Knots and Links

H.R.Morton

Spring 2002

1 Introduction

We are all able to tie a knot in a piece of rope. What exactly do we mean though when we say that a piece of rope is knotted?

Look at the pictures in figure 1.

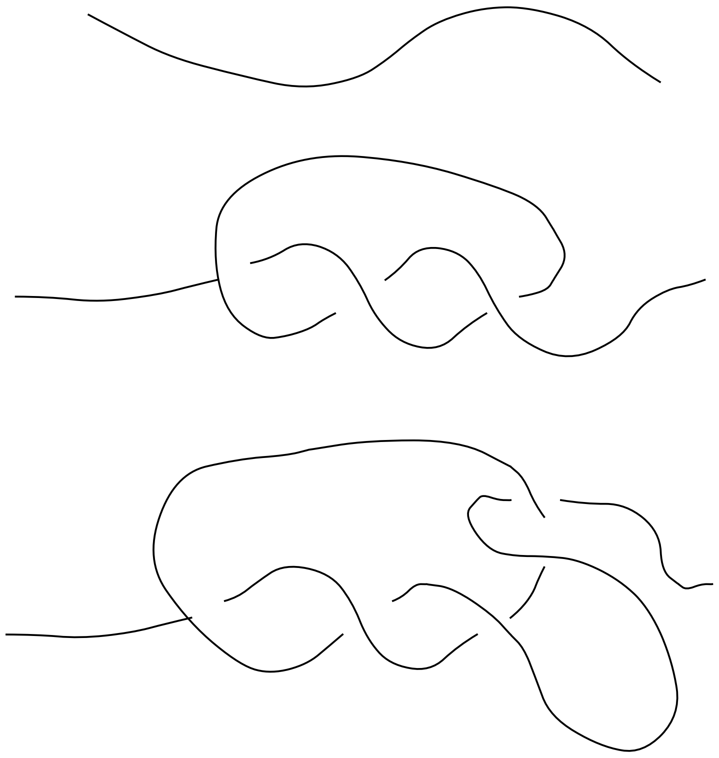


Figure 1:

Let us first stop the knots escaping by joining up the ends of the rope, as in figure 2. Compare what happens in the three cases.

In the first case we get a simple, or 'unknotted' circle, while in the second case we have a circle with what appears to be a knot in it.

Let us say that the rope is *knotted* if no possible manipulation of it will result in the unknotted circle. We do not allow cutting and rejoining.

The third example can clearly be undone by a little manipulation to form the simple circle, so again the rope is unknotted.

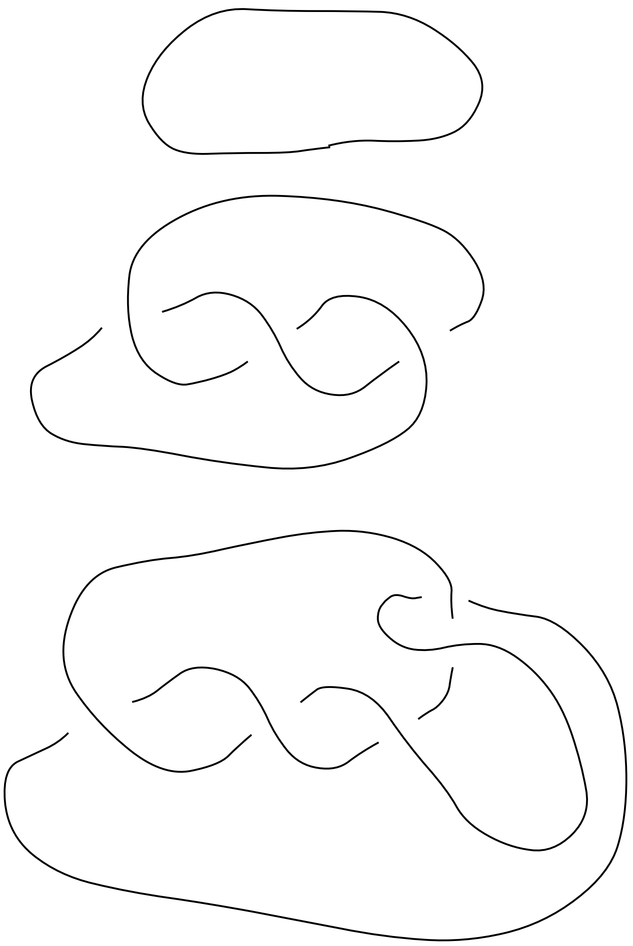


Figure 2:

We model this notion of a knot mathematically by referring to a closed curve in $\mathbf{R}^3$ as a knot, with the special case of the simple circle, lying say as the unit circle in a plane, known as the *trivial knot* or *unknot*. Knot theory in the mathematical sense is then the study of closed curves in space.

We call two knots *equivalent* if one can be manipulated, without passing one strand through another, to become the other knot. I give a more formal technical description of this below, but essentially anything is allowed which could be done with a rather stretchable piece of rope. The one manoeuvre which must be excluded is the analogue of the bachelor's technique for ignoring knots on a piece of cotton – pull it so tight that you can hardly see it! Using this technique on a curve with no physical thickness would get rid of any knot.

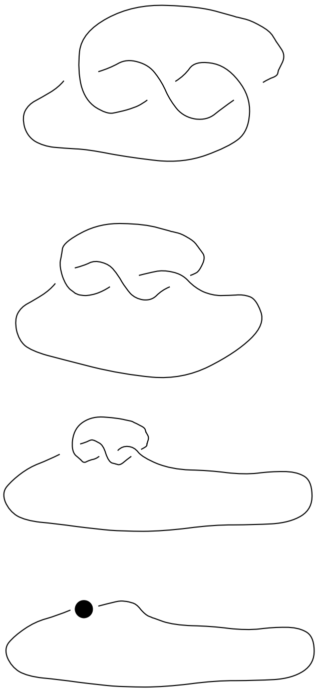


Figure 3: Bachelor's unknotting

We would like to know for a start if there are any knots which are not equivalent to the trivial knot. If so, are there lots of different knots, and how might we distinguish between them? It is easy to imagine that you have been given two knots and by a little patient work you manage to manipulate one to look like the other, e.g. the first and third knots in figure 2. What happens though if you find that even after a lot of trying you can't make them look the same – does it follow that the knots are inequivalent, or have you just not been dextrous enough? There is clearly a problem here, and something else will be needed, as there is no way that failure to manipulate can show that it is actually impossible to do so.

It should be realised that the question of how the rope is knotted isn't an intrinsic question about the rope alone, but rather a matter of how the rope is placed in space. Every closed loop of rope looks the same to an ant inside the rope. Some of the techniques developed for the study of knots have proved fruitful in other 'placement problems', i.e. in studying the different ways in which one particular geometric object, here a closed curve, may lie inside a larger one.

Background. The idea of looking at knotted and unknotted closed curves goes back to Gauss and beyond. Kelvin had some idea of trying to relate different types of atoms to knotted curves in the ether; this was taken up by a Scottish physicist Tait, who set out to enumerate all possible different knots in the hope of tallying them against different atoms. His lists of knots soon showed that the task of systematically enumerating all knots was hopelessly complicated; among other problems there are infinitely many. It is still true today that no practical framework exists for producing a comprehensive list, although Thistlethwaite has devised a fairly good means of handling the simpler knots. Various mathematicians in the 1920s and 1930s developed methods to show up a number of general properties shared by all knots, using some very elegant geometrical techniques and exploiting the growing interplay between algebra and this style of geometry. From this period has come the Alexander polynomial, and interpretations of it, as well as group theoretic invariants. Much more recently knot theory and theoretical physics have again had close contacts.

Definition. A *knot* is a simple closed curve $K \subset \mathbf{R}^3$ or in S^3 (more about this later).

Definition. The *complement* of K is $S^3 - K$.

We shall only deal with *tame* knots, e.g. smooth or polygonal curves, and we assume that K has a solid torus neighbourhood V with

$$(V, K) \cong (S^1 \times D^2, S^1 \times \{0\}).$$

This is like insisting on using a piece of rope, although its exact thickness will not matter.

It is often convenient to deal with $S^3 - \text{int}V = \text{ext}K$, the *exterior* of K, which is a *compact* 3-manifold with boundary $\partial(\text{ext}K) = \partial V \cong$ torus $S^1 \times S^1$.

From the point of view of topological invariants there is not much difference between $S^3 - K$, $\text{ext}K$ and $S^3 - V$.

Definition. Knots K_0 and K_1 are *homeomorphic* if there exists a homeomorphism $h : \mathbf{R}^3 \to \mathbf{R}^3$ such that $h(K_0) = K_1$.

Remark. A *homeomorphism* from A to B is a continuous bijective map from A to B whose inverse is also continuous.

A homeomorphism h from $\mathbf{R}^3$ to itself is either orientation preserving or orientation reversing. If it is orientation reversing then its composite with a reflection will be orientation preserving. Every orientation preserving homeomorphism of $\mathbf{R}^3$ is known to be *isotopic* to the identity, i.e. there exists a 1-parameter family $h_t, 0 \leq t \leq 1$, of homeomorphisms with $h_0 =$ identity and $h_1 = h$. Then if h is orientation preserving we can deform K_0 to K_1 through a family of knots $K_t = h_t(K_0)$. We shall call K_0 and K_1 *equivalent* when they are related in this way. (The term *ambient isotopic* is also used.)

Conversely a 1-parameter sliding of a neighbourhood V of K_0 to one of K_1 through $\mathbf{R}^3$ can be extended to such a family h_t of homeomorphisms, and models quite well the physical notion of equivalence by manipulation of a closed loop of rope.

We then have the result, by composing with a reflection if necessary, that two knots K_0 and K_1 are homeomorphic if and only if K_0 is equivalent to K_1 or its mirror-image.

Remark. Some knots, for example the trefoil, are not equivalent to their mirror image, while others such as the figure-eight knot are.

Many questions about equivalence of knots can be answered theoretically by looking at the fundamental group $G_K = \pi_1(\mathbf{R}^3 - K)$, the *group of* K, which is well-defined up to isomorphism.

It is true that

$$\begin{aligned} K_0, K_1 \text{ equivalent} \quad &\Rightarrow \quad \mathbf{R}^3 - K_0 \cong \mathbf{R}^3 - K_1 \\ &\Rightarrow \quad G_{K_0} \cong G_{K_1}. \end{aligned}$$

So knots are different (inequivalent, indeed not homeomorphic) if their groups are not isomorphic.

The group G_K is also isomorphic to $\pi_1(\text{ext}K)$; knowledge of this group *and* the subgroup coming from the torus $\partial(\text{ext}K) = \partial V$ is actually enough to theoretically determine K.

In later sections we shall discuss the fundamental group, and give explicit presentations of a knot group, starting from a diagram of the knot. It is not however easy in practice to decide when two non-abelian groups are isomorphic, so more readily compared invariants are sought to try to establish differences between given knots; these may be particularly effective when the knots are known to have certain geometric properties, such as lying on some specified closed surface in S^3, or forming the boundary of each of a family of surfaces. The geometric information either available as data or wanted as a property of the knot may not be accessible readily from its group – the behaviour of the knot and its relation to others may equally be more clearly seen from some of its geometric rather than algebraic invariants. There is a long history of interplay, usually with a grey area of indecisiveness, between the various algebraic and geometric threads, and it has been the richness of examples, coupled with the elusive nature of the full picture in spite of immediate calculations being available in specific cases which has maintained interest over many years.

1.1 Knot diagrams and moves

For our subsequent analysis it is essential that we concentrate on tame knots, i.e. knots equivalent to finite polygonal curves or equally to regular smooth curves. It can be shown that if two polygonal curves are equivalent then one can be moved to the other through a family of *polygonal* curves, and further that the move can be made up of a finite sequence of moves in which one vertex is moved on a straight line and the rest are left alone. We are also able to view a polygonal or smooth curve by means of a knot diagram, which is a projection from some direction to a plane in which the image has only a finite number of simple crossings. Only a small subset of directions must be excluded (a set of dimension 1 in the 2-dimensional set of directions) in finding a diagram. At each crossing point the two branches are distinguished into over and under crossings. The effects of a polygonal move on a curve are then visible on a diagram as either leaving it essentially unchanged, or altering it by one of the three Reidemeister moves, which are shown in figure 4.

Theorem 1.1 (Reidemeister) *If two diagrams represent equivalent knots then one diagram can be converted to the other by a finite sequence of Reidemeister moves, along with isotopy (deformation) of the image within the projection plane.*

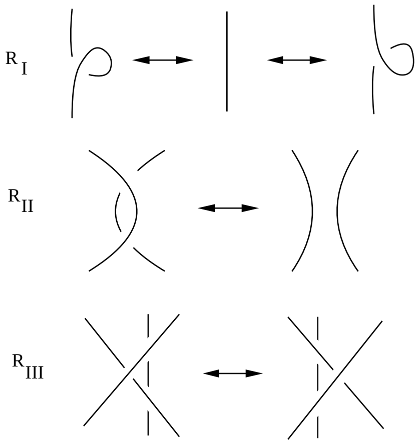


Figure 4: Reidemeister's moves

1.2 Links and linking number

We may enlarge our scope slightly and look, as Gauss did, not just at a single closed curve but at several at once.

Definition. A *link* of r components is a collection $L = L_1 \cup L_2 \cup \ldots \cup L_r$ of r closed non-intersecting curves.

When $r = 1$ we have a knot. In the case $r = 2$ we can very simply associate an integer with a link, which is the same for every equivalent link. This is called the *linking number* of the two components.

To define the linking number $\mathrm{lk}(L_1, L_2)$ we must first choose an orientation of each of the components, which we note on a diagram of the link by drawing arrows on the curves. Now look at one diagram of the link and consider only the crossings where L_1 crosses over L_2. Each of these crossings c_i can be given a sign $\varepsilon_i = \pm 1$, according to a conventional choice. The sum of these signs $\sum \varepsilon_i$ is unaltered when the diagram is changed by Reidemeister moves. For crossings of L_1 over L_2 are not affected by moves I and III, while if there are any involved in a move of type II they occur as a pair with opposite sign, so that the sum is unchanged.

Reidemeister's theorem holds also for links. We may then set $\mathrm{lk}(L_1, L_2) = \sum \varepsilon_i$ for any choice of diagram.

Proposition 1.2 *$lk(L_2, L_1) = lk(L_1, L_2)$.*

Proof : To calculate $\mathrm{lk}(L_2, L_1)$ we must count the crossings of L_2 over L_1 in some diagram. Start with a diagram in which we count the crossings c_i of L_1 over L_2. If we turn this diagram over and view it from the other side we get a new diagram of the link in which the crossings c_i become the crossings of L_2 over L_1. Each crossing, viewed from the other side has the same sign as it had initially, so the sum needed to calculate $\mathrm{lk}(L_2, L_1)$ from this diagram is identical to the sum calculating $\mathrm{lk}(L_1, L_2)$ in the original diagram. □

2 S^3 and $\mathbf{R}^3$.

We start by studying some explicit views of the unknot and some other simple links. We shall use stereographic projection which maps S^3 with a point removed to $\mathbf{R}^3$ to convert between links in $\mathbf{R}^3$ and links in S^3.

We first describe the stereographic projection homeomorphism $h : S^3 - N \to \mathbf{R}^3$, where we take

$$S^3 = \{(\mathbf{x}, t) \in \mathbf{R}^3 \times \mathbf{R}; |\mathbf{x}|^2 + t^2 = 1\}$$

and $N = (\mathbf{0}, 1)$, the 'North Pole'.

Definition. The map $h : S^3 - N \to \mathbf{R}^3$ defined by

$$h(\mathbf{x}, t) = \frac{1}{1-t}\mathbf{x}$$

is a homeomorphism, called *stereographic projection.*

Its inverse g is given by $g(\mathbf{X}) = (\lambda \mathbf{X}, \mu)$ where $\lambda = 1 - \mu$ and $\mu = \dfrac{|\mathbf{X}|^2 - 1}{|\mathbf{X}|^2 + 1}$.

Remark. Note that there is an orthogonal transformation of $\mathbf{R}^4$ carrying any chosen point on S^3 to N, so that $S^3 - \text{point} \cong S^3 - N$, for any other choice of point.

When we compare curves in $\mathbf{R}^3$ and curves in S^3 using h we may note that a homeomorphism τ say of $\mathbf{R}^3$ defines a homeomorphism $h\tau h^{-1}$ of $S^3 - N$ which can be extended to a homeomorphism of S^3 by mapping N to itself.

Given a curve $K \subset S^3$ not through N it is possible to find a homeomorphism of S^3 which fixes K and carries any chosen point of $S^3 - K$ to N. Hence if there is a homeomorphism of S^3 carrying one curve in S^3 not through N to another such curve then we can assume that the homeomorphism fixes N and consequently determines a homeomorphism of their images in $\mathbf{R}^3$ when projected from N.

It follows that equivalence of knots in S^3 or in $\mathbf{R}^3$ amount to essentially the same thing.

Definition. A *great circle* in S^3 is the intersection of S^3 with a 2-dimensional linear subspace W of $\mathbf{R}^4$.

Great circles through N map to straight lines under h, while other great circles map to circles in planes through the origin of $\mathbf{R}^3$. Each great circle meets the equatorial sphere $t = 0, |\mathbf{x}|^2 = 1$ in a pair of antipodal points; its image under h also passes through these two points in $\mathbf{R}^3$.

Two simple examples are the great circles $C_1 = \{(\mathbf{x}, t); x_1 = x_2 = 0\}$ and $C_2 = \{(\mathbf{x}, t); x_3 = t = 0\}$. Here $h(C_1 - N) = x_3$-axis while $h(C_2)$ is the unit circle $\{\mathbf{x}; x_1^2 + x_2^2 = 1, x_3 = 0\}$ in the plane $x_3 = 0$.

We now consider the complement of C_1 and of the link $C_1 \cup C_2$. The restriction map $h|S^3 - C_1 \to \mathbf{R}^3 - x_3$-axis is a homeomorphism. We can follow this with a further homeomorphism

$$k : \mathbf{R}^3 - x_3\text{-axis} \to S^1 \times P$$

to the product of S^1 with an open half-plane $P = \{(r, z); r > 0\}$, defined by

$$k(x_1, x_2, x_3) = \left(\frac{1}{\sqrt{x_1^2 + x_2^2}}(x_1, x_2), (\sqrt{x_1^2 + x_2^2}, x_3) \right).$$

It may be helpful to think of k as the cylindrical polar map taking (x, y, z) to the pair $e^{i\theta} \in S^1$, $(r, z) \in P$, where r, θ, z are the usual cylindrical polar coordinates based on the z-axis.

By composing these two homeomorphisms we then have a homeomorphism $S^3 - C_1 \cong S^1 \times P$. Restricted to C_2 this map carries C_2 first to $h(C_2) =$ unit circle in $x_3 = 0$ and then to $k(\text{unit circle}) = S^1 \times \{(1, 0)\}$. Restriction of the composite of the homeomorphisms to the complement of $C_1 \cup C_2$ then gives

$$S^3 - (C_1 \cup C_2) \cong \mathbf{R}^3 - (x_3\text{-axis} \cup \text{unit circle}) \cong S^1 \times (P - \{(1, 0)\}).$$

Any two non-intersecting great circles in S^3 form a link in S^3 which can be carried homeomorphically to $C_1 \cup C_2$ using a linear isomorphism of $\mathbf{R}^4$ followed by radial projection to the unit sphere S^3. The image under stereographic projection from a point not on the link will be a link of two circles in $\mathbf{R}^3$. We could take, for example, the great circles C_2 and $\{(\mathbf{x}, t); x_1 = at, x_2 = 0, a \neq 0\}$ which project stereographically from N to form two circles, K_1 in the plane $x_2 = 0$ and the unit circle K_2 in $x_3 = 0$, as shown in figure 5. We call this or any equivalent link the *Hopf link.*

Aside. We may write $\mathbf{R}^4$ as $\mathbf{C}^2$, taking points $(z_1, z_2) \in \mathbf{C}^2$ with $z_1 = x_1 + ix_2, z_2 = x_3 + it$ to correspond with our earlier choice of coordinates. Then S^3 is given by $|z_1|^2 + |z_2|^2 = 1$ and the circles C_1 and C_2 in S^3 are given respectively by $z_2 = 0$ and $z_1 = 0$. Their union satisfies the equation $z_1 z_2 = 0$.

