V$ for some orientable word V . This in turn gives $U \sim P^{\#(m+1)} \# V$, as required. \square

Proposition 13.19. *Let U be an orientable surface word with at least two letters. Then either*

- (a) *There is a word V that is strictly shorter than U such that $U \sim V$; or*
- (b) *There is a word V that is strictly shorter than U such that $U \sim D \# V$; or*
- (c) *There is a word V that is strictly shorter than U such that $U \sim T \# V$.*

Proof. If U contains no matching pairs, then all letters occur only once and we can perform a merger to shorten the word, as in case (a). Suppose instead that U contains a matching pair, and choose a pair which is as close together as possible. After rotating U if necessary we will have $U = xA\bar{x}B$. We can assume that A is not longer than B (otherwise we rotate some more, and consider $\bar{x}BxA$). If A is the empty word, then we can cancel x and \bar{x} to see that $U \sim B$, so case (a) is satisfied. If A contains only unmatched letters then we can merge them to get a single letter, say t , and we have

$$U \sim xt\bar{x}B \sim t\bar{x}Bx \cong txB\bar{x} = D \# B,$$

so case (b) is satisfied. This just leaves the case where A contains a matched letter, say y . It must be matched by a \bar{y} (not another y) because U is orientable. The \bar{y} cannot occur in A , because then the (y, \bar{y}) pair would be closer than the (x, \bar{x}) pair, which is contrary to our choice of x . The general form is thus

$$U \sim xKyL\bar{x}M\bar{y}N.$$

Put $V = NMLK$, which is shorter than U . We have

$$\begin{aligned} U &\sim x(K|yL)\bar{x}M\bar{y}N \sim xy(LK|\bar{x}M)\bar{y}N \\ &\sim xy\bar{x}(MLK|\bar{y}N) \sim xy\bar{x}\bar{y}NMLK \\ &= TV \sim T \# V. \end{aligned}$$

Thus, case (c) is satisfied. \square

Corollary 13.20. *If U is an orientable surface word, then $U \sim T^{\#n} \# D^{\#m}$ for some natural numbers $n, m \geq 0$. Moreover, U is closed if and only if $m = 0$.*

Proof. We prove this by induction on the length of U . If the length is zero then $U = S \sim T^{\#0} \# D^{\#0}$. If the length is one then $U = D \sim T^{\#0} \# D^{\#1}$. Otherwise, the length is at least two, and we can choose V as in the proposition. As V is shorter than U , the induction hypothesis means that $V \sim T^{\#j} \# D^{\#k}$ for some j and k . In case (a) we have $U \sim T^{\#j} \# D^{\#k}$, in case (b) we have $U \sim T^{\#j} \# D^{\#(k+1)}$, and in case (c) we have $U \sim T^{\#(j+1)} \# D^{\#k}$. \square

Proposition 13.21. $P \# T \sim P \# P \# P$

Proof. In view of Proposition 13.16, it will be enough to show that $PT \sim PPP$. Here we implicitly need to do some renaming before we join these words: the symbol PT really means $xyzy\bar{y}\bar{z}$, and PPP really means $xyyzz$. We have

$$\begin{aligned} PT &= (xx|y)z\bar{y}\bar{z} \sim yx[x|z]\bar{y}\bar{z} \\ &\sim yx\bar{z}[x\bar{y}]\bar{z} \sim yx[\bar{z}\bar{z}y\bar{x}] \\ &\sim \bar{z}\bar{z}y[\bar{x}]yx \sim \bar{z}\bar{z}yyxx \\ &\cong xxyyzz = PPP. \end{aligned}$$

\square

Corollary 13.22. *Let U be a nonorientable surface word. Then $U \sim P^{\#n} \# D^{\#m}$ for some integers $n > 0$ and $m \geq 0$.*

Proof. Using Corollaries 13.18 and 13.20, we see that $U \sim P^{\#i} \# T^{\#j} \# D^{\#k}$ for some $i, j, k \geq 0$. As U is not orientable, we must have $i > 0$. As at least one P is present, we can use Proposition 13.21 to replace each T by $P \# P$, giving $U \sim P^{\#(i+2j)} \# D^{\#k}$. \square

We stated as Proposition 12.16 that every connected closed PL surface is homeomorphic to $\Sigma(W)$ for some W . We will now prove this.

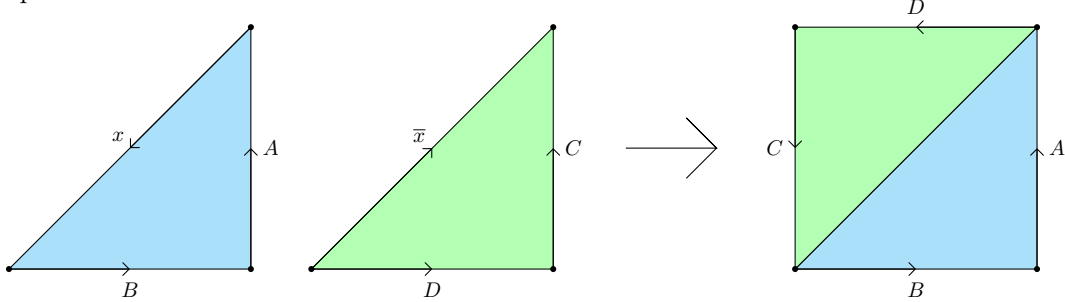
Proof of Proposition 12.16. By a *surface sentence* we mean a finite list of surface words, in which each letter occurs at most twice even if we take all the words together. Suppose we have a sentence $U = (W_1, \dots, W_d)$, where W_i has length n_i . We define a surface $\Sigma(U)$ by taking d separate discs D_1, \dots, D_d , dividing the boundary of D_i into n_i arcs, and then gluing arcs together when they have matching labels. Equivalently, we can start by forming the surfaces $\Sigma(W_1), \dots, \Sigma(W_d)$, and then glue together any boundary edges in these surfaces that have matching labels.

If X is any PL surface, then X can be formed from some list of triangles by gluing some of the edges together in pairs. This means that $X \simeq \Sigma(U)$ for some sentence U which consists of a word of length three for each triangle. Our problem is just to find a single word W such that $\Sigma(U) \simeq \Sigma(W)$.

We claim that the following operations do not change $\Sigma(U)$ up to homeomorphism.

- (a) We can apply surface word moves to individual words in the sentence.
- (b) We can replace a word W_i by the inverse \overline{W}_i .
- (c) If $W_i = AxB$ and $W_j = C\bar{x}D$ for some $j \neq i$, then we can replace W_i and W_j by the single word $ADCB$.

Indeed, (a) is clear, and (b) is valid because we can just turn over the disc D_i . Move (c) just involves gluing the two copies of x :



Now suppose we have a sentence U in which some letter x appears in two different words. After applying a move of type (b) if necessary, we can assume that it occurs with a bar in one word, and without a bar in the other word. We can thus combine the two words as in (c), giving a new sentence with fewer words that represents the same surface. After doing this repeatedly, we obtain a sentence $U' = (V_1, \dots, V_r)$ in which no letter appears in two different words. This means that there is no additional gluing, and the surface $\Sigma(U) \simeq \Sigma(U')$ is just the disjoint union of the surfaces $\Sigma(V_1), \dots, \Sigma(V_r)$. Thus, if the original surface is connected, then we must have $r = 1$, so U' consists of a single word, as required. \square

Theorem 13.23. *Let X be a connected PL surface (possibly with boundary).*

- (a) *If X is orientable, then it is homeomorphic to $T^{\#n} \# D^{\#m}$ for some $n, m \geq 0$ (where T is the torus and D is the disc). Moreover, X is closed if and only if $m = 0$.*
- (b) *If X is nonorientable, then it is homeomorphic to $P^{\#n} \# D^{\#m}$ for some $n > 0$ and $m \geq 0$ (where P is the projective plane). Moreover, X is closed if and only if $m = 0$.*