A very elegant theory elaborated by Milnor shows that if $f : \mathbf{C}^2 \to \mathbf{C}$ is a polynomial with $f(0,0) = 0$ then all small enough spheres S_ε centre $(0,0)$ meet $f^{-1}(0)$ in a curve or curves forming a link, with the link being independent of ε for sufficiently small ε. Milnor shows further that on the complement $S^3_\varepsilon - f^{-1}(0)$ the map $p : S^3_\varepsilon - f^{-1}(0) \to S^1$ defined by $p(\mathbf{z}) = f(\mathbf{z})/|f(\mathbf{z})| \in S^1$ is a fibration. This means that the fibres $F_\theta = p^{-1}(e^{i\theta}) = \{\mathbf{z}; \arg f(\mathbf{z}) = \theta\}$ are all homeomorphic; in fact they are all surfaces, and they fit nicely round the curve(s) $f^{-1}(0)$ rather like leaves round the spine of a book.

We can see something of this in the previous examples.

Example. Take $f(z_1, z_2) = z_1$. Here the result will look identical for all choices of radius ε and we may as well use the unit sphere. The link in question is just the curve $C_2 = f^{-1}(0)$. Then p is defined on the complement of C_2 by $p(z_1, z_2) = z_1/|z_1|$ and the surface F_θ with $\arg z_1 = \theta$ maps stereographically into the half-plane with cylindrical polar coordinate θ in $\mathbf{R}^3$. The surfaces F_θ then consist of open discs fitting around C_2.

Another view of the same family is given by looking at C_1 instead, taking $f(\mathbf{z}) = z_2$. The surfaces F_θ then satisfy $\arg z_2 = \theta$ and so form part of the great sphere in S^3 with equation $t = kx_3$ where $k = \tan\theta$. This projects to the sphere $x_1^2 + x_2^2 + x_3^2 = 1 + 2kx_3$ through the unit circle in $\mathbf{R}^3$ and the surfaces F_θ give a family of discs spanning the trivial knot.

Example. We may also look at the case $f(z_1, z_2) = z_1 z_2$ (or equally $z_1^2 - z_2^2$). Here $f^{-1}(0) = C_1 \cup C_2$, the Hopf link. To get a good view of the sets F_θ in this case it is helpful to consider the function defined on the complement of C_1 by $g(z_1, z_2) = z_1/z_2$. The inverse image $g^{-1}(\lambda)$ consists for each $\lambda \in \mathbf{C}$ of the great circle $z_1 = \lambda z_2$ and the complement of $C_1 \cup C_2$ is filled up by these circles

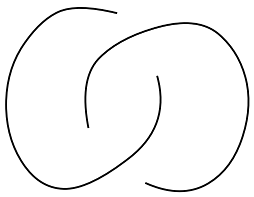


Figure 5: The Hopf link

with $\lambda \neq 0, \infty$. Each pair of such circles will form a Hopf link, as we showed above. The subset G_θ defined as for polynomials to consist of the points where $g/|g| = e^{i\theta}$ is then made up of those circles with $\arg\lambda = \theta$. Under stereographic projection the image of each G_θ is given by rotating the image of G_0 around the x_3-axis through the angle θ. Now G_0 consists of the circles where $\lambda \in \mathbf{R}, \lambda > 0$, given by $x_1 = \lambda x_3, x_2 = \lambda t$. The image in $\mathbf{R}^3$ forms the intersection of the plane $x_1 = \lambda x_3$ with the sphere $x_1^2 + x_2^2 + x_3^2 = 1 + 2\lambda^{-1}x_2$ and the whole set G_0 makes up an annulus. Moving the projection point away from C_1 allows a bounded view of a surface G_θ, along with the link forming its boundary, as a ribbon with a single full twist.

The surfaces $F_\theta = \{\mathbf{z}; \arg f(\mathbf{z}) = \theta\}$ defined by Milnor's method for $f = z_1z_2$ in place of g are mapped to G_θ by the reflection r of $\mathbf{R}^4$ with $r(z_1, z_2) = (z_1, \overline{z}_2)$, since $g(r(\mathbf{z})) = f(\mathbf{z})/|z_2|^2$ and both will have the same arg.

Generally the knots and links which arise from Milnor's construction as $f^{-1}(0)$ have very nice properties. The next most simple case of this can be given by $f = z_1^3 - z_2^2$. Then $f^{-1}(0)$ in S^3 consists of a curve lying on a torus $|z_1| = \text{const.}, |z_2| = \text{const.}$ in S^3. The torus may be parametrised by the pair $(\arg z_1, \arg z_2)$. The relation $z_1^3 = z_2^2$ on $f^{-1}(0)$ shows that $3\arg z_1 = 2\arg z_2$. When viewed in $\mathbf{R}^3$ the torus is symmetric about the x_3-axis, and the curve lies on it like the boundary of a ribbon with 3/2 twists, forming a trefoil knot.

Knots which arise in this way form a very restricted class, and a complete (infinite) list can be given. The figure-eight knot does not appear on the list, although it has a number of very similar properties to the trefoil.

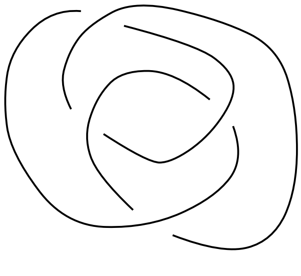


Figure 6: The figure-eight knot

3 The fundamental group

The classic technique of algebraic topology is to use algebraic objects — numbers, polynomials, groups or more complicated structures — as a means of studying geometric or topological features of a problem.

We have already seen the use of linking number in considering 2-component links. We showed that equivalent links have the same linking number, giving us a partial algebraic test for equivalence.

In a similar vein I shall describe a more complicated topological invariant which is defined in a very wide context, namely the fundamental group $\pi_1(X)$ of a topological space X. To say that the fundamental group is a 'topological invariant' means that two homeomorphic spaces must have *isomorphic* fundamental groups. It then gives a potential means for showing that two spaces are not homeomorphic.

We shall apply it in particular to the case of knot complements $\mathbf{R}^3 - K$, where the fundamental group $\pi_1(\mathbf{R}^3 - K)$ is often known as the 'group of the knot K'. Since equivalent knots have homeomorphic complements we may be able to establish that two knots are inequivalent by showing that their groups are not isomorphic.

3.1 Basic ideas

We shall now formulate the ideas and techniques needed to discuss the fundamental group of any subset of $\mathbf{R}^n$; these can be carried over essentially unaltered to any topological space.

First some brief terminology relating to paths in subsets A of $\mathbf{R}^n$.

Definition. For $x_0, x_1 \in X$ a *path in X* from x_0 to x_1 is a continuous map $a : [0,1] = I \to X$ such that $a(0) = x_0$ and $a(1) = x_1$.

Definition. The set X is *path-connected* if for each choice of $x, y \in X$ there is a path in X from x to y.

Example. Given $x, y \in \mathbf{R}^n$, the *straight-line path* from x to y is given by $a(s) = sy + (1-s)x$, $s \in I$.

Let $a, b : I \to X$ be paths in X from x_0 to x_1 and from x_1 to x_2 respectively.

Define a path $\overline{a}$ by

$$\overline{a}(s) = a(1-s).$$

Then $\overline{a}$ is a path in A from x_1 to x_0 called the *reverse* of the path a.

Define also a path $a.b$ from x_0 to x_2 by

$$a.b(s) = \begin{cases} a(2s), & 0 \leq s \leq \frac{1}{2}, \\ b(2s-1), & \frac{1}{2} \leq s \leq 1. \end{cases}$$

This path is called the *composite* of the paths a and b, and consists of tracing out the path a followed immediately by b. Its continuity follows from the 'piecing-together' theorem.

We aim to capture some information about a path-connected set X by looking at the collection of loops in X. A *loop* (based at $x_0 \in X$) is simply a path in X from x_0 back to x_0.

The set of *all* loops in X is inconveniently huge; we will generally get a much more manageable view by regarding two loops as 'the same' if we can deform one to the other within X.

More formally, we use the term 'homotopy' to denote the sort of deformation to be considered (given shortly), and then sort the loops into 'homotopy classes'.

These homotopy classes of loops in X will make up the fundamental group of X (based at x_0).

There are a few technical and logical things to check before everything works out, but the major features are that

- Each element of the fundamental group is represented by a loop in X.
- Homotopic loops represent the same element.
- Composition in the group comes from composing loops.
- The inverse of an element represented by a loop a is represented by the reverse loop $\overline{a}$.
- A continuous map $f : X \to Y$ induces a group homomorphism from $\pi_1(X)$ to $\pi_1(Y)$.
- A homeomorphism induces an *isomorphism* of fundamental groups.
- For many spaces it is easy to give a good algebraic description of the fundamental group.

3.2 Technical details

Here is a brief summary of some of the technical points.

Write $p(X, x_0, x_1)$ for the set of all paths in X from x_0 to x_1.

Definition. Let a and b be paths in $p(X, x_0, x_1)$. Say that a is *homotopic* to b, written $a \simeq b$, if there is a continuous map $F : I \times I \to X$, with

$$F(s, 0) = a(s), \quad F(s, 1) = b(s), \quad \text{for all } s \in I,$$

and

$$F(0, t) = x_0, \quad F(1, t) = x_1 \quad \text{for all } t \in I.$$

The map F is called a *homotopy* from a to b, and s and t are the *path parameter* and *homotopy parameter* respectively. We can think of F as defining a family of intermediate paths a_t in $p(X, x_0, x_1)$ by $a_t(s) = F(s, t)$, with $a = a_0$ and $b = a_1$.

Proposition 3.1 *If X is* **convex** *then any two paths a and b in $p(X, x_0, x_1)$ are homotopic.*

Proof : Use the *straight-line homotopy* from a to b, given by

$$F(s, t) = (1 - t)\, a(s) + t\, b(s).$$

For then $F(s, t) \in X$ when $(s, t) \in I \times I$, and F satisfies the boundary conditions. □

On the other hand, the straight-line homotopy can't always be used in a space such as $\mathbf{R}^2 - \{0\}$.

Notation. Write ℓ_{PQ} for the straight line path from P to Q, so that

$$\ell_{PQ}(s) = sQ + (1 - s)P.$$

We shall mainly be considering *loops* in a path-connected space X, that is paths with the same beginning and end point, x_0 say.

Definition. A path-connected space X is *simply-connected* if every two loops in $p(X, x_0, x_0)$ are homotopic, for all choices of x_0. (We shall see later that it is enough to show this for just *one* choice of x_0).

Example. Any convex set is simply-connected.

Proposition 3.2 *Let $f : X \to Y$ be continuous, and let $a \simeq b \in p(X, x_0, x_1)$. Then the paths $f \circ a$ and $f \circ b$ in Y from $f(x_0)$ to $f(x_1)$ are homotopic.*

Proof : Let $F : I \times I \to X$ be the homotopy from a to b, then $f \circ F : I \times I \to Y$ is the required homotopy from $f \circ a$ to $f \circ b$.
(Just check the boundary conditions). □

Corollary 3.3 *Let $q : I \to I$ be any continuous map with $q(0) = 0$ and $q(1) = 1$, and let b and $a = b \circ q$ be paths in Y (both from $a(0)$ to $a(1)$). Then $a \simeq b$.*

Proof : Take a as the map f, with I in the role of X.

Since I is convex we have $q \simeq 1_I$, as paths in I with the same endpoints, so that $a = b \circ q \simeq b \circ 1_I = b$. □

We can think of a and $a \circ q$ as paths which cover the same ground in X, but do so at different rates, related by q. This may even involve some backtracking, if q is not monotone.

A useful case of this occurs if we have paths $a_1, \ldots, a_n$ with $a_{i+1}(0) = a_i(1)$, which we compose in order. By judicious choice of a comparison map $q : I \to I$ we can show that the resulting path is independent, up to homotopy, of the exact rates of travel of the constituent paths.

Given $a_1, \ldots, a_n$ as above we can define a path $a = a_1.a_2.\cdots.a_n$ by

$$a(s) = a_i(ns - i + 1), \quad (i-1)/n \le s \le i/n, \ i = 1, \ldots, n.$$

Then $a \circ \ell_{(i-1)/n, i/n} = a_i$.

If we choose any other dissection of I as $0 = t_0 < t_1 < \ldots < t_n = 1$ we can define a path b for which $b \circ \ell_{t_{i-1}, t_i} = a_i$, i.e. we travel along a_i while the path parameter s lies in the interval $[t_{i-1}, t_i]$.

Theorem 3.4 *Where paths a and b are defined as above, then $a \simeq b$.*

Proof : Choose q with $q(i/n) = t_i$, to be affine on each interval $[(i-1)/n, i/n]$. Then

$$q \circ \ell_{(i-1)/n, i/n} = \ell_{t_{i-1}, t_i},$$

and the graph of q consists of line segments joining the points $z_i = (i/n, t_i) \in \mathbf{R}^2$ in order.

Now $a = b \circ q$, since

$$a \circ \ell_{(i-1)/n, i/n} = (b \circ q) \circ \ell_{(i-1)/n, i/n}$$

for each i.

Since q is continuous, and $q(0) = 0, q(1) = 1$ it follows that $a \simeq b$. □

Corollary 3.5 $(a_1.\cdots.a_j).(a_{j+1}.\cdots.a_n) \simeq a_1.\cdots.a_n$.

Proof : The left-hand side is given as the composite of the paths $a_1.\cdots.a_n$ taking $t_j = \frac{1}{2}$, with intervals $[t_{i-1}, t_i]$ of length $1/2j$, $i \leq j$ and length $1/2(n-j)$ otherwise. □

Corollary 3.6 $(a_1.a_2).a_3 \simeq a_1.(a_2.a_3)$.

Proof : Both are homotopic to $a_1.a_2.a_3$. □

3.2.1 Homotopy classes of paths

For $a \in p(X, x_0, x_1)$ write $[a]$ for the class of all paths homotopic to a. (Then $b \in [a]$ means $a \simeq b$).

Remark. In fact $[a] = [b] \Leftrightarrow a \simeq b$, since homotopy is an equivalence relation. A complete proof of this fact requires us to show that

1. $a \simeq a$. (Use the homotopy F, with $F(s,t) = a(s)$ for all t.)
2. If $a \simeq b$ then $b \simeq a$. (Given a homotopy F from a to b take $G(s,t) = F(s, 1-t)$ to give a homotopy from b to a.)
3. If $a \simeq b$ and $b \simeq c$ then $a \simeq c$. (Suitable piecing together of the two given homotopies will produce a homotopy from a to c.)

Notation. Write $\pi(X, x_0, x_1)$ for the set of homotopy classes of paths from x_0 to x_1.

Then $\alpha \in \pi(X, x_0, x_1)$ can be written as $\alpha = [a]$, for some path $a : I \to X$, *representing* α. Any other $a' \simeq a$ is equally a representative of α.

Proposition 3.7 *Let $\alpha \in \pi(X, x_0, x_1)$, $\beta \in \pi(X, x_1, x_2)$, be represented as $\alpha = [a]$, $\beta = [b]$. Then we can define $\alpha.\beta$ depending* **only** *on α and β, by $\alpha.\beta = [a.b]$.*

Proof : We must show that $\alpha.\beta$ does **not** depend on the choice of representative paths a and b, i.e. if $\alpha = [a'] = [a]$ and $\beta = [b'] = [b]$ then $[a'.b'] = [a.b]$. This amounts to showing that if $a \simeq a'$ and $b \simeq b'$ then $a.b \simeq a'.b'$.

So let F, G be homotopies from a to a' and from b to b' respectively. Construct the homotopy $H = F.G$ by

$$H(s,t) = \begin{cases} F(2s,t), & 0 \le s \le \frac{1}{2} \\ G(2s-1,t), & \frac{1}{2} \le s \le 1 \end{cases}$$

This map is continuous, by the piecing together theorem, applied to its definition on the two closed sets $[0, \frac{1}{2}] \times I$ and $[\frac{1}{2}, 1] \times I$. Check the boundary conditions to see that the definitions on the intersection of these two sets agree and that H is indeed a homotopy from $a.b$ to $a'.b'$. □

Theorem 3.8 *For $x_0 \in X$ the set $\pi(X, x_0x_0)$ forms a* **group** *under the operation described above (setting $\alpha.\beta = [a.b]$ where $\alpha = [a],\ \beta = [b]$).*

Definition. This group is called the *fundamental group* of X (based at x_0), and is written as $\pi_1(X, x_0)$.

Proof : We must show

1. that the operation is associative,
2. that there is an identity element $e_{x_0} \in \pi_1(X, x_0)$,
3. that there is an inverse α^{-1} for each $\alpha \in \pi_1(X, x_0)$.

1. Let $\alpha, \beta, \gamma \in \pi_1(X, x_0)$, and write $\alpha = [a],\ \beta = [b],\ \gamma = [c]$. Then $(a.b).c \simeq a.(b.c)$. So

   $$[(a.b).c] = [a.(b.c)] = [a].[b.c] = [a].([b].[c]) = \alpha(\beta\gamma).$$

   Now

   $$[(a.b).c] = [a.b].[c] = ([a].[b]).[c] = (\alpha\beta)\gamma,$$

   showing that $(\alpha\beta)\gamma = \alpha(\beta\gamma)$.

   [The same is true for homotopy classes of paths (not necessarily loops) which can be composed in this order].

2. Define e_{x_0} to be $e_{x_0} = [c_{x_0}]$, where c_{x_0} is the *constant path* at x_0, given by $c_{x_0}(s) = x_0,\ 0 \le s \le 1$.

3. For $\alpha = [a]$ we may take $\alpha^{-1} = [\overline{a}]$, where $\overline{a}$ is the *reverse* of a, i.e. $\overline{a}(s) = a(1-s)$.

These give identity and inverse elements in a wider context, as shown by the two following lemmas, which complete the proof of the theorem. □

Lemma 3.9 *Let $a \in p(X, x_0, x_1)$. Then $c_{x_0}.a \simeq a \simeq a.c_{x_1}$.*

Corollary 3.10 *For $\alpha = [a] \in \pi(X, x_0, x_1)$ we have $\alpha = [a] = [c_{x_0}.a] = [c_{x_0}].[a] = e_{x_0}.\alpha$.*
Similarly $\alpha = \alpha.e_{x_1}$.

Lemma 3.11 *Let $a \in p(X, x_0, x_1)$. Then $a.\overline{a} \simeq c_{x_0}$ (and, replacing a by $\overline{a}$, $\overline{a}.a \simeq c_{x_1}$).*

Corollary 3.12 *For $\alpha = [a] \in \pi(X, x_0, x_1)$ write $\beta = [\overline{a}] \in \pi(X, x_1, x_0)$. Then $\alpha\beta = e_{x_0}$ and $\beta\alpha = e_{x_1}$.*
Again, for $\alpha \in \pi_1(X, x_0, x_1)$ we have $\alpha\beta = e_{x_0} = \beta\alpha$. Then β is inverse to α, and we can write $[\overline{a}] = \alpha^{-1}$.