Proof. By Proposition 12.16, there is a word W such that $X \simeq \Sigma(W)$. Using Corollary 13.20 or Corollary 13.22 we see that there is a standard word W' of the form $T^{\#n} \# D^{\#m}$ or $P^{\#n} \# D^{\#m}$ that is equivalent to W . Proposition 13.5 tells us that X is again homeomorphic to $\Sigma(W')$. Proposition 13.11 also tells us that the connected sum of words gives the connected sum of surfaces, and the claim follows from that. \square

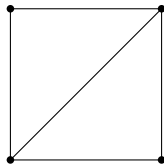
14. THE EULER CHARACTERISTIC

There is still one thing missing from our classification of surfaces. To explain this, let us restrict attention to closed orientable surfaces. We have shown that every such surface is homeomorphic to $T^{\#n}$ for some n .

However, we have not proved that n is unique: as far as we currently know, $T^{\#7}$ could be homeomorphic to $T^{\#11}$, for example. To prove that this is not the case, we need a new invariant, called the *Euler characteristic*. This is a number $\chi(X) \in \mathbb{Z}$ associated to a space X , which can be defined in various different ways. Different definitions work for different kinds of spaces, but any two definitions will give the same answer on any space for which they are both defined. The most sophisticated definitions (using algebraic topology) work for a very general class of spaces, but they are hard to set up. We will content ourselves with a more basic version.

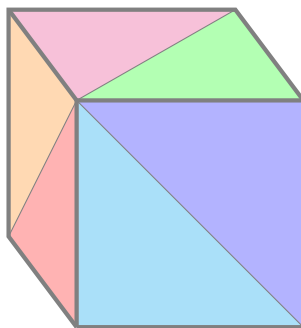
Definition 14.1. Let X be a subset of \mathbb{R}^N , and let T_1, \dots, T_r be a linear surface triangulation of X , as in Definition 11.2. Let p be the number of distinct vertices among these triangles, and let q be the number of distinct edges. We define $\chi = p - q + r$, and call this the *Euler characteristic* of the triangulation.

Example 14.2. Consider the following triangulation of a square:



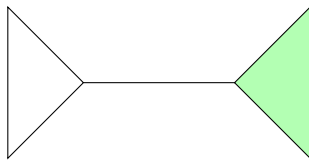
There are four vertices, five edges and two triangles, giving $(p, q, r) = (4, 5, 2)$ and $\chi = p - q + r = 1$.

Example 14.3. Now consider the surface of a cube, triangulated as follows.



There are eight vertices, one at each corner of the cube, so $p = 8$. There are six faces divided into two triangles each, giving $r = 12$. There are four edges around the top square, four around the bottom square, four vertical edges, and six diagonal edges cutting across the faces; this gives $q = 18$ in total. We therefore have $\chi = 8 - 18 + 12 = 2$.

Remark 14.4. Our definition of the Euler characteristic actually works for any subspace $X \subseteq \mathbb{R}^N$ that is composed of vertices, edges and triangles that fit together nicely, even if X is not a surface. For example, consider the following space:

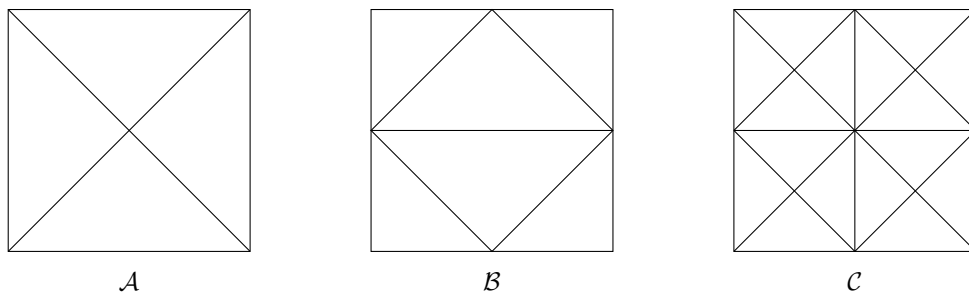


(The interior of the right hand triangle is part of the space, but the interior of the left hand triangle is not.) We have $p = 6$ and $q = 7$ and $r = 1$ so $\chi = 0$.