Proof of Lemma 3.9: Take $q : I \to I$ defined by

$$q(s) = \begin{cases} 0, & 0 \le s \le \frac{1}{2}, \\ 2s-1,, & \frac{1}{2} \le s \le 1. \end{cases}$$

Then q is a path in the convex set I from 0 to 1, and is thus homotopic to the uniform path 1_I, so that $a \circ q \simeq a \circ 1_I = a$. We have $a \circ q = c_{x_1}.a$, giving $c_{x_1}.a \simeq a$, for

$$a \circ q(s) = \begin{cases} a(0) = x_0, & 0 \le s \le \frac{1}{2}, \\ a(2s-1), & \frac{1}{2} \le s \le 1, \end{cases} = c_{x_0}.a(s)\,.$$

□

Proof of Lemma 3.11: Take $r : I \to I$ defined by

$$r(s) = \begin{cases} 2s, & 0 \le s \le \frac{1}{2} \\ 2-2s, & \frac{1}{2} \le s \le 1. \end{cases}$$

Then r is a path in I from 0 to 0, and hence $r \simeq c_0$, the constant path at 0. We have that $a.\overline{a} = a \circ r$ and $a \circ r \simeq a \circ c_0 = c_{x_0}$ □

Proposition 3.13 *For $a, b \in p(X, x_0, x_1)$ we have*

$$a \simeq b \quad \Leftrightarrow \quad a.\overline{b} \simeq c_{x_0}.$$

Proof : Write $\alpha = [a]$, $\beta = [b]$. Then $\alpha.\beta^{-1} = e_{x_0} \ \Leftrightarrow \ \alpha = \beta$. □

Corollary 3.14 *If all loops at x_0 are homotopic in X then so are all paths in X from x_0 to x_1 and conversely.*

3.2.2 Change of base point

Theorem 3.15 *For a path-connected space X, $\pi_1(X, x_1) \cong \pi_1(X, x_0)$ for any $x_0, x_1 \in X$, i.e. the fundamental group of X is independent, up to group isomorphism, of the choice of basepoint.*

Proof : Choose a path c in X from x_0 to x_1, and put $\gamma = [c]$. Define a map $\theta_\gamma : \pi_1(X, x_0) \to \pi_1(X, x_1)$ by $\theta_\gamma(\alpha) = \gamma^{-1}\alpha\gamma$.
Then θ_γ is a group isomorphism.

1. θ_γ is a homomorphism, for $\theta_\gamma(\alpha\beta) = \gamma^{-1}\alpha\beta\gamma = \gamma^{-1}\alpha\gamma\gamma^{-1}\beta\gamma = \theta_\gamma(\alpha)\,\theta_\gamma(\beta)$.
2. θ_γ is bijective, since $\varphi(= \theta_{\gamma^{-1}})$ given by $\varphi(\delta) = \gamma\delta\gamma^{-1}$ is its inverse.

□

Note that the isomorphism **does** in general depend on the path c, up to homotopy.

Corollary 3.16 *A space X is simply-connected $\Leftrightarrow$ $\pi_1(X, x_0) = \{e\}$ for any choice of basepoint x_0.*

3.2.3 Homomorphisms

Theorem 3.17 *Given a continuous map $f : X \to Y$ there is a group homomorphism $f_* : \pi_1(X, x_0) \to \pi_1(Y, f(x_0))$ defined for $\alpha = [a]$ by $f_*(\alpha) = [f \circ a]$.*

Proof : The map f_* is well-defined, for if $\alpha = [a] = [a']$ we have $a \simeq a'$, and so $f \circ a \simeq f \circ a'$, giving $[f \circ a] = [f \circ a']$.

Now let $\alpha = [a]$ and $\beta = [b]$ be elements of $\pi_1(X, x_0)$. We have $f \circ (a.b) = (f \circ a).(f \circ b)$, on comparing their values at $s \in I$.

Then $\alpha\beta = [a.b]$ so

$$f_*(\alpha\beta) \quad = [f \circ (a.b)] = [(f \circ a).(f \circ b)] = [f \circ a].[f \circ b] \qquad = f_*(\alpha)\, f_*(\beta).$$

□

Theorem 3.18 *If $f : X \to Y$ and $g : Y \to Z$ are continuous, then $(g \circ f)_* = g_* \circ f_*$.*

Proof : Let $\alpha = [a] \in \pi_1(X, x_0)$. Then $f_*(\alpha) = [f \circ a]$, so

$$g_* \circ f_*(\alpha) = g_*(f_*(\alpha)) = [g \circ (f \circ a)] = [(g \circ f) \circ a] = (g \circ f)_*(\alpha)$$

for each α, giving the result. □

Remark. Take $f = 1_X$, i.e. $f(x) = x$ for all $x \in X$. Then $f \circ a = a$ so that $f_*(\alpha) = \alpha$ for all $\alpha \in \pi_1(X, x_0)$.

In other words, $f_* = 1_{\pi_1(X)}$.

Corollary 3.19 *If $f : X \to Y$ is a homeomorphism then f_* is a group isomorphism (with inverse g_* where g is the inverse homeomorphism to f).*

Proof : $f \circ g = 1_Y$, so $f_* \circ g_* = (f \circ g)_* = (1_Y)_* = 1_{\pi_1 Y}$, while similarly $g_* \circ f_* = 1_{\pi_1(X)}$. □

3.3 Some examples

Convex sets, for example the whole of $\mathbf{R}^2$, $\mathbf{R}^3$, any half-plane or interval, are all simply-connected, (their fundamental group is trivial).

Probably the simplest non-trivial example is the plane with a point removed. It is not too difficult to show that the unit circle $S^1 \subset \mathbf{R}^2$ and $\mathbf{R}^2 -$ origin have isomorphic fundamental groups. (Prove this!)

What can be established with a bit more work is that this group is infinite cyclic. This means that it is isomorphic to C_∞ (powers t^k of some non-trivial t) in multiplicative notation, or equally $\mathbf{Z}$ in additive notation. Because fundamental groups do not have to be abelian I shall generally write them multiplicatively. In this case the generator t can be represented by the simple loop a around the unit circle with $a(s) = (\cos 2\pi s, \sin 2\pi s)$.

The proof relies on making a well-defined count of the 'winding number' or 'degree' of a loop ℓ around the origin in $\mathbf{R}^2$, and showing that

1. homotopic loops have the same degree,

2. loops with the same degree are homotopic.

It can be shown that the loop a above has degree 1, and that degree adds under composition. A loop of degree k then represents t^k.

There is in fact a nice way to count the degree of a loop ℓ which crosses a radial line from the origin a finite number of times. Count the number, k_+, of anti-clockwise crossings, and the number, k_-, of clockwise crossings: the degree of ℓ is then $k = k_+ - k_-$.

We can now combine this result with our previous study of the complements of the trivial knot and the Hopf link to calculate their groups once we know the simple description of the fundamental group of a product.

Theorem 3.20 *For two spaces A and B the fundamental group of their product $A \times B$ is given by*

$$\pi_1(A \times B) \cong \pi_1(A) \times \pi_1(B).$$

Proof : Any loop ℓ in $A \times B$ determines loops a, b in A, B by $\ell(s) = (a(s), b(s))$, and homotopy is respected by this decomposition, i.e. $\ell \simeq \ell' \iff a \simeq a'$ and $b \simeq b'$. □

Recall that we have previously constructed a homeomorphism from the complement of a great circle C_1 in S^3 to a product, $S^3 - C_1 \cong S^1 \times P$. where P is a half-plane, and therefore simply-connected. Now $\pi_1(S^1 \times P) \cong C_\infty \times \{e\} \cong C_\infty$, generated by the class of the loop $\ell(s) = (a(s), p)$, where a is the loop of degree 1 round the circle and p is some fixed point of P. Hence $\pi_1(S^3 - C_1)$ is infinite cyclic, generated by the image in S^3 of the loop ℓ; this may be imagined from its image in $\mathbf{R}^3 - z$-axis, where it runs once round the axis.

Under the homeomorphism to the product $S^1 \times P$ the great circle C_2 becomes the circle $S^1 \times \{(1, 0)\}$ so that $S^3 - (C_1 \cup C_2) \cong S^1 \times (P - \{(1, 0)\})$. Now $\pi_1(P - \{(1, 0)\}) \cong C_\infty$, generated by a loop b in P around $(1, 0)$. Then $\pi_1(S^1 \times (P - \{(1, 0)\})) \cong C_\infty \times C_\infty$ and the two infinite cyclic groups are generated by loops $(a(s), \text{constant})$ and $(\text{constant}, b(s))$ respectively. Carrying these back to S^3 shows that the group of the Hopf link is isomorphic to $C_\infty \times C_\infty$, and is generated by two loops, one encircling C_1 and the other encircling C_2.

It is not possible to give such a simple description of a knot complement in general. However I shall give a straightforward prescription in the next section for a presentation of the group of a knot K starting from a diagram of K. The justification for the method, and also for the lack of distinction between the group of a knot in S^3 and in $\mathbf{R}^3$ is a result of the important theorem of van Kampen, which is summarised next.

3.4 van Kampen's theorem

This result gives a general and powerful method for building up knowledge of the fundamental group of a space X in terms of the fundamental groups of some constituent pieces of X.

Given $X = U_1 \cup U_2$, with $U_0 = U_1 \cap U_2$ and each of the three sets U_i open rel X and *path-connected.*
The goal of van Kampen's theorem is a description of $\pi_1(X, u_0)$ in terms of $\pi_1(U_1), \pi_1(U_2)$ and $\pi_1(U_0)$, for a choice of base point $u_0 \in U_0$.

Theorem 3.21 (van Kampen's theorem I, (generators): Geometric form) *Let $X = U_1 \cup U_2$, with $U_0 = U_1 \cap U_2$ and $u_0 \in U_0$ and suppose that each of the three U_i is open rel X and* path-connected.

Then every loop a at u_0 in X is homotopic (in X) to a composite of loops $a_1.a_2 \ldots a_k$ with each loop a_i lying either in U_1 or U_2.

Proof : Apply Lebesgue's lemma to the continuous map $a : I \to X$ to find k so that each subinterval of i of length $\leq 1/k$ is mapped by a into either U_1 or U_2. Write $b_i = a \circ \ell_{(i-1)/n,i/n}$, where ℓ_{PQ} is the straight line path from P to Q. Then each b_i is a path in either U_1 or U_2 and $a = b_1.b_2 \ldots b_k$.

Convert each of these paths into a *loop* at u_0 lying in the same U_i, while altering their product a only up to homotopy as follows. For each point $u_i = a(i)$ choose a path d_i from u_i to u_0 lying entirely in U_1 if $u_i \in U_1$, and entirely in U_2 if $u_i \in U_2$, (hence in U_0 if $u_i \in U_0$). Take d_0 and d_k to be the constant path at u_0.

Then $a_i = \overline{d}_{i-1}.b_i.d_i$ is a loop at u_0 lying either in U_1 or U_2 and $a_1.a_2 \ldots a_k \simeq b_1.b_2 \ldots b_k = a$ as required, writing $\overline{d}$ for the reverse of the path d. □

Corollary 3.22 *If $X = U_1 \cup U_2$ as in the hypotheses of the theorem (both open rel X with $U_1 \cap U_2$ path connected) and U_1 and U_2 are both* simply connected *then X is simply connected.*

Proof : Any loop a at $u_0 \in X$ is homotopic to $a_1.a_2 \dots a_k$ with each a_i being a loop in U_1 or U_2. Hence each a_i is homotopic, in U_1 or U_2 and thus in X, to the constant loop at u_0. Then the composite loop $a_1.a_2 \dots a_k$ is homotopic in X to the constant loop at u_0. □

Example. The sphere S^n is simply-connected, for $n \geq 2$.

For we need only apply the corollary taking U_1 and U_2 to be $S^n - \{N\}$ and $S^n - \{S\}$, the complements of the North and South poles, and observing that $U_0 = S^n - \{N, S\} \cong \mathbf{R}^n - \{0\}$ is path connected for $n \geq 2$.

To get a good algebraic view, use the language of **generators** and **relations** for groups.

Definition. A set of elements B in a group G (written multiplicatively) *generate* G if each $g \in G$ can be written as $g = g_1.g_2 \dots g_m$ with either $g_i \in B$ or $g_i^{-1} \in B$ for each i.

Such an expression $g_1.g_2 \dots g_m$ is called a *word* in the generators B.

Example. If $B = \{x, y\}$ generates G then elements of G consist of products such as $x^2y^{-1}x^{-2}y^3$.

Definition. G is *finitely-generated* if it has a finite generating set $\{b_1, \dots, b_k\}$.

Remark. A *cyclic* group is a group with a single-element generating set $\{b_1\}$.

Remark. If $f : G \to H$ is a group homomorphism and G is generated by $\{b_1, \dots, b_k\}$ then the image $f(G)$ is generated by $\{f(b_1), \dots, f(b_k)\}$, hence the image of a cyclic group is always cyclic.

Note. In the case of an *abelian* group G, where the group operation is written additively, the set of words in $\{b_1, \dots, b_k\}$ are just the elements $\Sigma\lambda_i b_i,\ \lambda_i \in \mathbf{Z}$. For example, $G = \mathbf{Z} \times \mathbf{Z}$ can be generated by $b_1 = (1, 0),\ b_2 = (0, 1)$, or by $\{(2, 1), (3, 1)\}$ but not by $\{(2, 0), (0, 1)\}$.

It follows from standard linear algebra that $G = \mathbf{Z}^k$ can never be generated by fewer than k elements, for a generating set must also generate $\mathbf{R}^k$ in the usual sense of linear algebra. (Why?)

We can then formulate van Kampen's theorem algebraically

Theorem 3.23 (van Kampen's theorem I, (generators): Algebraic form) *Let $X = U_1 \cup U_2$, with $U_0 = U_1 \cap U_2$ and $u_0 \in U_0$ and suppose that each of* *the three U_i is open rel X and* path-connected. *Write $j_1 : U_1 \to X$ and $j_2 : U_2 \to X$ for the inclusion maps.*

Then $\pi_1(X, u_0)$ is generated by the union of the two subsets $j_{1*}(\pi_1(U_1, u_0))$ and $j_{2*}(\pi_1(U_2, u_0))$, i.e. every element of $\pi_1(X, u_0)$ is a product of elements or their inverses taken from these two subsets.

Proof : Note that an element of $j_{1*}(\pi_1(U_1, u_0))$ is simply the homotopy class in X of a loop which is homotopic to a loop in U_1. Starting with $[a] \in \pi_1(X, u_0)$ we can use the geometric formulation to write $[a] = [a_1].[a_2] \dots [a_k]$ where each a_i is a loop in U_1 or U_2 so that $[a_i]$ lies in one of the two subsets claimed. □

Corollary 3.24 *Suppose that $\pi_1(U_1, u_0)$ and $\pi_1(U_2, u_0)$ are generated by $b_1, \dots, b_r$ and $c_1, \dots, c_s$ respectively, then $\pi_1(X, u_0)$ is generated by $\overline{b}_1, \dots, \overline{b}_r, \overline{c}_1, \dots, \overline{c}_s$, where $\overline{b}_1 = j_{1*}(b_1)$ etc.*

Example. We can write $\mathbf{R}^2 - \{k \text{ points}\}$ as $U_1 \cup U_2$ with $U_1 \cong \mathbf{R}^2 - \{r \text{ points}\}$ and $U_2 \cong \mathbf{R}^2 - \{s \text{ points}\}$, where $k = r + s$, and $U_1 \cap U_2$ is path-connected (in fact convex).

Then, by induction on k there is a generating set of k elements for $\pi_1(\mathbf{R}^2 - \{k \text{ points}\})$ ($= G_k$ say). Indeed it is not difficult to use the theorem to find k explicit loops whose homotopy classes generate the fundamental group, G_k. We can show also that G_k can not be generated by **fewer** than k elements, by considering the homomorphism $f : G_k \to \mathbf{Z}^k$ given by

$$f([a]) = (d_1(a), \dots, d_k(a))$$

where $d_i(a)$ is the winding number of a about the ith point.

Now f is surjective, because we can find a loop a_i for each i with $d_i(a_i) = 1$, $d_j(a_i) = 0$, $j \neq i$. The image of any generating set for G_k will then generate $\mathbf{Z}^k$, and we noted above that $\mathbf{Z}^k$ requires at least k generators.

It is not clear yet just what the fundamental group of $\mathbf{R}^2 - \{2 \text{ points}\}$ is. While we have a set of generators $[a_1], [a_2]$ we do not know if they commute, showing that G_2 is abelian, or whether they have any other relations between them. The only piece of information from the previous example is that it is impossible to find one single generator for G_2.

The next part of van Kampen's theorem gives us a much clearer view of the group, by showing how to find **relations** as well as generators.

Definition. A group G has a *presentation* with generators $b_1, \ldots, b_k$ and relations $v_1 = w_1, \ldots, v_s = w_s$, where each v_i and w_i is a word in the generators, if

1. G is generated by $\{b_1, \ldots, b_k\}$, and
2. Two words in the generators give the same element of $G \Leftrightarrow$ one word can be changed to the other by a sequence of the following moves or their inverses:

   (i) replace $b_r b_r^{-1}$ by e (the identity element),

   (ii) remove e,

   (iii) replace the subword v_i by w_i.

We then write

$$G =< b_1, \ldots, b_k : v_1 = w_1, \ldots, v_s = w_s > .$$

Examples.

1. $G =< x : x^2 = e >$, the cyclic group of order 2.
2. $G =< x : >$, the infinite cyclic group.
3. $G =< x, y : x^3 = e, \, y^2 = e, \, yx = x^{-1}y > \cong S_3$ (the permutation group on 3 objects).

The great thing about a presentation of a group G is that it enables you to write down homomorphisms from G to other groups.

If $f : G \to H$ is a homomorphism then we can find $f(g)$ for any element $g \in G$ once we know $f(b_1), \ldots, f(b_k)$ for a set of generators.

Theorem 3.25 (Presentation theorem) *If* $G =< b_1, \ldots, b_k : v_1 = w_1, \ldots, v_s = w_s >$ *and we can find* $\overline{b}_1, \ldots, \overline{b}_k \in H$ *such that* $v_i(\overline{b}_1, \ldots, \overline{b}_k) = w_i(\overline{b}_1, \ldots, \overline{b}_k)$ *for each* i *then there is a homomorphism* $f : G \to H$ *with* $f(b_j) = \overline{b}_j$ *for each* j.

Proof : Given $g \in G$ write it as a word in the generators, say $g = g_1 \dots g_n$ where each $g_j = b_i^{\pm 1}$ for some i . Define $f(g) = f(g_1) \dots f(g_n)$ where $f(g_j) = \overline{b}_i^{\pm 1}$.

By checking through the moves allowed in 2. it can be seen that this definition does not depend on how g has been written as a word in the generators. It is then immediate that f is a homomorphism, for if $g = g_1 \dots g_n$ and $h = h_1 \dots h_k$ are written as words in the generators we can then write $gh = g_1 \dots g_n h_1 \dots h_k$ and, by the definition of f,we have $f(gh) = f(g_1) \dots f(g_n) f(h_1) \dots f(h_k) = f(g)f(h)$. □

Example. We can define a homomorphism $f : G \to Gl(2, \mathbf{R})$ where $G =< x, y : xyx = yxy >$ by choosing $\overline{x} = \begin{pmatrix} 1 & 0 \\ -1 & 1 \end{pmatrix}$ and $\overline{y} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$.

The only check required is that $\overline{x}\,\overline{y}\,\overline{x} = \overline{y}\,\overline{x}\,\overline{y}$.

Important groups from this point of view are the *free groups*, for example $G =< x_1, \dots, x_k : >$, with **no** relations. A homomorphism from the free group G is determined by a free choice of the images of $x_1, \dots, x_k$.

When $k = 1$ the free group is infinite cyclic; for $k > 1$ the free group is non-abelian.

To return to the setting of van Kampen's theorem, suppose that $\pi_1(U_1)$ has generators $x_1, \dots, x_k$ and relations $v_1 = w_1, \dots, v_s = w_s$, where each v_i and w_i is a word in the generators. Write $\overline{x}_i = j_{1*}(x_i) \in \pi_1(X)$ etc., and put $\overline{v}_i = v_i(\overline{x}_1, \dots, \overline{x}_k)$ etc. for the corresponding words in the elements $\overline{x}_1, \dots, \overline{x}_k$.