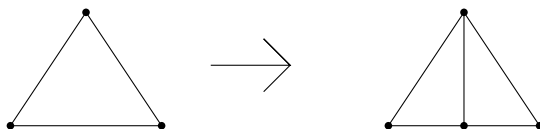
The following fact is crucial:

Proposition 14.5. *Any two triangulations of the same space have the same Euler characteristic.*

We will not give a full proof of this, but we will give some indication of why it is true. Consider the following three triangulations of the unit square:



The proposition says that these should have the same Euler characteristic. The best order to approach this is to show that \mathcal{A} has the same Euler characteristic as \mathcal{C} , and then that \mathcal{B} has the same Euler characteristic as \mathcal{C} . The point here is that the triangles in \mathcal{C} can be obtained by cutting up the triangles in \mathcal{A} , and they can also be obtained by cutting up the triangles in \mathcal{B} . One can show that this kind of approach is always possible: given two linear triangulations \mathcal{A} and \mathcal{B} for the same subset $X \subseteq \mathbb{R}^N$, there is always a third triangulation \mathcal{C} which can be obtained by subdividing \mathcal{A} , and can also be obtained by subdividing \mathcal{B} . If we believe this, then the problem is just to show that subdivision does not change the Euler characteristic. One basic kind of subdivision is like this:



In the right hand picture:

- There is one extra vertex, so the number p of vertices increases by one.
- The horizontal edge has been split in two, and there is a new vertical edge. This means that the number q of edges increases by two.
- The triangle has been split in two, so the number r of triangles increases by one.

It is clear from this that the combination $\chi = p - q + r$ does not change at all. There are various other kinds of subdivision, but they can all be analysed in a similar way, and we conclude that a single step of subdivision does not change the Euler characteristic. It follows inductively that no finite process of subdivision can change the Euler characteristic; so any two different triangulations will have the same characteristic, but the argument outlined above.

Here is another important fact along the same lines as Proposition 14.5.

Theorem 14.6. *If there is a homeomorphism $f: X \rightarrow Y$, then $\chi(X) = \chi(Y)$.*

If the homeomorphism f is itself piecewise-linear, then we can apply f to a sufficiently fine linear triangulation of X to get a linear triangulation of Y , and so the theorem will follow from Proposition 14.5 in this case. However, a more complex argument is needed to cover the case where f is not piecewise-linear; we will not discuss this further.

Example 14.7. (a) If X is homeomorphic to the disc D , then $\chi(X) = 1$. Indeed, we saw in Example 14.2 that $\chi = 1$ for one triangulation of the square (which is homeomorphic to D) and the general claim follows by Proposition 14.5 and Theorem 14.6.
 (b) If X is homeomorphic to the sphere S^2 , then $\chi(X) = 2$. This follows from Example 14.3, by the same argument as for (a).
 (c) If X is homeomorphic to a circle, then $\chi(X) = 0$. Indeed, it will be enough to prove this when X is the boundary of an n -gon, in which case we have $(p, q, r) = (n, n, 0)$ and so $\chi = n - n + 0 = 0$ as claimed.

The following result often makes it easy to calculate Euler characteristics.

Proposition 14.8. *Let X be a surface, and suppose that $X = X_0 \cup X_1$, where X_i is the union of some subset of the triangles in X . Then*

$$\chi(X) = \chi(X_0) + \chi(X_1) - \chi(X_0 \cap X_1).$$

Proof. Let p_i, q_i and r_i denote the numbers of vertices, edges and faces in X_i . Similarly, let p, q and r denote the numbers in $X = X_0 \cup X_1$, and let p', q' and r' denote the numbers in $X_0 \cap X_1$. Now $p_0 + p_1$ counts the total number of vertices in X , except that vertices in $X_0 \cap X_1$ get counted twice, so we must subtract p' to compensate for that; this gives $p = p_0 + p_1 - p'$. Similarly, we have $q = q_0 + q_1 - q'$ and $r = r_0 + r_1 - r'$. It follows that

$$\begin{aligned}\chi(X) &= p - q + r = (p_0 + p_1 - p') - (q_0 + q_1 - q') + (r_0 + r_1 - r') \\ &= (p_0 - q_0 + r_0) + (p_1 - q_1 + r_1) - (p' - q' + r') \\ &= \chi(X_0) + \chi(X_1) - \chi(X_0 \cap X_1).\end{aligned}$$

□

Lemma 14.9. *If M^* is obtained from M by removing an open disc, then $\chi(M^*) = \chi(M) - 1$.*

Proof. We can write M as the union of M^* with a disc D , where $M^* \cap D$ is a circle S^1 . This gives

$$\chi(M) = \chi(M^*) + \chi(D) - \chi(S^1) = \chi(M^*) + 1 - 0 = \chi(M^*) + 1.$$

□

Example 14.10. The cylinder and the torus both have $\chi = 0$. Indeed, if we remove two discs from a sphere S^2 then the result is homeomorphic to a cylinder C , so $\chi(C) = \chi(S^2) - 2 = 2 - 2 = 0$. Next, the torus T can be written as the union of two bent cylinders C_0 and C_1 , where $C_0 \cap C_1$ consists of two circles S_0 and S_1 . This means that $\chi(C_0) = \chi(C_1) = \chi(C_0 \cap C_1) = 0$, so the formula $\chi(T) = \chi(C_0) + \chi(C_1) - \chi(C_0 \cap C_1)$ gives $\chi(T) = 0$ as well.

Corollary 14.11. *If M and N are surfaces, then $\chi(M \# N) = \chi(M) + \chi(N) - 2$.*

Proof. $M \# N$ is the union of M^* and N^* , glued along a circle S^1 . This gives

$$\chi(M \# N) = \chi(M^*) + \chi(N^*) - \chi(S^1) = (\chi(M) - 1) + (\chi(N) - 1) - 0 = \chi(M) + \chi(N) - 2.$$

□

It is convenient to rewrite the above relation in terms of a slightly different number:

Definition 14.12. The *genus* of M is $g(M) = 1 - \chi(M)/2$ (so $\chi(M) = 2 - 2g(M)$).

(This definition is completely standard for orientable closed surfaces X , but not so standard for nonorientable surfaces or surfaces with boundary.)

Proposition 14.13. *We have $g(S^2) = 0$ and $g(T) = 1$ and $g(M \# N) = g(M) + g(N)$.*

Proof. The results for $g(S^2)$ and $g(T)$ are immediate from our calculation of the Euler characteristics. We also have $\chi(M \# N) = \chi(M) + \chi(N) - 2$, so

$$\begin{aligned}g(M \# N) &= 1 - (\chi(M) + \chi(N) - 2)/2 = 1 - (2 - 2g(M) + 2 - 2g(N) - 2)/2 \\ &= g(M) + g(N)\end{aligned}$$

as claimed. □

Lemma 14.14. *The projective plane $\mathbb{R}P^2$ has Euler characteristic 1.*

It will be convenient to postpone the proof of this until Example 14.22.

Corollary 14.15. *If the surface word W is equivalent to $T^{\#i} \# P^{\#j} \# D^{\#k}$ then $\Sigma(W)$ has genus $i + (j + k)/2$, and Euler characteristic $2 - 2i - j - k$.*

Proof. As the Σ construction sends connected sums to connected sums, we see that $g(\Sigma(W)) = i g(T) + j g(P) + k g(D) = i + (j + k)/2$. The relation $\chi = 2 - 2g$ therefore gives $\chi(\Sigma(W)) = 2 - 2i - j - k$. □

We need one more numerical invariant:

Definition 14.16. For any surface M , we write $b(M)$ for the number of boundary circles in M .