Since j_{1*} is a homomorphism it follows that $\overline{v}_i = j_{1*}(v_i)$, so the relations $\overline{v}_1 = \overline{w}_1$ etc. will hold in $\pi_1(X)$.

Similarly, suppose that

$$\pi_1(U_2) =< y_1, \dots, y_r : t_1 = u_1, \dots, t_\ell = u_\ell >$$

and write $\overline{y}_i = j_{2*}(y_i)$ etc.

Then we know already, by van Kampen I, that $\pi_1(X)$ is generated by $\overline{x}_1, \dots, \overline{x}_k$ and $\overline{y}_1, \dots, \overline{y}_r$ so it remains to give a sufficient set of relations.

These come from the relations in the presentations of $\pi_1(U_1)$ and $\pi_1(U_2)$, together with further relations which take care of the fact that an element of $\pi_1(X)$ which is represented by a loop in $U_0 = U_1 \cap U_2$ will appear as the image of an element of $\pi_1(U_1)$, written in the generators $\overline{\mathbf{x}}$, and also of an element in $\pi_1(U_2)$, written in the generators $\overline{\mathbf{y}}$.

Write $i_1 : U_0 \to U_1$ and $i_2 : U_0 \to U_2$ for the inclusions, and suppose that $\pi_1(U_0)$ has generators $z_1, \ldots, z_m$. Then $\overline{i_{1*}(z_j)} = \overline{i_{2*}(z_j)}$ for each j.

Theorem 3.26 (van Kampen's theorem II (relations)) *Let X, U_0, U_1, U_2 be as before, and let*

$\pi_1(U_1) =< \mathbf{x} : \mathbf{v} = \mathbf{w} >,\ \pi_1(U_2) =< \mathbf{y} : \mathbf{t} = \mathbf{u} >, \pi_1(U_0) =< \mathbf{z} :$ *relations* $>$.

Then $\pi_1(X) =< \overline{\mathbf{x}}, \overline{\mathbf{y}} : \overline{\mathbf{v}} = \overline{\mathbf{w}}, \overline{\mathbf{t}} = \overline{\mathbf{u}}, \overline{i_{1*}(\mathbf{z})} = \overline{i_{2*}(\mathbf{z})} >$.

Proof : This requires an analysis of the way in which two homotopic loops in X can be written as a product of loops in U_1 and U_2 by breaking the homotopy up into small squares each of which maps into one or other of the subspaces, and using these to find a sequence of representations of the element of $\pi_1(X)$ as words in the generators, differing only as specified by the claimed relations. □

Remark. The result gives a presentation of $\pi_1(X)$ as the 'union' of presentations of $\pi_1(U_1)$ and $\pi_1(U_2)$ with extra relations to say that the elements determined by any generator of $\pi_1(U_0)$ in these two groups are equal.

There are many cases where it does in fact work for suitable *closed* subsets U_1, U_2, although it does not always apply.

Example. The group $G_2 = \pi_1(\mathbf{R}^2 - \{2 \text{ points}\})$ is free on two generators, $G_2 =< x_1, x_2 : >$.

For we can take U_1 and U_2 as open half-spaces with a point removed, each with infinite cyclic fundamental group $< x_1 : >, < x_2 : >$.

We can also arrange that U_0 is convex, hence simply-connected, so that there are no generators of $\pi_1(U_0)$. Then van Kampen's theorem gives the presentation claimed, where $\overline{x}_1, \overline{x}_2$ have been replaced by x_1, x_2.

The example of $G_k = \pi_1(\mathbf{R}^2 - \{k \text{ points}\})$, discussed earlier after Corollary 3.24, can similarly be completed by the use of Theorem 3.26 to show that G_k is free on k generators, each represented by a simple loop round one of the missing points.

Remark. The fact that S^2 and similarly $\mathbf{R}^3$ − point are simply-connected (noted earlier) allow a general proof that a point can be removed from any open subset of $\mathbf{R}^3$ without altering its fundamental group.

Try using van Kampen's theorem to establish this. Then deduce that the group of the unknot in $\mathbf{R}^3$ (rather than S^3) is infinite cyclic.

4 The group of a knot

Many properties of a knot can best be defined without reference to a specific diagram, for example the group $G_K = \pi_1(\mathbf{R}^3 - K)$. It is, however, particularly useful if detailed calculations can be made using just one diagram. This is the case for G_K, and I shall now give an explicit way to find a presentation of the group, starting from any given diagram of K.

4.1 Wirtinger's presentation

Given a diagram of K, with k crossing points, we can think of the curve K as divided into k arcs by the undercrossing points. Choose an orientation for K, and select one crossing point. Start on the undercrossing arc from the chosen crossing point c_0 and label the arcs successively $1, \ldots, k$, labelling the crossing points as they appear in order as undercrossings, so that arc i runs from c_{i-1} to c_i (possibly passing several overcrossings on the way). Finish by setting $c_k = c_0$.

For a link L with several components, $L_x, L_y, L_z, \ldots$ say, label each component in turn, with crossings $c_0^{(x)}, \ldots, c_{k_x}^{(x)}$ on L_x, $c_0^{(y)}, \ldots, c_{k_y}^{(y)}$ on L_y and so on, where $c_i^{(y)}$ is then a crossing point at which the component L_y is the undercrossing.

Theorem 4.1 *Given a diagram of a knot K with k crossings the group G_K can be presented with generators $x_1, \ldots, x_k$ and k relations, one for each crossing in the diagram, to be described shortly.*

Proof : Take the base point of the fundamental group G_K to be the point from which the diagram is viewed. The generator x_i is defined as the element represented by a loop, which I also call x_i, passing from the base point, (i.e. our viewpoint), by a straight line from the eye followed by a path crossing directly under arc i from right to left and then a straight line back to the eye, as in figure 7. Any two such loops around the arc i are homotopic in $\mathbf{R}^3 - K$; another loop also representing x_i is shown in figure 8.

Any loop in $\mathbf{R}^3 - K$ which is made up of a path joined to the eye (base-point) by two straight-line segments will appear in the diagram simply as the image of the path – the loop x_i for example will simply look like a path crossing under the arc i.

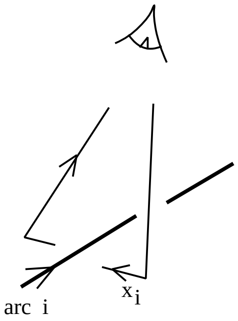


Figure 7:

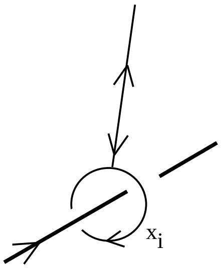
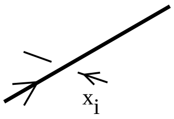


Figure 8:

Every path in $\mathbf{R}^3 - K$ whose endpoints are visible defines a loop, by joining the ends to the eye along the line of sight, and hence determines an element of G_K. If we choose an intermediate point in a path, also visible, and join this also to the eye, we can define two loops, one from each part of the path, whose composite is homotopic to the loop defined by the original path. The two new loops then determine elements of G_K whose product is the original element. Any path which is completely visible defines a loop that is homotopic in $\mathbf{R}^3 - K$ to the constant loop at the eye, simply by pulling all points back to the eye along straight lines, and hence determines the identity element of G_K. It is then possible to read off the element in G_K represented by any path which crosses finitely often under K in the diagram, as a word in the elements $x_i^{\pm 1}$. Any loop in $\mathbf{R}^3 - K$ is homotopic to a loop of this type, so we have established that the elements $x_i^{\pm 1}$ can be taken to generate G_K. $\square$

Remark. A fully detailed version of this result would use van Kampen's



Plan view (from the eye)

Figure 9:

theorem; further use of the theorem will guarantee that the relations given below are sufficient to give a complete presentation of G_K.

At the crossing c_i, where the incoming arc is i and the outgoing arc is $i+1$, let us suppose that the arc $j(i)$ forms the overcrossing arc and that it crosses with sign $\varepsilon(i)$, under the same convention as for the discussion of linking numbers.

Consider a loop in the form of a square lying underneath c_i and crossing once under each of the four pieces of arc which meet there. Join one corner of the square to the eye to give a loop which represents an element of G_K. This element must be the identity $e \in G_K$ because the loop is homotopic to a constant, by simply pulling the square out to its corner, when it is all visible and can be moved back to the eye. On the other hand, the technique above allows us to read off the element represented by the square as a product of elements coming one from each side of the square by joining up to the eye. Depending on how we read round the square the elements determined by the sides will be $x_i^{\pm 1}, x_{i+1}^{\pm 1}$ and $x_{j(i)}^{\pm 1}$.

We can alternatively read the relation as giving x_{i+1} in terms of x_i and $x_{j(i)}$, when we read round three sides of the square in place of one. We then have the relation

$$x_{i+1} = x_{j(i)}^{-\varepsilon(i)} x_i x_{j(i)}^{\varepsilon(i)}$$

from the ith crossing. The case of a positive crossing is illustrated in figure 10.

Theorem 4.2 *Wirtinger's full presentation theorem says that G_K has a presentation*

$$G_K \cong < x_1, \ldots, x_k; x_{i+1} = x_{j(i)}^{-\varepsilon(i)} x_i x_{j(i)}^{\varepsilon(i)} > .$$

Remark. In the case of a link with several components the presentation is exactly similar, with one generator $x_1, \ldots, x_{k_x}, y_1, \ldots, y_{k_y}, \ldots$ for each arc

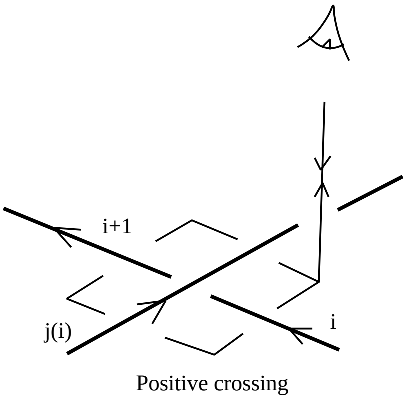


Figure 10:

and one relation for each crossing; for a crossing $c_i^{(y)}$ in which the undercrossing component is L_y the relation will have the form $y_{i+1} = g^{-1}y_i g$ where g depends on the overcrossing arc at $c_i^{(y)}$, and will be the generator for that arc, or its inverse.

It follows at once that all the generators of G_L coming from a particular component, L_y for example, are conjugate, i.e. $y_j = g^{-1}y_i g$ for each i, j and some $g \in G_L$, since y_{i+1} is conjugate to y_i for each i.

Remark. This illustrates a general result about fundamental groups, namely that any closed loop in $\mathbf{R}^3 - L$ not necessarily through the base point will define an element of G_L by first choosing a path from the base point to a point of the loop, then going round the loop, and retracing the path to the base point. The element of G_L so defined depends on the original loop *and* on the path chosen, but alteration of the path alters the element of G_L only up to conjugacy, as does any movement of the original loop in $\mathbf{R}^3 - L$. Then loops, without a special restriction on base point, correspond to conjugacy classes in G_L; the conjugacy class of y_1 for example is related to a *meridian loop* about L_y, i.e. a loop around the edge of a small disc which crosses L_y transversely, as illustrated in figure 8 above. The other generators y_i then clearly belong to the same conjugacy class, as they too can be represented by meridian loops.

4.2 The group of the trefoil

To find a presentation for the group G_K of the trefoil knot K we can apply Wirtinger's method to the diagram illustrated in figure 11. We start with three generators x_1, x_2, x_3, one for each arc. The relation from crossing c_0 allows us to rewrite $x_3 = x_2 x_1 x_2^{-1}$ and so present the group with just two generators. The relation from crossing c_1 then reads $x_2 x_1 x_2^{-1} = x_1^{-1} x_2 x_1$ and the remaining relation, which is equivalent to this, reads $x_2 x_1 x_2^{-1}.x_2 = x_1.x_2 x_1 x_2^{-1}$.

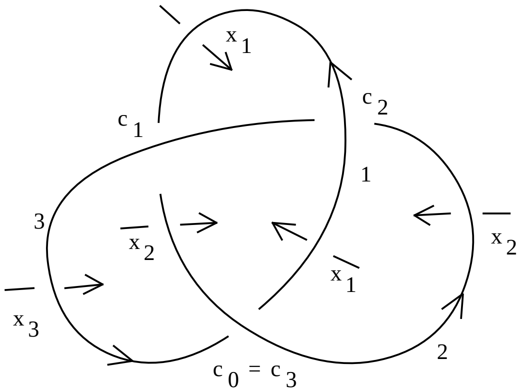


Figure 11:

We can rewrite these relations to give

$$G_K = < x_1, x_2 : x_1 x_2 x_1 = x_2 x_1 x_2 > .$$

The curves x and y illustrated in figure 12 represent $x_2 x_1$ and $x_1 x_2 x_1 x_2^{-1} x_2$ respectively, when completed to a loop by a straight line path to and from the point P.

Now $x = x_2 x_1$ and $y = x_1 x_2 x_1$ so $x_1 = y x^{-1}$ and $x_2 = x x_1^{-1} = x^2 y^{-1}$ so G_K is equally generated by x and y while the relation becomes $y^2 = x^3$ in terms of the new generators.

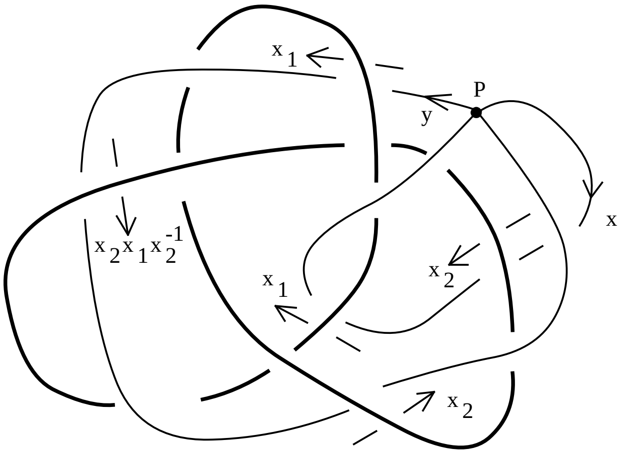


Figure 12:

5 Representations of knot groups

We shall now look at possible homomorphisms from knot groups to other groups.

In this way we may be able to show that two knot groups are not isomorphic (and hence the knots are not equivalent) by comparing the possible homomorphisms from the two knot groups to some (simpler) fixed group.

We look first at homomorphisms to abelian groups.

5.1 Abelianisation of a group

Every group G has an 'abelianisation', G/G', defined formally by factoring out the commutator subgroup G'; this is the normal subgroup of G generated by the elements $aba^{-1}b^{-1}$ with $a, b \in G$, (see any text on group theory for further discussion and properties). The effect of applying the natural homomorphism $\varphi : G \to G/G'$ is simply to map every pair of elements in G to a pair which commute. A presentation for G/G' can be given readily from a presentation of G by adding the relations that all the generators commute.

One immediate result about abelianisations is that every homomorphism $\theta : G \to H$ where H is abelian must factor through the abelianisation G/G'. That is, there exists a homomorphism $\overline{\theta} : G/G' \to H$ with $\theta = \overline{\theta} \circ \varphi$.

5.1.1 Knot and link groups

Let us distinguish between the generators of G and its abelianisation G/G' by writing $\overline{g} = \varphi(g) \in G/G'$ for the element arising from g in the abelianisation. For a link L a Wirtinger presentation for G_L then gives an immediate presentation for the group G_L/G'_L, with generators $\overline{x}_1, \ldots, \overline{y}_{k_y}, \ldots$. These generators commute, and the Wirtinger relations give further relations of the form $\overline{y}_{i+1} = \overline{g}^{-1}\overline{y}_i\overline{g}$, which can be rewritten simply as $\overline{y}_{i+1} = \overline{y}_i$. Set $t_x = \overline{x}_1 = \overline{x}_2 = \ldots, t_y = \overline{y}_1 = \overline{y}_2 = \ldots, \ldots$. Then the abelianisation is generated as an abelian group by r elements $t_x, t_y, \ldots$, one for each component of L, with no further relations, and so it is isomorphic to the free abelian group $C_\infty \times \ldots \times C_\infty$ of rank r , where L has r components. In particular the group G_K of any knot K abelianises to the infinite cyclic group, $C_\infty \cong \mathbf{Z}$, generated by $t = t_x$, represented by any meridian loop.

For a general space X the abelianisation of the fundamental group $\pi_1(X)$ is known as the first homology group $H_1(X)$. Recall that any continuous map

$f : X \to Y$ induces a homomorphism $f_* : \pi_1(X) \to \pi_1(Y)$. By considering the homomorphism $\varphi_Y \circ f_*$ from $\pi_1(X)$ to the abelian group $H_1(Y)$ we see that there is a homomorphism, which we also write as f_*, from $H_1(X)$ to $H_1(Y)$, with $\varphi_Y \circ f_* = f_* \circ \varphi_X$.

The fact that the abelianisation of the group of a link L depends only on the number of components of the link shows that the homology group $H_1(\mathbf{R}^3 - L)$ of a link complement is not a useful invariant in distinguishing between links.

Remark. Note that for the trefoil in figure 12, the abelianisation map $\varphi : G_K \to G_K/G'_K$ maps the generators x_1, x_2 to the generator $t = \varphi(x_1)$, while $\varphi(x) = t^2, \varphi(y) = t^3$. These powers in fact correspond to the linking numbers of the curves x and y with K, as we shall shortly see in general.

5.1.2 Linking numbers revisited

We have already looked at the linking number of two curves $K_1 \cup K_2$ as an example of an invariant which was defined originally from one diagram of the link, and then shown not to depend on the diagram chosen. We can give an alternative proof of independence by use of the group of one component, K_1 say.

For any knot K, the group $H_1(\mathbf{R}^3 - K)$ is infinite cyclic, generated by an element t which can be represented by any positively oriented meridian curve, assuming that an orientation of K has been chosen. (Choice of the opposite orientation of K would lead to the use of t^{-1} in place of t.) An equivalence h from K to K' determines isomorphisms h_* from G_K to $G_{K'}$ and from $H_1(\mathbf{R}^3 - K)$ to $H_1(\mathbf{R}^3 - K')$. A positively oriented meridian curve of K will be carried by h to a positively oriented meridian of K', since h is orientation preserving. The induced isomorphism from $H_1(\mathbf{R}^3 - K)$ to $H_1(\mathbf{R}^3 - K')$ then carries the generator t to the generator t', where both are represented by positively oriented meridians. (An orientation reversing homeomorphism, such as a reflection, would carry t to t'^{-1}.)

Suppose now that $K_1 \cup K_2$ is a link, with a chosen orientation for each component. Regard K_2 as an oriented curve in the complement of K_1. Then K_2 represents an element $k_2 \in G_{K_1}$ up to conjugacy, and so determines an element $\varphi(k_2) = t^\ell \in G_{K_1}/G'_{K_1} = H_1(\mathbf{R}^3 - K_1)$ which depends only on K_2. Choose a diagram of $K_1 \cup K_2$ and use it to give a Wirtinger presentation of G_{K_1}. The curve K_2 will then give an element k_2 as a product of generators of G_{K_1}, one for each crossing of K_2 under K_1 in the diagram, with sign depending on the sign of the crossing. Thus

$$k_2 = x_{i_1}^{\varepsilon(i_1)} x_{i_2}^{\varepsilon(i_2)} \dots x_{i_s}^{\varepsilon(i_s)}$$

say, where K_2 crosses under the arcs $i_1, i_2, \dots, i_s$ of K_1 in turn. Since $\varphi(x_i) = t$ for each arc i of K_1 we have $\varphi(k_2) = t^\ell$, where

$$\ell = \sum_{j=1}^{s} \varepsilon(i_j) = \mathrm{lk}(K_1, K_2).$$

This gives an alternative proof of the invariance of linking number, since any equivalence h carrying $K_1 \cup K_2$ to $K_1' \cup K_2'$ will take k_2 to $h(k_2)$, and $\varphi(h(k_2)) = h_*(\varphi(k_2)) = h_*(t^\ell) = t'^\ell$. In this setting the linking number of K_2 with K_1 may then be defined as the exponent ℓ where K_2 represents t^ℓ in $H_1(\mathbf{R}^3 - K_1)$, and t is the generator represented by the positively oriented meridian.

5.2 Knot colouring

We move now from abelian representations of a knot group (which can only have cyclic image) to look at some other homomorphisms from knot groups that can sometimes be used to tell knots apart. The existence of such homomorphisms will be shown to be detectable from a single diagram of the knot.

Definition. The knot diagram D_K can be n-coloured if we can assign a colour d_i to each arc i, drawn from a palette of n colours labelled $0, \dots, n-1$, so as to satisfy the following requirements:

1. At the crossing c_i the colour of the overcrossing arc $j(i)$ must be the average $\bmod n$ of the colours of the other two incident arcs i and $i+1$, in other words
$$2d_{j(i)} = d_i + d_{i+1} \mod n,$$
2. The colours d_i must not all be congruent $\bmod r$ for any factor $r > 1$ of n.

Theorem 5.1 *(a) If any one diagram D_K of a knot K can be n-coloured then there is a surjective homomorphism $d : G_K \to D_n$, where D_n is the dihedral group of symmetries of a regular n-gon.*

(b) If there is a surjective homomorphism $d : G_K \to D_n$ then every *diagram of K can be n-coloured.*

Proof : There is a presentation of the group D_n with two generators, a rotation a, having $a^n = e$, and a reflection $b = b_0$, say. Set $b_r = a^r b$, so that the reflections in D_n are written $b_0, \ldots, b_{n-1}$. Note that $b_r b_s = a^{r-s}$.
(a) Define d by $d(x_i) = p_i$ where p_i is the reflection b_{d_i}. This will determine a group homomorphism provided that the relations in the Wirtinger presentation of G_K are respected. Thus we require $p_{j(i)} p_{i+1} = p_i p_{j(i)}$ for each i. Now

$$\begin{aligned} p_{j(i)} p_{i+1} &= a^{d_{j(i)} - d_{i+1}} \\ p_i p_{j(i)} &= a^{d_i - d_{j(i)}} \end{aligned}$$

while condition (1) for the diagram shows that

$$d_{j(i)} - d_{i+1} = d_i - d_{j(i)} \bmod n.$$

Condition (2) guarantees that the homomorphism d is surjective, since all powers $a^{d_i - d_j}$ lie in the image of d and, by (2), they generate the whole cyclic subgroup generated by a. Together with any reflection p_i the whole of D_n then lies in the image of d.
(b) Use the Wirtinger presentation for a given diagram of K. Since all meridians of G_K are conjugate it follows that $d(x_i)$ must be a reflection for each i, otherwise $d(x_i)$ is a rotation for each i and so the image of d consists entirely of rotations, and is not the whole of D_n.

Write $d(x_i) = b_{d_i}$ to define d_i, the colour for the i-th arc. This determines an n-colouring for the diagram, since condition (1) follows from the relations in G_K while (2) follows from the surjectivity of d. □

Examples. It is clear equally from the group-theoretic test, or from the diagram check, that the unknot cannot be n-coloured for any n.

The trefoil can be 3-coloured, but cannot be n-coloured for any larger n. (prove this)

The figure-eight knot cannot be 3-coloured, but can be 5-coloured.

Show how to distinguish the knot in figure 13 from the trefoil using n-colouring for some suitable choice of n.

Remark. For n prime, the condition (2) is the same as saying that the same colour must not be used for all arcs, i.e. that at least two colours must be used.

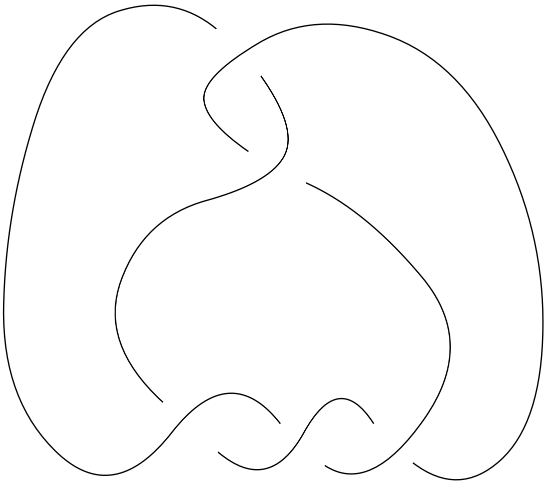


Figure 13:

We can write an equivalent version of condition (1) in terms of linear equations over $\mathbf{Z}_n$.

Suppose that the diagram which we are using has k crossings. Write A for the $(k-1)\times k$ matrix with entries $a_{i\,i} = a_{i\,i+1} = -1, a_{i\,j(i)} = 2$ and other entries zero. We require a solution

$$\mathbf{d} = \begin{pmatrix} d_1 \\ \vdots \\ d_k \end{pmatrix}$$

to the equations $\mathrm{A}\mathbf{d} = \mathbf{0}$ which must not be a multiple of $\begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$.

When n is prime this last restriction is sufficient for condition (2), and it is then equivalent to asking that A should have rank $< k-1$, i.e. that the matrix given by deleting a column of A should have determinant equal to zero mod n. It can be shown from other considerations that this matrix will always have non-zero determinant (as an integer), and in fact will be non-zero mod 2. There are thus only a finite number of possible prime n for which any given K can be n-coloured.

Note for later. The matrix $\mathrm{A}(t)$ with entries $a_{i\,i} = -1, a_{i\,i+1} = t, a_{i\,j(i)} = 1 - t$ for a positive crossing, and $a_{i\,i+1} = -1, a_{i\,i} = t, a_{i\,j(i)} = 1 - t$ for a negative crossing is the Alexander matrix derived from the Wirtinger presentation with the given knot diagram. The Alexander polynomial $\Delta_K(t)$ is the determinant of the $(k-1) \times (k-1)$ matrix given by deleting a column. Now the matrix A above is $\mathrm{A}(-1)$, so for n prime

$$K \text{ is } n\text{-colourable} \iff \Delta_K(-1) = 0 \bmod n.$$

6 Knots and surfaces

We shall be able to find information about a knot K in $\mathbf{R}^3$ by constructing a surface F lying in $\mathbf{R}^3$ of which K forms the boundary ∂F. We call any such surface a *spanning surface* for K. Anyone familiar with vector calculus will recall the relations given by Stokes' theorem between integrals of vector fields across a surface F and integrals around its boundary ∂F for example.

As in Stokes' theorem, we shall only consider orientable spanning surfaces. I shall summarise some facts about surfaces, considered first as 'abstract' surfaces, that is simply as topological spaces not necessarily lying in $\mathbf{R}^3$, and then look at some features of surfaces when they are embedded in $\mathbf{R}^3$. For further background consult a text such as Armstrong's 'Basic Topology'.

6.1 Surfaces on their own

There are a number of equivalent definitions of surfaces, using either local properties as 2-dimensional topological or smooth manifolds, or a more combinatorial approach in terms of triangulations. For our purposes the combinatorics will give the most directly applicable view. We take a compact surface with boundary F to be made up of a finite union of pieces $\{P_i\}$, each homeomorphic to a closed disc, D_i say, in $\mathbf{R}^2$. Two pieces only meet along their boundary (the image of the bounding circle of D_i); the boundary of each P_i is made up of a finite number of intervals in such a way that each interval forms part of the boundary of at most one other P_j.

We take the orientation of every simple closed curve in $\mathbf{R}^2$ to be anticlockwise. Then the boundary of each disc D_i in $\mathbf{R}^2$ has a chosen orientation, which defines, by the homeomorphism with P_i, an orientation of the boundary curve of P_i. Every interval $P_i \cap P_j$ common to two pieces is oriented in two ways, once as part of ∂P_i and once as part of ∂P_j. We call the surface *orientable* if the homeomorphisms can be chosen so that these two orientations are opposite for every interval $P_i \cap P_j$. We then say that we have made a *consistent* choice of homeomorphisms.

Remark. The possibility of making a consistent choice can be shown to hold for all dissections of a given F if it holds for one dissection. For example, if F is itself just a disc in $\mathbf{R}^2$ dissected into a union of polygons $\{P_i\}$ then we may take $D_i = P_i$, with the identity map, so that the boundary of each P_i is oriented anticlockwise. The common boundary between any P_i and P_j

will then inherit ‘cancelling’ orientations from the polygons on each side.

The boundary of F, consisting of one or more closed curves, is made up of the parts of the boundaries of P_i which occur in only one P_i. For an orientable surface each of these curves inherits an orientation from a consistent choice of homeomorphisms.

Euler characteristic. Any dissection of a surface F has an *Euler characteristic* χ defined by $\chi = P - E + V$ where P is the number of pieces, E is the number of edges, i.e. intervals on the boundaries of the pieces, and V is the number of vertices, the end points of the boundary intervals.

Theorem 6.1 *Any two dissections for a given F have the same Euler characteristic.*

Theorem 6.2 *Every orientable surface with r boundary curves is homeomorphic to $F_{g,r}$ for some g, where $F_{g,r}$ is an explicit surface, realisable in $\mathbf{R}^3$ as a sphere with g handles, and with r discs removed.*

Corollary 6.3 *An orientable surface is determined up to homeomorphism by its Euler characteristic and number of boundary components.*

Proof : $\chi(F_{g,r}) = 2 - r - 2g$ so we can find g from r and χ. □

Figure 14 shows a fairly standard view of $F_{g,1}$.

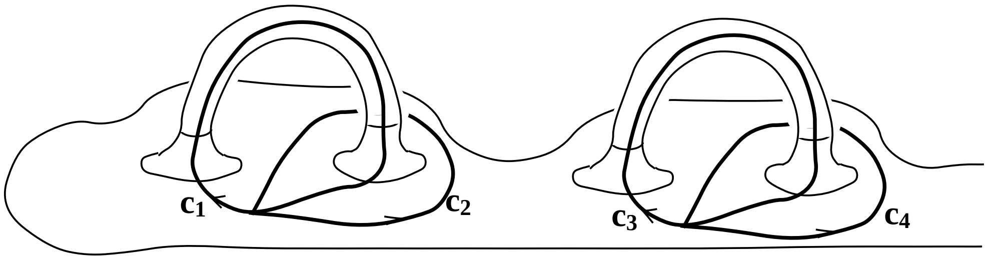


Figure 14:

The surface $F_{g,r}, r \geq 1$ may alternatively be viewed as the neighbourhood of a family of closed curves $c_1, \ldots, c_{2g}, d_1, \ldots, d_{r-1}$, and some arcs joining them together. The curves themselves may be taken to be disjoint, except that the pairs c_{2i-1}, c_{2i} each meet in a single point, and we may take the arcs to join the points of intersection of successive pairs, and then single points on each of the curves d_j, as in the illustrations below.

Figure 15 shows the alternative view of the surface $F_{g,1}$, drawn as a neighbourhood of the curves $\{c_i\}$.

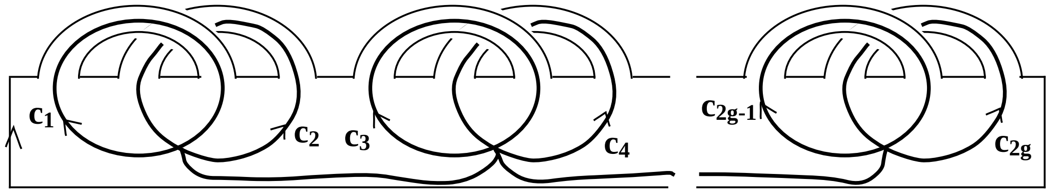


Figure 15:

For the surface $F_{g,r}$ add the modification in figure 16 to the right hand end of the picture.

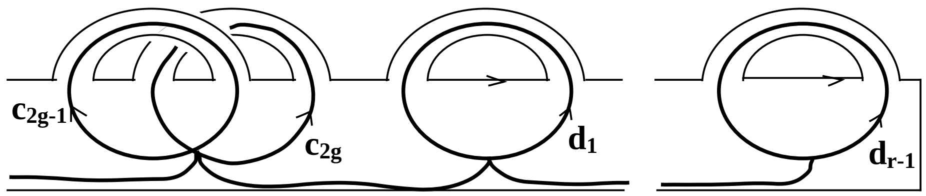


Figure 16:

The fundamental group $\pi_1(F_{g,r})$, based say at the first point of intersection, is a free group, generated by elements $c_1, \ldots, c_{2g}, d_1, \ldots d_{r-1}$, using the path determined by the arcs to pass from the base point to a point on the relevant closed curve. It is thus free of rank $2g + r - 1 = 1 - \chi$. Its abelianisation, $H_1(F_{g,r})$, is then free abelian of this rank. The r boundary curves can be taken to represent elements of $\pi_1(F)$ when we orient them and choose a path to each one from the base point. With the orientations shown in the illustration these curves represent

$$d_1, d_2, \ldots, d_{r-1} \text{ and } [c_1, c_2]\,[c_3, c_4] \ldots [c_{2g-1}, c_{2g}]\, d_1^{-1} d_2^{-1} \ldots d_{r-1}^{-1}.$$

Here $[a, b]$ stands for the commutator $[a, b] = aba^{-1}b^{-1}$. Consequently in $H_1(F)$ the boundary curves represent $\overline{d}_1, \ldots, \overline{d}_{r-1}$ and $\prod \overline{d}_j^{-1}$, in multiplicative notation, $(-\overline{d}_1 - \ldots - \overline{d}_{r-1}$ in additive notation). When $r = 1$ the single boundary curve represents the trivial element of $H_1(F)$.

Note. The orientations used above for the curves are consistent with a choice of orientation for F, so we see that in general, when F has been oriented and its boundary curves are taken with consistent orientation they represent elements of the group $H_1(F)$ whose product (sum, in additive notation) is trivial. (This in fact is the original idea behind homology, which set out to handle oriented curves in X, adding two curves by taking their union, and declaring a curve to be equivalent to zero (null-homologous) if it bounded an orientable surface lying in X.)

6.2 Surfaces lying in $\mathbf{R}^3$

We shall now consider orientable surfaces lying in 3-dimensional space. The boundary of any such surface is then a knot or link with a consistently chosen orientation. We shall aim to study properties of the boundary curve by means of features of a 'spanning surface'. One point which must now be established is that we can always find a spanning surface for a given oriented link.

First let us apply the observation above about the boundary curves in a surface in the case when the surface lies in $\mathbf{R}^3$.

Corollary 6.4 *Suppose that L is a curve in the complement of a knot K, and that $F \subset \mathbf{R}^3 - K$ is an orientable surface with boundary L. Then $lk(L,K) = 0$.*

More generally, suppose that $L_1, \ldots, L_r$ are curves in the complement of a knot K, and that $F \subset \mathbf{R}^3 - K$ is an orientable surface with boundary $L_1 \cup \ldots \cup L_r$. Then $\sum lk(L_j, K) = 0$ when the curves L_j are oriented consistently with an orientation of F.

Proof : The curve L_j in F determines an element $[L_j] \in H_1(F)$, and $\prod [L_j] = 1$ in $H_1(F)$, when the curves are consistently oriented.

Consider the homomorphism $i_* : H_1(F) \to H_1(\mathbf{R}^3 - K)$ induced by the inclusion $i : F \to \mathbf{R}^3 - K$. Now $H_1(\mathbf{R}^3 - K)$ is infinite cyclic, with generator t, and the curve L_j represents $t^{\mathrm{lk}(L_j,K)}$ in this group. Thus $i_*[L_j] = t^{\mathrm{lk}(L_j,K)}$.

Because i_* is a group homomorphism we then have $i_*(\prod [L_j]) = t^{\sum \mathrm{lk}(L_j,K)}$. But $i_*(\prod [L_j]) = i_*(1) = 1 = t^0$, and so $\sum \mathrm{lk}(L_j, K) = 0$. □

6.2.1 Seifert's construction

We now show how to construct an orientable surface $F \subset \mathbf{R}^3$ spanning any given knot K. The same construction, applied to an oriented link L will give a surface whose oriented boundary is L.

Construction. Start with an oriented diagram for K. In place of each crossing join the overpass to the underpass by two arcs respecting the orientation so as to get a number of oriented closed curves which form a new diagram without any crossings. Arrange these closed curves, $C_1, \ldots, C_k$ say, to lie in planes at different levels in $\mathbf{R}^3$. If one curve lies inside another in the diagram then place it at a higher level in $\mathbf{R}^3$. Each curve C_j bounds a disc D_j in its plane. Choose the orientation of D_j consistently with that of C_j, (this simply involves using projection to a fixed copy of $\mathbf{R}^2$ in the case where C_j is oriented anticlockwise, or projection followed by a reflection in the other case, to formally identify D_j as an oriented surface.) Finally connect the discs by a twisted rectangle R_i at each crossing, in which two edges of R_i form the part of K which was removed, while the other two edges are joined to the discs. The result is a surface F whose boundary is K. The orientation on R_i can be chosen so that the edge orientations cancel on each of the two adjoining discs, so that the complete surface F is oriented, and determines the chosen orientation on the boundary K.

The surface F has one boundary curve, so $F \cong F_{g,1}$ for some g, called the *genus* of F. We can find g easily from a calculation of χ for F, since $\chi = 1 - 2g$. Now let us calculate $\chi(F)$ by adding the contributions from D_j and R_i separately. To avoid counting edges and vertices twice, we shall count common edges and vertices only as the contribution from the discs D_j, so that each rectangle will contribute two edges only, and no vertices, to the sum. There is then a net contribution of $1 - 2 + 0 = -1$ for each R_i, and a contribution of 1 from each D_j, since there are the same number of edges and vertices on the boundary of a disc. The result is then

$$\chi(F) = k - n$$

where there are n crossings in the diagram, and k 'Seifert circles', (the closed curves C_j constructed by cutting out the crossings).

This construction shows that every knot K has an orientable spanning surface; I shall call a spanning surface which arises from this construction on some diagram a *projection surface* for K.

Definition. The *genus* g_K of K is the minimum genus $g(F)$ among all orientable surfaces F which span K.

Although the genus of the projection surface for a diagram of K depends on the diagram, the genus of K is an invariant of K, because any homeomorphism carrying K to an equivalent knot K' will carry a spanning surface of K to one of K'.

Remark. If K is spanned by a surface $F \cong D^2$ then we can use the homeomorphism from F to D^2 to see how K could be moved through F to lie in a small, essentially planar, part of F. This isotopy of K within F can be extended to $\mathbf{R}^3$ to give an equivalence of K with a curve bounding a disc in a plane, so that

$$\begin{aligned} K \text{ is unknotted} \quad &\iff \quad K \text{ is spanned by a disc} \\ &\iff \quad g_K = 0. \end{aligned}$$

It is not always easy to find the genus for a given knot. Seifert's construction certainly will give an upper bound, just by calculating the genus of some projection surface, but this may not be the minimum, and there are even cases where no projection surface has the minimum genus. We shall shortly find a readily calculated lower bound for the genus.

The trefoil has genus ≥ 1, since it is known to be knotted. The simplest projection surface has genus 1, so we can be sure that the trefoil has genus 1.