If M is homeomorphic to N , then $b(M) = b(N)$. This is harder to prove rigorously than you might think, but it should be clear geometrically.

Theorem 14.17. *None of the standard surfaces mentioned in Theorem 13.23 are homeomorphic to each other. In more detail:*

- (a) *If $T^{\#i} \# D^{\#j} \simeq T^{\#k} \# D^{\#l}$, then $i = k$ and $j = l$.*
- (b) *If $P^{\#i} \# D^{\#j} \simeq P^{\#k} \# D^{\#l}$, then $i = k$ and $j = l$.*
- (c) *$T^{\#i} \# D^{\#j}$ is never homeomorphic to $P^{\#k} \# D^{\#l}$.*

Proof. It is easy to see that

$$b(T^{\#i} \# D^{\#j}) = b(P^{\#i} \# D^{\#j}) = j.$$

Thus, if $T^{\#i} \# D^{\#j} \simeq T^{\#k} \# D^{\#l}$ then we can use b to deduce that $j = l$. We can then use the Euler characteristic to see that $2 - 2i - j = 2 - 2k - l$, and it follows that $i = k$. The same argument works for (b). It should be reasonably clear from the definitions that an orientable surface cannot be homeomorphic to a nonorientable one. \square

Now suppose we start with a surface word W , and we want to understand the homeomorphism type of $\Sigma(W)$. For simplicity, we will only consider the case where W is a closed surface word. It is easy to see whether W is orientable or not, so the only problem is to determine the Euler characteristic. We could do this by using word moves to convert W to standard form, but it is also possible to use a more direct method, which we will explain in a slightly more general setting.

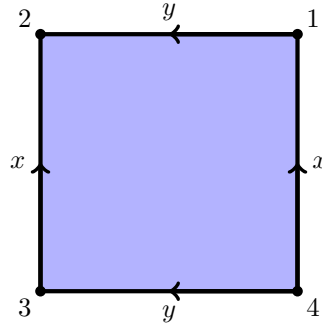
Definition 14.18. Let M be a surface. A *covering pattern* for M is a finite list of subsets $P_k \subseteq M$, together with homeomorphisms which identify each P_k as a regular n -gon with some edges identified in pairs, such that:

- (a) The sets P_k cover all of M .
- (b) For all $i \neq j$, the intersection $P_i \cap P_j$ consists of a finite number of vertices and complete edges.

The sets P_k will be called *polygonal cells*.

Proposition 14.19. *Suppose that M has a covering pattern with p vertices, q edges and r cells. Then $\chi(M) = p - q + r$.*

Remark 14.20. Here the edges and cells are supposed to be counted *after* identifying together the edges that are supposed to be identified. For example, consider the usual picture for the torus T :



This is a covering pattern with only one cell.

The original square has four vertices and four edges. After gluing the two edges marked x become the same, and the two edges marked y become the same, so we end up with only two edges. When we glue the two x s, vertex 1 becomes the same as 2, and vertex 3 becomes the same as 4. However, when we glue the two y s, vertex 1 becomes the same as 4, and 2 becomes the same as 3. This means that all four vertices become the same in the torus. We thus have $p = 1$ and $q = 2$ and $r = 1$, giving $\chi = 1 - 2 + 1 = 0$. This agrees with the answer we obtained earlier in Example 14.10.

Proof of Proposition 14.19. Let P_1, \dots, P_r be the polygonal cells in our covering pattern. Let n_i be the number of sides of P_i , and put $n = \sum_i n_i$. We could try to construct a triangulation by drawing extra edges from each of the vertices of P_i to the centre, which divides P_i into n_i triangles. This adds r new vertices

(the centres of the polygons P_i), so the new number of vertices is $p' = p + r$. Similarly, it adds n_i new edges to P_i , so $q' = q + \sum_i n_i = q + n$. The total number of triangles is $r' = n$. From this we get

$$\chi = p' - q' + r' = (p + r) - (q + n) + n = p - q + r,$$

as required.