6.2.2 General spanning surfaces

We shall restrict our attention for the moment to the case of knots, and consider a knot K with an orientable spanning surface F lying in $\mathbf{R}^3$. If F has genus g then it is homeomorphic to the standard surface $F_{g,1}$, which we regard as made up of $2g$ ribbons surrounding embedded curves $c_1, \ldots, c_{2g}$. Let $h : F_{g,1} \to F$ be a homeomorphism, and consider the corresponding curves $h(c_1), \ldots, h(c_{2g})$ in F. These curves may well be knotted and intertwined in $\mathbf{R}^3$, but within F they lie just like $c_1, \ldots, c_{2g}$ do in $F_{g,1}$. The surface F, with boundary K, then consists of $2g$ (possibly rather uneven) ribbons around these curves. It is possible to move its boundary K within F so as to lie as close as desired to the curves $h(c_1), \ldots, h(c_{2g})$, using F as a guide to the movement. In the course of moving K within F we can ensure that K is never moved through itself in $\mathbf{R}^3$, and so the final curve is a knot equivalent to the original curve K, which bounds a surface consisting of ribbons close to the curves $h(c_1), \ldots, h(c_{2g})$, and a connecting set of arcs, as shown in figure 17

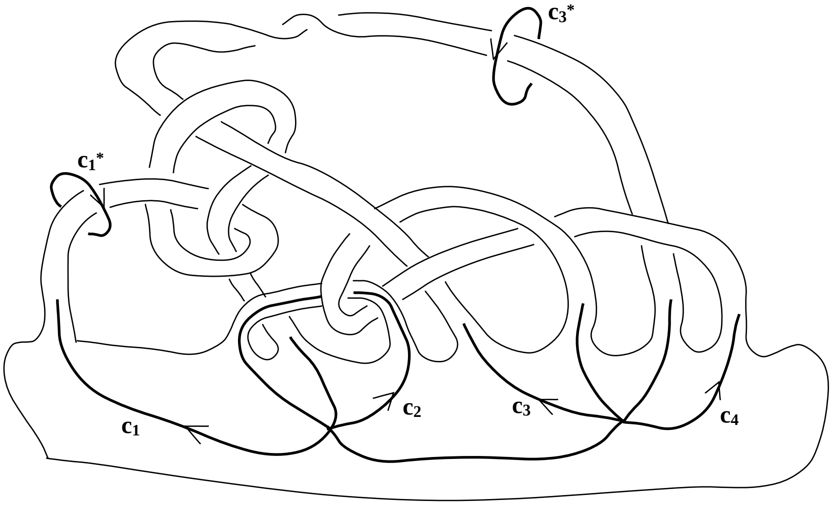


Figure 17:

We can think of the neighbourhood of the arcs as a disc, to which $2g$ ribbons have been attached, remembering that the individual ribbons may well be twisted and knotted and interlinked with the others as they lie in $\mathbf{R}^3$. To save notation in what follows, write c_i in place of $h(c_i)$ for the curves in the surface F.

The complement $\mathbf{R}^3 - F$ is essentially the same as $\mathbf{R}^3-$core curves and arcs, certainly as regards its fundamental group. We can use a similar construction to the Wirtinger presentation to find a presentation for this fundamental group, in terms of loops around pieces of the core curves, and consequently we can find $H_1(\mathbf{R}^3 - F)$. It turns out that $H_1(\mathbf{R}^3 - F)$ is free abelian of rank $2g$, generated by 'meridians' $c_1^*, \ldots, c_{2g}^*$, where c_i^* encircles the i-th ribbon. The main feature of this presentation which we shall use later is that $\text{lk}(c_i, c_j^*) = \delta_{ij}$. In what follows we shall not need to view F explicitly as ribbons, nor find curves c_i exactly as here, but we shall use the existence of the generating systems $c_1, \ldots, c_{2g}$ for $H_1(F)$ and $c_1^*, \ldots, c_{2g}^*$ for $H_1(\mathbf{R}^3 - F)$, with their linking properties.

Let c^* be any curve in $\mathbf{R}^3$ which does not meet F. Any curve x in F will have linking number $\mathrm{lk}(x, c^*)$ determined by its image, when regarded as an element of $H_1(F)$, under the group homomorphism $i_* : H_1(F) \to H_1(\mathbf{R}^3 - c^*)$. Suppose then that $x = x_1c_1 + \ldots + x_{2g}c_{2g}$ in $H_1(F)$, written additively, for some $x_i \in \mathbf{Z}$. Then $\mathrm{lk}(x, c^*) = \sum x_i \,\mathrm{lk}(c_i, c^*)$, so that

$$\mathrm{lk}(x, c_j^*) = x_j.$$

It follows that we can find the coefficients x_j, and hence the element of $H_1(F)$ which the curve x represents once we know the linking numbers of x with each of the curves c_j^*. Similarly, for $c^* = \sum y_j c_j^*$ in $H_1(\mathbf{R}^3 - F)$ we have $\mathrm{lk}(c_i, c^*) = y_i$, and again the element $c^* \in H_1(\mathbf{R}^3 - F)$ is determined by the linking numbers of c^* with the basis elements c_i.

Aside. We may note that the linking number gives a bilinear pairing

$$\mathrm{lk} : H_1(F) \times H_1(\mathbf{R}^3 - F) \to \mathbf{Z}$$

taking a pair (x, c^*) to $\mathrm{lk}(x, c^*)$. The bases chosen for $H_1(F)$ and $H_1(\mathbf{R}^3 - F)$ are dual bases with respect to this pairing, and the pairing is non-degenerate.

We shall not always use the basis above in describing elements of $H_1(F)$ and $H_1(\mathbf{R}^3 - F)$, but its existence guarantees certain results which we shall call on later. Firstly, both groups can be written additively as $\mathbf{Z}^{2g}$, and any other choice of basis $\{a_i\}$ for $H_1(F)$ will determine a 'dual' basis $\{a_i^*\}$ of $H_1(\mathbf{R}^3 - F)$ with respect to the linking pairing, that is $\mathrm{lk}(a_i, a_j^*) = \delta_{ij}$. For, given the basis $\{a_i\}$ we may write $c_j = \sum p_{ij} a_i$ for some invertible integer matrix $P = (p_{ij}) \in GL(2g, \mathbf{Z})$. Then set $a_j^* = \sum p_{ji} c_i^*$ to define a new basis $\{a_j^*\}$ for $H_1(\mathbf{R}^3 - F)$. Now a little linear algebra shows that for $c = \sum x_i a_i = \sum y_j c_j \in H_1(F)$ and $d = \sum x_i^* a_i^* = \sum y_j^* c_j^* \in H_1(\mathbf{R}^3 - F)$ we can think of c as having coordinate vectors $\mathbf{x}$ and $\mathbf{y}$ related by $\mathbf{x} = P\mathbf{y}$, while for d the coordinate vectors $\mathbf{x}^*$ and $\mathbf{y}^*$ in the two bases are related by $\mathbf{y}^* = P^T\mathbf{x}^*$.

Then $\mathrm{lk}(c, d) = (\mathbf{y}^*)^T\mathbf{y} = (\mathbf{x}^*)^T P\mathbf{y} = (\mathbf{x}^*)^T\mathbf{x}$. Take $c = a_i$ and $d = a_j^*$ to confirm that $\mathrm{lk}(a_i, a_j^*) = \delta_{ij}$.

7 The Conway polynomial

In this section we shall show how to define the Conway polynomial of a knot, starting from a spanning surface.

7.1 Seifert matrices

Take an orientable surface F spanning a knot K. At any interior point of a surface F in $\mathbf{R}^3$ there are locally two 'sides' to the surface, for a small enough ball around the point is separated into two components by the surface. For an orientable surface we can globally distinguish these two sides as 'top' and 'bottom', by saying that a line crossing F transversely passes from bottom to top if the curve in F going round its intersection point in the right handed sense has come from an anticlockwise curve in $\mathbf{R}^2$ under the homeomorphism from the appropriate disc D_i. The choice of consistent homeomorphisms dictates which side will be which, with the alternative choice interchanging top and bottom, and at the same time reversing the orientation of all the boundary curves of F. There is then a map $i^+ : F \to \mathbf{R}^3 - F$ defined by pushing F in the direction of the positive normal, i.e. the normal which crosses the surface from bottom to top.

Choose embedded curves $a_1, \ldots, a_{2g}$ in F forming a basis for $H_1(F)$. Translate them by i^+ into $\mathbf{R}^3 - F$ to give curves $a_1^+ = i^+(a_1), \ldots, a_{2g}^+$ in $\mathbf{R}^3 - F$.

Definition. The $2g \times 2g$ integer matrix $A = (a_{ij})$, where $a_{ij} = \mathrm{lk}(a_i, a_j^+)$, a *Seifert matrix* for the surface F.

This matrix can be thought of as the matrix of the linear map $i_*^+ : H_1(F) \to H_1(\mathbf{R}^3 - F)$ with respect to the basis $\{a_i\}$ of $H_1(F)$ and the dual basis $\{a_i^*\}$ of $H_1(\mathbf{R}^3 - F)$, where both groups are written additively as $\mathbf{Z}^{2g}$, since the j-th column of A consists of the coordinates, in the basis $\{a_i^*\}$, of the image a_j^+ of the j-th basis element.

A different choice of basis for $H_1(F)$, to a basis $\{b_j\}$ say, will lead to a Seifert matrix $B = Q^T A Q$, where the matrix Q is the invertible integer matrix (with integer inverse), which relates the two bases by $b_j = \sum q_{ij} a_i$. Note that $\det Q = \pm 1$, since Q^{-1} must also have integer entries.

We could interpret A as the matrix of the bilinear form $H_1(F) \times H_1(F) \to \mathbf{Z}$ pairing (x, y) to $\mathrm{lk}(x, y^+)$. We can translate the curves x and y^+ along the normal to F so that x is moved off to the negative side, to become x^- while

y^+ returns to lie on F as the curve y. The link $x \cup y^+$ is then equivalent to the link $x^- \cup y$, so that $\mathrm{lk}(x, y^+) = \mathrm{lk}(x^-, y) = \mathrm{lk}(y, x^-)$ by symmetry of linking number. It follows that $a_{ji} = \mathrm{lk}(a_i, a_j^-)$ so that the matrix A^T represents $i^- : H_1(F) \to H_1(\mathbf{R}^3 - F)$ in the same bases as for i^+.

7.2 Polynomials

Notation. For any $n \times n$ matrix A, set $G(s, u) = \det(sA + uA^T)$.

Take a Seifert matrix A for the surface F, with some choice of basis for $H_1(F)$.

Theorem 7.1 *The polynomial $G(s, u)$ defined from a Seifert matrix for the surface F is independent of the choice of basis of $H_1(F)$.*

Proof : For another choice of basis we have matrix $B = Q^T AQ$ and then

$$\det(sB + uB^T) = \det Q^T(sA + uA^T)Q = (\det Q)^2 G(s, u) = G(s, u).$$

□

Notation. Write $G(s, -s^{-1}) = F(s)$.

Theorem 7.2 *We can write $F(s) = \nabla(z)$ as a polynomial in $z = s - s^{-1}$.*

Proof : We have

$$G(u, s) = \det(uA + sA^T) = \det(uA + sA^T)^T = G(s, u).$$

Now G is a homogeneous polynomial of degree n in s and u which is symmetric. Hence it can be rewritten as a polynomial $D(z, v)$, say, in the elementary symmetric functions $z = s + u$ and $v = su$. Put $u = -s^{-1}$ to get $v = -1$, $z = s - s^{-1}$ and $G(s, -s^{-1}) = F(s) = \nabla(z)$. □

Clearly $F(s^{-1}) = (-1)^n F(s)$. Now $F(s^{-1}) = \nabla(s^{-1} - s) = \nabla(-z)$. So

$$\nabla(-z) = \begin{cases} \nabla(z), & n \text{ even} \\ -\nabla(z), & n \text{ odd} \end{cases}$$

and it is thus either an even or an odd polynomial in z depending on n.

Where K is a knot then $n = 2g$ and so $\nabla(z)$ is a polynomial in z^2.

Theorem 7.3 *$\nabla(z)$ depends only on K and not on the chosen spanning surface F.*

The proof will be given later, in a slightly different form, but a version, depending on a further assumption about spanning surfaces, follows shortly.

Definition. The polynomial $\nabla(z)$ is known as the *Conway polynomial* of the knot K. A similar definition can be made for any oriented link.

Definition. The *Alexander polynomial* is closely related, and is basically $F(s)$ multiplied by a large enough power of s to clear it of negative powers of s, and written in terms of $t = s^2$.

Since $F(s) = s^{-n}\det(s^2A - A^T)$ we have, up to a power of t, that the Alexander polynomial $\Delta(t)$ is $\det(tA - A^T)$.

7.3 Properties of Conway and Alexander polynomials

Theorem 7.4 *For a knot K, deg$(\nabla(z)) \leq 2g_K$.*

Proof : For an $n \times n$ matrix A, the polynomial $G(s, u) = \det(sA + uA^T)$ is homogeneous of degree n in s and u, and hence has degree $\leq n$ in $z = s + u$ when rewritten in terms of z and $v = su$. Where the knot K has a spanning surface of genus g the Seifert matrix A is a $2g \times 2g$ matrix, so that $\nabla(z) = G(s, -s^{-1})$ has degree $\leq 2g$ in z. This is true for all spanning surfaces, since $\nabla(z)$ does not depend on the surface chosen, and hence $\deg(\nabla(z)) \leq 2g_K$ where the genus g_K is the minimum genus over all spanning surfaces. □

Suppose that $\deg\nabla(z) = k$. Then k is even for a knot, since $\nabla(z)$ is an even function. Then $s^k\nabla(s - s^{-1})$ is a polynomial in $s^2 = t$, which we write as $\Delta(t)$, the Alexander polynomial of K. The Alexander polynomial also has degree k, which is again a lower bound for $2g_K$.

Theorem 7.5 *The Alexander polynomial for a knot has the form*

$$\Delta(t) = c_0 + c_1 t + \ldots + c_k t^k, \text{ with } c_i = c_{k-i}.$$

Proof : For a knot, $\nabla(-z) = \nabla(z)$, so

$$\begin{aligned} \Delta(t^{-1}) = s^{-k}\nabla(s^{-1} - s) &= s^{-k}\nabla(s - s^{-1}) \\ &= s^{-2k}\Delta(t) \\ &= t^{-k}\Delta(t). \end{aligned}$$

Then $\Delta(t) = \sum c_i t^i = t^k \Delta(t^{-1}) = \sum c_i t^{k-i}$. □

Theorem 7.6 *For a knot K, $\nabla(0) = 1$.*

Proof : The constant term $\nabla(0) = G(1, -1)$ can be calculated from any choice of basis for $H_1(F)$, where F is a spanning surface and A is the resulting Seifert matrix, as $\nabla(0) = \det(A - A^T)$.

Take as basis the family of curves $c_1, \ldots, c_{2g}$, with the property that c_{2r-1} and c_{2r} only meet in one point, and otherwise pairs of curves don't meet at all. Write $\mathrm{D} = A - A^T$. Then $d_{ij} = \mathrm{lk}(c_i, c_j^+) - \mathrm{lk}(c_j, c_i^+)$.

If c_i and c_j don't meet, then the link $c_i \cup c_j$ is equivalent to the links $c_i \cup c_j^+$ and $c_i^+ \cup c_j$. Then

$$\mathrm{lk}(c_i, c_j^+) = \mathrm{lk}(c_i, c_j) = \mathrm{lk}(c_j, c_i^+)$$

and so $d_{ij} = 0$.

Suppose now that c_i and c_j cross at one point in F, as in the remaining cases with $i = 2r - 1, j = 2r$. We can view $c_i \cup c_j$ so that the intersection point is seen from the top of the surface F, while otherwise we assume that we see a finite number of crossings of c_i with c_j. We may then view the link $c_i \cup c_{j^+}$ as having the same diagram as $c_i \cup c_j$ except at the intersection point, where c_j^+ passes above c_i. The link $c_i^+ \cup c_j$ has the same diagram, except this time c_i^+ passes above c_j at the intersection point. Consequently, the difference in linking number,

$$\begin{aligned} d_{ij} &= \mathrm{lk}(c_i, c_j^+) - \mathrm{lk}(c_j, c_i^+) \\ &= \mathrm{lk}(c_i, c_j^+) - \mathrm{lk}(c_i^+, c_j) \\ &= \pm 1, \end{aligned}$$

depending on the sign of the crossing at the intersection point, which features in counting crossings of c_j under c_i in $\mathrm{lk}(c_i^+, c_j)$ but not in $\mathrm{lk}(c_i, c_j^+)$. Since $d_{ji} = -d_{ij}$ the matrix D then has the form

$$D = \pm \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \oplus \pm \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \oplus \ldots$$

giving $\det D = 1$. □

Example. The trefoil and figure eight knots have $\nabla(z) = 1 + z^2$ and $\nabla(z) = 1 - z^2$ respectively. The only possible polynomials for knots of genus 1 are $\nabla(z) = 1 + nz^2$ for some $n \in \mathbf{Z}$, and all of these occur.

Seifert showed that all even polynomials $\nabla(z)$ with $\nabla(0) = 1$ occur as the polynomial of some knot. It is possible that the genus is strictly larger than $\frac{1}{2}\mathrm{deg}\nabla$ and indeed there are very many non-trivial knots with $\nabla(z) = 1$.

To prove that $\nabla(z)$ is independent of the choice of spanning surface it is possible to use the result of Trotter on 'S-equivalence' of spanning surfaces. This relies on the notion of modifying a spanning surface by 'adding a hollow handle', where F is modified by embedding a copy of $D^2 \times I$ in $S^3 - K$ and avoiding F, except that the two ends $D^2 \times \{0\}$ and $D^2 \times \{1\}$ meet F in two disjoint discs. A new spanning surface F' is formed by deleting the discs and replacing them by the cylinder $S^1 \times I$.

Two surfaces in $\mathbf{R}^3$ which can be changed from one to the other by a sequence of moves either adding or removing hollow handles are called S-equivalent. Trotter shows that any two orientable surfaces with the same boundary are S-equivalent.

It is then enough to show that $\nabla(z)$ is unaltered when the spanning surface is changed by adding a hollow handle.

Start from a spanning surface F, and add a hollow handle to give a surface F'. Choose generators for $H_1(F')$ to consist of $2g$ generators for $H_1(F)$ together with two extra generators c and d, where d is the meridian curve $S^1 \times \{\frac{1}{2}\}$ lying entirely on the hollow handle and c runs across the handle, meeting d in just one point. Then d spans a disc which meets none of the generators of $H_1(F)$, so that $\mathrm{lk}(a_i, c^+) = \mathrm{lk}(a_i, d^+) = \mathrm{lk}(d, a_i^+) = 0$, while $\mathrm{lk}(d, d^+) = 0$.

We also have $\mathrm{lk}(c, d^+) = \pm 1$ while $\mathrm{lk}(d, c^+) = 0$. The Seifert matrix for F' then has the form

$$\begin{pmatrix} A & \mathbf{v} & \mathbf{0} \\ \mathbf{w^T} & k & \pm 1 \\ \mathbf{0^T} & 0 & 0 \end{pmatrix}.$$

We have

$$\begin{aligned} G'(s,u) &= \det \begin{pmatrix} sA + uA^T & s\mathbf{v} + u\mathbf{w} & \mathbf{0} \\ s\mathbf{w^T} + u\mathbf{v^T} & (s+u)k & \pm s \\ \mathbf{0^T} & \pm u & 0 \end{pmatrix} \\ &= -su \det(sA + uA^T) \\ &= -su \, G(s,u), \end{aligned}$$

on expanding the determinant by the last row and then by the last column.

Putting $u = -s^{-1}$ gives $G'(s, -s^{-1}) = G(s, -s^{-1})$ as required.

Remark. I am grateful to Pedro Manchon for pointing out that my definition of a hollow handle in the original version of these notes was too restrictive.

7.4 Seifert matrices again

When looking for a basis of curves for the homology of an explicit spanning surface F for a knot it is useful not to have to manipulate the surface into any sort of standard form. The following sufficient condition is then helpful in ensuring that the curves chosen can be used in constructing a Seifert matrix.

Proposition 7.7 *Let F be a surface in $\mathbf{R}^3$ of genus g, and one boundary component. If we find curves $a_1, \ldots, a_{2g}$ in F and $a_1^*, \ldots, a_{2g}^*$ in $\mathbf{R}^3 - F$ with $lk(a_i, a_j^*) = \delta_{ij}$ then $a_1, \ldots, a_{2g}$ forms a basis for $H_1(F)$.*

Proof : We know that there exists a basis $c_1, \ldots, c_{2g}$ of $H_1(F)$ with a dual basis $c_1^*, \ldots, c_{2g}^*$ of $H_1(\mathbf{R}^3 - F)$. We may then write the elements a_i and a_j^* in terms of these bases, as $a_i = \sum p_{ik} c_k$ and $a_j^* = \sum q_{j\ell} c_\ell^*$, for some integer matrices $P = (p_{ik})$ and $Q = (q_{j\ell})$. Then

$$\delta_{ij} = \mathrm{lk}(a_i, a_j^*) = \sum p_{ik} q_{j\ell} \mathrm{lk}(c_k, c_\ell^*) = \sum p_{ik} q_{j\ell} \delta_{k\ell} = \sum p_{ik} q_{jk},$$

so that $I = PQ^T$. It follows that $\det P \det Q = 1$, so $\det P = \det Q = \pm 1$, and P has an *integer* inverse. We can then express the elements $\{c_k\}$ as integer combinations of $\{a_i\}$, showing that $\{a_i\}$ also form a basis for $H_1(F)$. □

We may choose curves $a_1, \ldots, a_{2g}$ for a projection surface so that for each a_i there is at least one 'crossing rectangle' through which a_i passes once, while no other a_j pass through at all. Then we can choose a curve a_i^* around this rectangle in $\mathbf{R}^3 - F$ to satisfy the linking conditions above, and ensure that $a_1, \ldots, a_{2g}$ chosen in this way forms a basis for $H_1(F)$.

Example. In the knot shown in figure 18, and more generally in figure 19, where there are $2m$ and $2k$ half-twists as indicated, the projection surface has genus $g = 1$, (from a calculation of χ), so that we need to find two curves a_1 and a_2. These can be chosen so that one passes along the ribbon with m twists, and the other along the ribbon with k twists, meeting in just one point in the surface F.

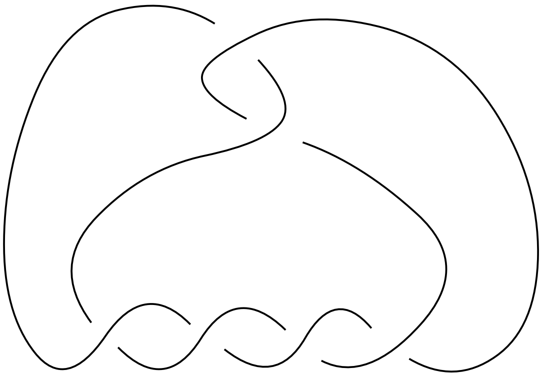


Figure 18:

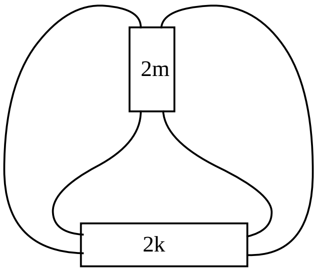


Figure 19:

The entries $a_{ij} = \mathrm{lk}(a_i, a_j^+)$ in the resulting Seifert matrix can then be found; it is enough in general, for $i \neq j$, to move a_j away from F only near the intersection points of a_i with a_j, to give two disjoint curves. The surface F can be ignored after this in calculating their linking number. In the case $i = j$ note that the curve a_i^+ can be moved down, without meeting a_i, so as to lie on the surface F just to one side of a_i, giving a pair of curves in F which are equivalent to the pair $a_i \cup a_i^+$. It is easy to draw these curves alone, without the rest of the surface F, noting simply that as a_i passes through a rectangle then any curve lying beside it must follow the twist in the rectangle. The diagonal entry a_{ii} is then the linking number of the curve a_i with a neighbouring parallel curve in F.

In the example of figure 19 we have $a_{12} = 0, a_{21} = 1, a_{11} = m, a_{22} = k$, so that

$$sA + uA^T = \begin{pmatrix} mz & u \\ s & kz \end{pmatrix}$$

and $\nabla(z) = 1 + mkz^2$. This Conway polynomial will then not serve to distinguish between the cases $m = 6, k = 1$ and $m = 3, k = 2$, for example.

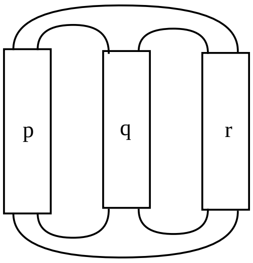


Figure 20:

The knots in figure 20 with p, q, r all odd (positive or negative) also have a projection surface of genus 1, and $\nabla(z) = 1 + f(p, q, r)z^2$. It is possible to choose p, q, r so that $f(p, q, r) = 0$, while avoiding knots which are obviously trivial. Try $(p, q, r) = (-3, 5, 7)$.

In fact the knots with $|p|, |q|, |r| > 1$ can be shown to be non-trivial by exhibiting a homomorphism from the knot group onto a non-abelian group of 3×3 matrices.

7.5 Connected sums

The operation of tying two knots one after the other on a piece of rope can be realised for closed curves as follows. Take two knots K_1 and K_2 lying in disjoint half-spaces, separated by a plane $\mathbf{R}^2$. Draw an arc from a point on K_1 to a point on K_2 which crosses the separating plane $\mathbf{R}^2$ just once. Then break K_1 and K_2 at the ends of the arc, and join them by two edges of a ribbon which follows the chosen arc. The resulting curve K is called the *connected sum* of K_1 and K_2, written $K = K_1 + K_2$. It can be shown that K is independent, up to equivalence, of the choice of arc used in the construction.

7.5.1 Conway polynomial for a connected sum

Surfaces F_1 and F_2 spanning K_1 and K_2 in their separate half-spaces can be joined by the ribbon around an arc which avoids the surfaces to give a surface $F = F_1 + F_2$ spanning their connected sum $K = K_1 + K_2$. Then, by calculating the Euler characteristics, we have

$$\text{genus}(F) = \text{genus}(F_1) + \text{genus}(F_2).$$

A basis for $H_1(F)$ can be taken as $\{a_i^{(1)}\} \cup \{a_j^{(2)}\}$, where $\{a_i^{(1)}\}$ lie in F_1 and form a basis for $H_1(F_1)$ while $\{a_j^{(2)}\}$ similarly form a basis for $H_1(F_2)$. The curves $a_i^{(1)}$ and $a_j^{(2)}$ lie in separate half-spaces, so they are disjoint, and have linking number 0. The Seifert matrix arising from this choice of basis then has the form $A = A_1 \oplus A_2$, where A_1 and A_2 are Seifert matrices for F_1 and F_2. Hence

$$\nabla_K(z) = \det(sA+uA^T) = \det((sA_1+uA_1^T)\oplus(sA_2+uA_2^T)) = \nabla_{K_1}(z)\nabla_{K_2}(z).$$

7.5.2 The genus of a connected sum of knots

It is clear from the construction that there exists a spanning surface for $K = K_1 + K_2$ of genus $g_{K_1} + g_{K_2}$, so that $g_K \leq g_{K_1} + g_{K_2}$.

If K_1 and K_2 satisfy $g_{K_1} = \frac{1}{2}\text{deg}\nabla_{K_1}$ and $g_{K_2} = \frac{1}{2}\text{deg}\nabla_{K_2}$ then the inequality $g_K \geq \frac{1}{2}\text{deg}\nabla_K$ shows immediately that $g_K = g_{K_1} + g_{K_2}$.

It is possible to prove this in general, by consideration of how a spanning surface F' for K of minimal genus g_K meets the separating plane. This general result shows in particular that if $K_1 + K_2$ is unknotted (so that $g_K = 0$) then K_1 and K_2 must both be unknotted.

It can also be shown that connected sum of knots is an associative and commutative operation, so for example $K_1 + K_2 + K_3 = K_1 + K_3 + K_2$. Geometric arguments can be used to show further that every polygonal knot K has a decomposition, unique up to order of factors, as $K = K_1 + \cdots + K_r$, where the factors are non-trivial, and cannot themselves be decomposed further.

7.6 Links and the Conway polynomial

We may extend the calculation of Conway polynomials to cover the case of *oriented* links. We can again choose an oriented spanning surface F for a link L, to induce the chosen orientation on its boundary components. Note that a spanning surface will have to be quite different if the orientation of one boundary component is reversed.

Using a connected spanning surface F we may then find an $n \times n$ Seifert matrix A as before, with $n = 2g + r - 1$ when L has r components. Then take $\nabla(z) = \det(sA + uA^T)$, as a polynomial in z, even or odd, depending on the parity of n, and hence on the number of components of L. If L can be spanned by a union of several disjoint surfaces these may be connected by tubes to give a connected spanning surface.

Note that if $L = L_1 \cup L_2$ with $\mathrm{lk}(L_1, L_2) = 0$ then it is always possible to modify a surface spanning L_1 to give one which avoids L_2 by running tubes around L_2 to cut out pairs of punctures of the surface by L_2 with opposite signs. However it is not in general possible to do this simultaneously for L_1 and L_2 so that the two surfaces are disjoint.

Example. The link shown in figure 21 can be spanned by a ribbon F with $\mathrm{rank}\, H_1(F) = 1$, and $H_1(F)$ can be generated by the single core curve d_1 having $\mathrm{lk}(d_1, d_1^+) = m$. This gives $\nabla(z) = mz$.

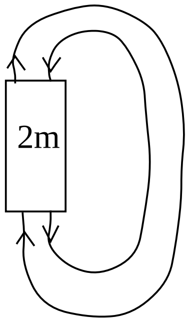


Figure 21:

In general, a spanning surface F for a 2-component link $L = L_1 \cup L_2$ has a basis $c_1, \ldots, c_{2g}, d_1$ for $H_1(F)$, where the curve d_1 lies parallel to the component L_1. We can then calculate the diagonal element $\mathrm{lk}(d_1, d_1^+)$ in the resulting Seifert matrix A as follows. The curve d_1^+ can be replaced by a parallel curve to d_1 in F, so for the purposes of calculating the linking number we may consider simply d_1 and the boundary curve L_1 itself. The surface F contains a ribbon parallel to L_1, whose other edge consists of the curve d_1. Remove this ribbon from F to get an orientable surface F' in $\mathbf{R}^3 - L_1$

whose oriented boundary consists of d_1 and L_2 with their given orientation. Then $d_1 + L_2 = 0$ as elements of $H_1(F')$ and hence of $H_1(\mathbf{R}^3 - L_1)$, so that $\mathrm{lk}(L_1, d_1) = -\mathrm{lk}(L_1, L_2)$, and thus $\mathrm{lk}(d_1, d_1^+) = -\mathrm{lk}(L_1, L_2) = m$ say. Now

$$\mathrm{lk}(c_i, d_1^+) = \mathrm{lk}(c_i, d_1) = \mathrm{lk}(c_i^+, d_1),$$

since c_i and d_1 are disjoint in F, so the Seifert matrix then has the form $A = \begin{pmatrix} B & \mathbf{v} \\ \mathbf{v}^T & m \end{pmatrix}$. Then

$$sA + uA^T = \begin{pmatrix} sB + uB^T & z\mathbf{v} \\ z\mathbf{v}^T & zm \end{pmatrix},$$

and

$$\nabla(z) = \det(sA + uA^T) = z \det \begin{pmatrix} sB + uB^T & \mathbf{v} \\ z\mathbf{v}^T & m \end{pmatrix}.$$

We can find the coefficient of z in $\nabla(z)$ by putting $s = 1, u = -1, z = 0$ in the determinant above, to get

$$\det \begin{pmatrix} B - B^T & \mathbf{v} \\ \mathbf{0} & m \end{pmatrix} = m \det(B - B^T).$$

As in the case of knots, with a standard choice of curves $c_1, \dots, c_{2g}$ we have

$$B - B^T = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \oplus \cdots \oplus \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}.$$

Then $\det(B - B^T) = 1$, and $\nabla(z) = mz+$ higher terms, with $m = -\mathrm{lk}(L_1, L_2)$.

Example. For the unlink with two (or more) components we have $\nabla(z) = 0$.

Remark. A similar analysis gives $\nabla(z) = mz^{r-1} +$ higher terms for a link with r components, using a basis $c_1, \dots, c_{2g}, d_1, \dots, d_{r-1}$, with curves $d_1, \dots, d_{r-1}$ parallel to the boundary curves $L_1, \dots, L_{r-1}$ of L. The coefficient m can be shown in this case to depend simply on the collection of linking numbers of the r components. In fact it can be written as a polynomial in the linking numbers of degree $r - 1$.

7.7 Skein relations

Let $L_\pm$ and L_0 be oriented links with diagrams which differ only as shown in the neighbourhood of one crossing.

Theorem 7.8 (Conway) *The polynomials of $L_\pm$ and L_0 satisfy the linear (skein) relation*

$$\nabla_{L_+}(z) - \nabla_{L_-}(z) = -z\nabla_{L_0}(z).$$

Proof : The diagrams for the three links have the same set of Seifert circles. Construct spanning surfaces $F_\pm$ for $L_\pm$ from the projection surface F_0 for L_0 by adding one extra twisted rectangle. Having chosen a basis of curves for $H_1(F_0)$ we just need one further curve for $H_1(F_\pm)$ which we can choose to go once across the extra rectangle. This curve and its parallel in F_+ will have an apparent negative crossing as they cross through the rectangle, while the corresponding curves in F_- will have a positive crossing, but elsewhere in $F_\pm$ they will look the same. If they have linking number k in F_+ then they will have linking number $k+1$ in F_-. The other linking numbers in the Seifert matrices for $F_\pm$ are the same in each case, so the Seifert matrices have the form

$$\begin{pmatrix} B & \mathbf{v} \\ \mathbf{w}^T & k \end{pmatrix}, \quad \begin{pmatrix} B & \mathbf{v} \\ \mathbf{w}^T & k+1 \end{pmatrix}.$$

Then

$$\nabla_{L_+} = \det\begin{pmatrix} sB + uB^T & s\mathbf{v} + u\mathbf{w} \\ s\mathbf{w}^T + u\mathbf{v}^T & zk \end{pmatrix}$$

and

$$\nabla_{L_-} = \det\begin{pmatrix} sB + uB^T & s\mathbf{v} + u\mathbf{w} \\ s\mathbf{w}^T + u\mathbf{v}^T & z(k+1) \end{pmatrix}.$$

Expand these determinants by the last row, and subtract, to get

$$\nabla_{L_+} - \nabla_{L_-} = -z\det(sB + uB^T) = -z\nabla_{L_0}.$$

□

Convention. It is usual in other contexts to replace z by $-z$ in $\nabla(z)$, so that the sign of ∇ is changed for links with an even number of components, and in a 2-component link the coefficient of z is then the linking number of the components. With this convention the skein relation becomes $\nabla_+ - \nabla_- = z\nabla_0$. For Alexander polynomials, with $s = \sqrt{t}$ this gives

$$\Delta_{L_+}(t) - \Delta_{L_-}(t) = (\sqrt{t} - 1/\sqrt{t})\Delta_{L_0}(t).$$

8 A $\mathbf{Z}[t,t^{-1}]$-module determined by the group of a knot

Every knot group G has infinite cyclic abelianisation, and thus there is a surjective homomorphism

$$\varphi : G \to C_\infty = < t : > .$$

Write K for the kernel $\ker\varphi = \{g \in G : \varphi(g) = 1\}$. In fact K is the commutator subgroup $G' \subset G$. Write $\Lambda = \mathbf{Z}[C_\infty] = \mathbf{Z}[t,t^{-1}]$ for the ring of integer linear combinations of elements of C_∞.

I shall describe how the abelianisation K/K' of K can be viewed as a module over the ring Λ. The main requirement is to define scalar multiplication $\lambda.u \in K/K'$ for $u \in K/K'$ and $\lambda \in C_\infty$, for example $t^2.u$ and $t^{-1}.u$, and to be sure that $t^2.u = t.(t.u)$, etc.

Starting with any $u \in K$ and $g \in G$ we may set

$$g.u := gug^{-1} \in K.$$

Notice that $gh.u = g.(h.u)$ for $g, h \in G$.

Suppose now that $\varphi(g) = \varphi(g')$. Then $g'g^{-1} \in K$ and

$$\begin{aligned} g'.u &= g'g^{-1}gug^{-1}(g'g^{-1})^{-1} \\ &= gug^{-1}(g'g^{-1})(g'g^{-1})^{-1} \bmod K' \\ &= g.u \bmod K', \end{aligned}$$

meaning that they give the same element when K is abelianised. Hence $g.u$, when regarded as an element of the abelianisation K/K', depends only on $\varphi(g) \in C_\infty$. We can thus define $t.u$ for any $u \in K$ by choosing $x \in G$ with $\varphi(x) = t$ and setting $t.u = xux^{-1}$. When regarded as an element of the abelian group K/K', (which we shall write additively), this definition for $t.u$ does not depend on the particular choice of x, by the calculations above. Equally, the element x^2ux^{-2} will then represent $t^2.u$ and $x^{-1}ux$ will represent $t^{-1}.u$, etc.

Example. The following element of K can then readily be written out as an element of the module K/K': $xux^{-1}x^3u^2x^{-3}u^{-1}$ represents $t.u + 2t^3.u - u = (t + 2t^3 - 1).u$, where scalar multiplication by a general scalar in Λ is defined to be an appropriate integer linear combination in the abelian group.

Definition.

The Λ-module defined in this way from the knot group is determined, up to isomorphism of Λ-modules, by the knot, since the group G, the map φ and K are all determined up to isomorphism. It is known as the *Alexander module* of the knot.

Theorem 8.1 *From a presentation for G and a surjective homomorphism*

$$\varphi : G \to < t : >$$

we can give a presentation for the Λ-module K/K', where $K = ker \varphi$.

Proof : Suppose that G can be presented as $G =< x_1, \ldots, x_n : r_1 = e, \ldots, r_m = e >$. Choose $x \in G$ with $\varphi(x) = t$, and write $u_j = x_j x^{-k_j}$, where $\varphi(x_j) = t^{k_j}$. Then $\varphi(u_j) = 1$ for each i, so that $u_j \in K$. G can then be generated by the elements $x, u_1, \ldots, u_n$, since $x_j = u_j x^{k_j}$. The relations can be rewritten to give m relations in these new generators, with one further relation coming from the expression of x in terms of $x_1, \ldots, x_n$.

Any word w in $x, u_1, \ldots, u_n$ can be written, by induction on its length, as a product of the form $x^k p$, where $\varphi(w) = t^k$ and p is a product of elements $x^r u_j^{\pm 1} x^{-r}$. Then K itself is generated by the elements $x^r u_j x^{-r}$, which represent $t^r u_j$ in the module K/K'. Thus K/K' is generated as a module by $u_1, \ldots, u_n$.

We can give the relations in K in terms of these generators, and thus the relations in K/K', as follows. For each relation $r_i = e$ write the word r_i as a word in the elements $x^r u_j x^{-r}$, which can be done since $\varphi(r_i) = 1$. These relations $r_i = e$, regarded as relations among the generators of K, together with their conjugates by powers of x, will provide sufficient relations to pass between words in K when viewed as words in the generators of K.

Read the word r_i in the generators of K in additive notation as a linear combination of the elements $t^r u_j$ in the abelianisation K/K' to get the relations for K/K'. Each word r_i will thus be rewritten in the form $\sum q_{ij} u_j$ for some $q_{ij} \in \Lambda$. The result will be a defining set of m relations $\sum q_{ij} u_j = 0$ for K/K'. and hence an $m \times n$ presentation matrix Q for the module, which can be used as shown below to define certain characteristic ideals $E_0, \ldots, E_n$ of the ring Λ, depending only on the module and not on Q. □

8.1 Module invariants

Certain features of a presentation matrix Q for a Λ-module are invariants of the module itself. Where the module has n generators w_j, and m relations $\sum q_{ij}w_j = 0$ these include, for each k, the ideal $E_k \subset \Lambda$ generated by all $(n-k)\times(n-k)$ minors (determinants of this sized submatrices) of the $m \times n$ matrix Q. (This just means all those elements of Λ which arise by taking Λ-linear combinations of the generators.)