However the above argument is only complete if the specified subdivision procedure actually gives a triangulation. This is not always the case, as we discussed in the proof of Lemma 12.13: in a proper triangulation the two endpoints of an edge are always distinct, and this need not hold for the attempted triangulation discussed above. However, the finer subdivision pattern used in the proof of Lemma 12.13 will always give a proper triangulation. One can check that the numbers of vertices, edges and triangles in this triangulation are

$$p' = p + 3q + r + 3n$$

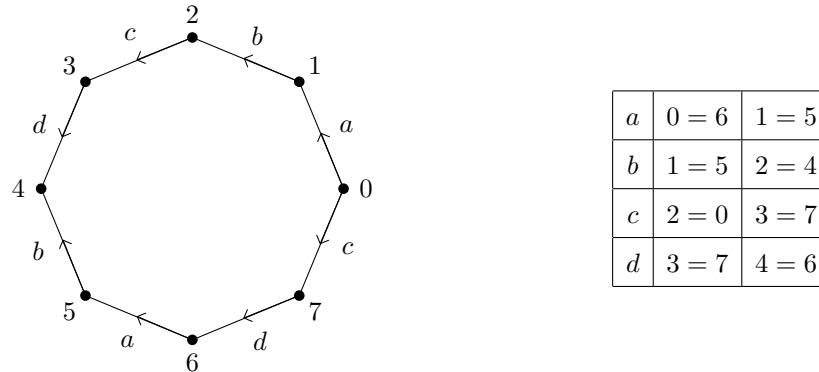
$$q' = 4q + 13n$$

$$r' = 10n.$$

This gives $\chi = p' - q' + r' = p - q + r$, as required. \square

Now consider a closed surface word W . The letters occur in matched pairs, so the length must be even, say $2q$. By construction, $\Sigma(W)$ has a covering pattern with one polygonal cell. After gluing, the number of edges is q . The number p of vertices can be found by the method explained in Example 12.15; we will give further examples below. The Euler characteristic is then $p - q + 1$.

Example 14.21. Consider the word $W = abcd\bar{b}\bar{a}\bar{d}\bar{c}$. This can be drawn as shown on the left below.



We need to analyse the vertices in $\Sigma(W)$. One of the edges marked a starts at vertex 0, and the other starts at vertex 6, so vertices 0 and 6 become the same in $\Sigma(W)$. Similarly, by considering the endpoints of the two edges marked a , we see that vertices 1 and 5 become the same. Repeating this for the other edges gives the table shown on the right above. Putting this together, we get $0 = 2 = 4 = 6$ and $1 = 5$ and $3 = 7$, and there are no further identifications, so there are three vertices in $\Sigma(W)$, and four edges. This gives $\chi(\Sigma(W)) = 3 - 4 + 1 = 0$, so the genus is $g = 1 - \chi/2 = 1$, so $\Sigma(W)$ must be homeomorphic to a torus. We can also do this by word moves, as follows:

$$W = a(bcd|\bar{b})\bar{a}\bar{d}\bar{c} \sim a(\bar{b}b)cd\bar{a}\bar{d}\bar{c} \sim ac(d\bar{a}|\bar{d})\bar{c} \sim ac(\bar{d}d)\bar{a}\bar{c} \sim ac\bar{a}\bar{c} \cong xy\bar{x}\bar{y} = T.$$

Example 14.22. We can construct $\mathbb{R}P^2$ as $\Sigma(xx)$. This involves a disc with two vertices that are identified together, and two edges that are identified together. Thus, we have a covering pattern with $(p, q, r) = (1, 1, 1)$ and therefore $\chi = 1 - 1 + 1 = 1$. This proves Lemma 14.14.