Since every $(n-k)\times(n-k)$ minor is a Λ-linear combination of generators of E_{k+1} it follows that $E_0 \subset E_1 \subset \cdots \subset \Lambda$, with $\Lambda = E_k$ eventually. Where Λ is a unique factorisation domain, as in our case, then there is a greatest common divisor, d_k say, for the generators of E_k, which is determined up to multiplication by *units* (invertible elements) of the ring Λ. In the case where $\Lambda = \mathbf{Z}[t,t^{-1}]$ the invertible elements are simply $\pm t^r$.

The ideals referred to above are sometimes called the 'elementary ideals' of the module.

Theorem 8.2 *The elementary ideal E_0 of the Alexander module for a knot is a principal ideal, whose generator is the Alexander polynomial. Thus the Alexander polynomial can be found, up to a unit in Λ, from a presentation of the group of the knot.*

Example. The trefoil has group $G =< x, y : xyx = yxy >$. Find a presentation for $K/K' = G'/G''$ as a Λ-module.

The abelianisation homomorphism $\varphi : G \to< t : >$ is given by $\varphi(x) = \varphi(y) = t$. So set $u = yx^{-1}$. Then G is generated by x and u, with $y = ux$, so the relation becomes $xux^2 = ux^2ux$. Rewrite this as $ux^2ux^{-1}u^{-1}x^{-1} = e$ and write the LHS as a product

$$u(x^2ux^{-2})(xu^{-1}x^{-1})$$

of generators of G'.

This product becomes, in additive notation in G'/G'',

$$u + t^2u - tu,$$

so the module has generator u and relation

$$(1 - t + t^2)u = 0,$$

giving a 1×1 relation matrix with entry $1 - t + t^2$.

The E_0 ideal is then generated by this single entry $d_0 = 1 - t + t^2 \in \Lambda$, so the Alexander polynomial of the trefoil is, up to a unit, $\Delta(t) = 1 - t + t^2$.

8.2 Further examples

(1) A presentation of $G = \pi_1(\mathbf{R}^3 - K)$ can be given for the figure-eight knot K, by

$$G =< x_1, x_2 : x_2x_1x_2^{-1}x_1x_2x_1^{-1} = x_1x_2x_1^{-1}x_2 > .$$

Take $x = x_2$, $u_1 = x_1x^{-1}$ to present G' by u_1 and its conjugates by x, with basic relation

$$xu_1^2xu_1^{-1}x^{-2} = u_1xu_1^{-1}x^{-1}.$$

The resulting module presentation for G'/G'' has one generator u_1, and the relation $(1-3t+t^2)u_1 = 0$, giving the Alexander polynomial $\Delta(t) = 1-3t+t^2$, (or better $-1 + 3t - t^2$, so that $\Delta(1) = 1$.)
[Confirm that this gives $\nabla(z) = 1 - z^2$.]
(2) The trefoil group can be presented by $< x, y : x^2 = y^3 >$, where on abelianising we have $\varphi(x) = t^3, \varphi(y) = t^2$. Put $X = xy^{-1}$, and $u_1 = xX^{-3}, u_2 = yX^{-2}$. Then

$$\begin{array}{rcl} u_1X^3u_1X^3 & = & u_2X^2u_2X^2u_2X^2 \\ \text{and } X & = & u_1X^3X^{-2}u_2^{-1}. \end{array}$$

These give

$$\begin{array}{rcl} u_1 + t^3u_1 & = & u_2 + t^2u_2 + t^4u_2 \\ \text{and } u_1 - tu_2 & = & 0. \end{array}$$

The resulting presentation matrix is then

$$Q = \begin{pmatrix} 1 + t^3 & -(1 + t^2 + t^4) \\ 1 & -t \end{pmatrix}$$

and the E_0 ideal is generated by $\det Q = 1 - t + t^2$. Thus $\Delta(t) = 1 - t + t^2$, up to a unit in Λ.

Remark. There is a mechanisation of this procedure, called the 'Free differential calculus', which was developed by Fox to give a quick passage to a presentation matrix for G'/G'' starting from the group relations in a presentation for G.

8.3 Trivial Alexander polynomial

In general the module G'/G'' can be presented by $u_1, \ldots, u_{2g}$, with $2g$ relations $\sum q_{ij}u_j = 0$, and then $\Delta(t) = \det Q$, up to a unit. Then

$$\begin{aligned}
\Delta(t) = 1 &\Leftrightarrow \det Q \text{ is a unit in } \Lambda,\\
&\Leftrightarrow Q \text{ has an inverse in } \Lambda,\\
&\Leftrightarrow u_1 = \cdots = u_{2g} = 0,\\
&\Leftrightarrow G'/G'' = \{0\},\\
&\Leftrightarrow G'' = G'.
\end{aligned}$$

Examples can be found where $\Delta(t) = 1$ but $G \not\cong \mathbf{Z}$, for instance the 'doubled' knots described earlier in terms of a genus 1 spanning surface, although I did not give a proof that these were in fact knotted. Other examples include certain pretzel knots, where a homomorphism from G to a non-abelian group of 3×3 matrices can be exhibited.

8.4 Other modules

Every group G with abelianisation $\varphi : G \to < t : >$ gives rise to a presentation for G'/G'' as a module, and thus an E_0 ideal, but such groups do not in general arise as the group of a knot. For example, $G =< x, u : xu = u^2x >$ has abelianisation with $\varphi(x) = t, \varphi(u) = e$, and module presented by u with the relation $tu = 2u$. The resulting 'Alexander polynomial' $t - 2$ fails to have the symmetry properties which the polynomial for a knot would have, (it should have even degree for a start.)

8.5 The Alexander polynomial from the Wirtinger presentation

Recall that from a diagram of a knot with k crossings there is a presentation for the group of the knot as

$$G =< x_1, \ldots, x_k : r_1 = e, \ldots, r_k = e >,$$

with one generator for each arc, and one relation for each crossing, of the form $x_{i+1} = x_{j(i)}^{-\varepsilon(i)} x_i x_{j(i)}^{\varepsilon(i)}$, with the convention that $x_{k+1} = x_1$. Any one of the crossing relations is a consequence of the others and may be omitted, to give a presentation with $k - 1$ relations.

Choose any $x_j = x$ and rewrite with generators $x, u_1, \ldots, u_k$, where $u_i = x_i x^{-1}$. There is then one relation $u_j = 1$, with $k - 1$ others of the form $u_{i+1}x = (u_{j(i)}x)^{-\varepsilon(i)} u_i x (u_{j(i)}x)^{\varepsilon(i)}$.

For $\varepsilon(i) = 1$ this gives $u_{i+1} = t^{-1}(-u_{j(i)} + u_i) + u_{j(i)}$, and so

$$tu_{i+1} - u_i + (1-t)u_{j(i)} = 0$$

in the module G'/G'', while for $\varepsilon(i) = -1$ we get

$$-u_{i+1} + tu_i + (1-t)u_{j(i)} = 0.$$

This gives a $k \times k$ presentation matrix Q, whose ith row has entries $-1, t$ and $1-t$ in appropriate columns for the first $k-1$ rows, while the last row has a single entry of 1 in the jth place. The Alexander polynomial is then the determinant of this $k \times k$ matrix.

8.5.1 Relation to colouring

Consider the equations $Q\mathbf{d} = \mathbf{0}$ for a column $\mathbf{d} = \begin{pmatrix} d_1 \\ \vdots \\ d_k \end{pmatrix}$. When we put $t = -1$ the equations become $d_{i+1} + d_i = 2d_{j(i)}$ together with one equation $d_j = 0$. A choice of $d_1, \ldots, d_k$ not all 0, satisfying the equations mod n, is exactly what is needed, for n prime, to give an n-colouring of the knot K.

Regard entries in the matrix $Q(-1)$ as elements of $\mathbf{Z}_n$; the equations then have a non-zero solution $\mathbf{d}$ if and only if $\det(Q(-1)) = 0$ in $\mathbf{Z}_n$. Now $\det(Q(-1)) = \pm\Delta(-1)$ which is equal to zero in $\mathbf{Z}_n$ if and only if $\Delta(-1)$ is divisible by n.

This gives an alternative way to see that the trefoil can only be 3-coloured, and the figure-eight knot can only be 5-coloured. It shows also that colouring information is always available from knowledge of Δ. In particular it shows that a knot with $\Delta(t) = 1$ can not be n-coloured for any n.

Aside. A similar analysis for the equations when t is chosen to be a different integer detects the possibility of representing the knot group G onto the metacyclic group

$$< a, b : a^n = e, b^{-1}ab = a^t > .$$

[See Fox's 'Quick trip through knot theory' for more details.]

When $\Delta(t) = 1$ it is more difficult to find non-trivial representations of G on straightforward groups, although there is a general theorem which implies that every knot group except $\mathbf{Z}$ can be mapped onto some finite non-abelian group.

8.6 The Seifert matrix route to the Alexander polynomial

I shall not give a proof of the theorem relating the Seifert matrix approach to the fundamental group calculations, but a sketch of how it arises will be in order.

It is possible to construct from the exterior X of a given knot another 3-dimensional manifold $\tilde{X}$ called its infinite cyclic cover. The fundamental group of this manifold is the group K, and hence the abelian group K/K' is just the group $H_1(\tilde{X})$. As part of the construction of $\tilde{X}$ there is a 'shift homeomorphism', which induces an isomorphism on $H_1(\tilde{X})$. When this isomorphism is used to define an action of the generator of C_∞ on $H_1(\tilde{X})$ the group can be viewed as a module over Λ. It can be established fairly readily that this module is isomorphic to the Alexander module as previously described. It is, however, possible to use any choice of spanning surface for the knot, with Seifert matrix A, in making an explicit geometric construction of the space $\tilde{X}$. From this construction a presentation of $H_1(\tilde{X})$ as a Λ-module can be found, with presentation matrix $Q = tA - A^T$. The E_0 ideal for the module then has a single generator $\det(tA - A^T)$.

9 New invariants

In 1984 V.F.R.Jones constructed a new invariant of oriented links $V_L(t) \in \mathbf{Z}[t^{\pm\frac{1}{2}}]$, which turned out to have the property that

$$t^{-1}V_{L_+} - tV_{L_-} = (\sqrt{t} - 1/\sqrt{t})V_{L_0}$$

for links related as in the Conway polynomial relation. This was quickly extended to a 2-variable invariant $P_L(v,z) \in \mathbf{Z}[v^{\pm 1}, z^{\pm 1}]$, with the property that

$$v^{-1}P_{L_+} - vP_{L_-} = zP_{L_0}.$$

The name 'Homfly polynomial' has come to be attached to P, being the initial letters of six of the eight people involved in this further development. The polynomial P contains both the Conway/Alexander polynomial, and Jones' invariant, and can be shown to contain more information in general than both of these taken together. We have

$$\begin{aligned} P(1,z) &= \nabla(z) \\ P(1, s-s^{-1}) &= \Delta(s^2) \\ P(s^2, s-s^{-1}) &= V(s^2) \\ P(s, s-s^{-1}) &= \pm 1 \end{aligned}$$

The skein relation can readily be shown to determine P and V once its value on the trivial knot is given. It has been usual to take $P = 1$ on the trivial knot, although in some recent applications a different normalisation can be more appropriate.

Given the existence of V and P we can then make some calculations. For example, the unlink with two components has

$$\begin{aligned} P &= \frac{v^{-1} - v}{z}, \\ V(s^2) &= -(s + s^{-1}), \end{aligned}$$

while the Hopf link with linking number +1 has

$$\begin{aligned} P &= vz + (v^{-1} - v)v^2 z^{-1}, \\ V(s^2) &= s^3 - s - (s + s^{-1})s^4 = -s(1 + s^4). \end{aligned}$$

The Hopf link with linking number -1 has

$$\begin{aligned} P &= -v^{-1}z + (v^{-1} - v)v^{-2}z^{-1}, \\ V(s^2) &= -s^{-1}(1 + s^{-4}). \end{aligned}$$

This illustrates the general feature that for the mirror image $\overline{L}$ of a link L, (where the signs of all crossings are changed), we have $P_{\overline{L}}(v, z) = P_L(v^{-1}, -z)$ and so $V_{\overline{L}}(s^2) = V_L(s^{-2})$. It is thus quite possible to use V in many cases to distinguish a knot from its mirror-image, while there will be no difference in their Conway polynomials. It is worth noting that although there are still knots which cannot be distinguished from each other by P in spite of being inequivalent, no non-trivial knot has so far been found for which $P = 1$, or even $V = 1$.

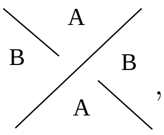
9.1 Kauffman's bracket polynomial

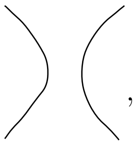
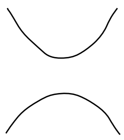
Before making any further calculations of the Jones polynomial V I shall give a derivation of it, due to Kauffman, which is remarkably straightforward, and which has subsequently been used to prove a long-standing conjecture about the class of knots which are known as 'alternating'. The methods adopted by Kauffman have also led to a very nice geometric way of recovering the algebra which Jones used in his original construction of the invariant V.

The construction starts from an unoriented link diagram D for a link L, and associates with it an integer polynomial $[D]$ in 3 variables A, B, and δ. Relations between the variables are then imposed which ensure that the polynomial is not changed when D is altered by Reidemeister moves II and III. Finally a suitable correction for the effect of Reidemeister move I can be made to give an invariant of the link L, which yields Jones' original invariant V.

The polynomial is defined inductively on the number of crossings in the diagram, using two rules.

The principal rule relates the polynomial $[D]$ to those of the two diagrams given from D by selecting one crossing, and cutting it out in each of the two possible ways. It is possible to distinguish these ways systematically by observing that the four quadrants defined around the chosen crossing can be labelled alternately as two types, A and B, say, by the convention that turning the overcrossing arc anticlockwise will sweep out the two regions to be labelled A. Then we can regard one of the two ways of cutting out the crossing as ‘opening the A-channel’, in Kauffman’s words, when the two A quadrants are connected to give a diagram D_A, while the other way opens the B-channel, to give a diagram D_B, as shown below.

$D =$ 

, $D_A =$  , $D_B =$  .

The main rule then says that

$$[D] = A[D_A] + B[D_B].$$

Repeated application of this rule to the remaining crossings in turn gives an expression of $[D]$ in terms of polynomials of diagrams with no crossings, which then consist of a number of disjoint simple closed curves in the plane.

The second rule allows us to finish the definition, by the assignment of a factor δ for every disjoint component of the diagram without crossings. This can be summarised as

$$[O\ D] = \delta[D],$$

where $O\ D$ is a diagram consisting of D together with a single disjoint curve without crossings.

While the calculation of $[D]$ appears to depend on the order in which crossings are removed it is fairly clear that, when D has c crossings, the resulting polynomial will be the sum of 2^c terms, each term arising by a choice, for each crossing, of either the A-channel or the B-channel. A formal definition of $[D]$ can then be given in terms of what Kauffman calls the ‘states’ of the diagram D.

Definition. A *marker* on a diagram is a selection at one crossing of the diagram of a pair of opposite quadrants at that crossing.

Definition. A *state* of a plane diagram D is a choice of one marker for each crossing of the diagram.

These definitions could equally be made for the projection of a diagram, in which over- and undercrossings are not distinguished.

A marker in a diagram D will be either an A-marker or a B-marker, depending on the quadrants which are selected by it. Any state S of D will then have $a(S)$ markers of type A, say, and $b(S)$ of type B, with $a(S)+b(S) =$

c. Splitting the diagram D apart using the markers of the state S will yield a number, $|S|$, say, of disjoint simple closed curves.

Then the polynomial $[D]$ is defined as

$$[D] = \sum_{\text{states } S} A^{a(S)} B^{b(S)} \delta^{|S|}.$$

The next step is to impose relations on A, B and δ so that $[D]$ is unaltered by Reidemeister moves on D. A quick calculation shows that invariance under Reidemeister move II is guaranteed by the choice of

$$AB = 1, \quad \delta = -A^2 - B^2$$

and further that invariance under Reidemeister III is ensured by invariance under Reidemeister II.

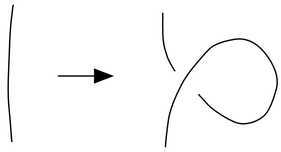
This provides Kauffman's definition of his bracket polynomial $< D >$, as a Laurent polynomial in $\mathbf{Z}[A^{\pm 1}]$ by taking

$$< D > = \delta^{-1}[D],$$

with $B = A^{-1}$ and $\delta = -A^2 - A^{-2}$.

Note that a factor of δ has been removed from $[D]$ so that the simple unknotted diagram O has $< O > = 1$. (There are reasons related to other appearances of these polynomials why the normalisation of $[D]$ with $[O] = \delta$ is often more appropriate).

This bracket polynomial is then invariant under Reidemeister moves II and III. It is not invariant under Reidemeister move I, so that it does not directly provide a link invariant. However it can quickly be shown that $< D >$ is multiplied by a fixed element $\alpha^{\pm 1}$ when D is altered by a Reidemeister move of type I. The element α is in fact $-A^{-3}$ for the Reidemeister move illustrated here.



It is then possible to compensate for Reidemeister move I and produce an invariant of an oriented link L as follows. Choose any diagram D of L and write $w(D)$ for the signed crossing number of the diagram, i.e. $w(D) = \sum \varepsilon(c)$ over crossings c in D. Now write

$$f_L = \alpha^{w(D)} < D >.$$

Then f_L does not depend on the choice of diagram, for the crossing number does not change under Reidemeister moves II and III, while the change under Reidemeister I exactly balances the change of the bracket polynomial.

Comparison of f_L with the defining properties of Jones original polynomial $V_L(t)$ shows that $f_L(A) = V_L(A^{-4})$, so that the bracket polynomial gives a direct approach to the Jones polynomial, requiring very little formal machinery in its construction.

9.2 Crossing number and the Jones polynomial

One of the most satisfying results proved using the new invariants has been the relation between the possible number of crossings in any diagram of the knot and a simple feature of the Jones polynomial. It singles out an alternating diagram for the knot (if it has one) as being particularly efficient as regards number of crossings, and provides one of the relatively few 'if and only if' relations between algebraic and geometric properties of knots.

Definition. For a Laurent polynomial

$$P = \sum_{r=m}^{n} c_r A^r \in \mathbf{Z}[A^{\pm 1}] \text{ with } c_m \neq 0, c_n \neq 0,$$

set $\text{span}(P) = n - m$.

Let L be a link. Then $\text{span}< L >$ can be defined using any diagram for L, since multiplication of a Laurent polynomial by a power of the variable does not alter its span.

Theorem 9.1 *Let L have a diagram D with $c(D)$ crossings. Then* $\text{span}< L > \leq 4c(D)$.

Corollary 9.2 $\text{span}< L > = \text{span} f_L \leq 4c$, *where c is the minimum number of crossings in any diagram of L.*

Corollary 9.3 $\text{span} V_L \leq c$, *as a Laurent polynomial in $t = A^{-4}$.*

Definition. A diagram is an *obvious sum* if a circle can be drawn in the plane of the diagram meeting in two points, with some crossing points lying on either side of the circle.

Theorem 9.4 *If L has a diagram with $c(D)$ crossings which is not an obvious sum, then* span$< L > = 4c(D)$ *if and only if the diagram is alternating.*

Corollary 9.5 *For an alternating diagram of L which is not an obvious sum the number of crossings in the diagram is the minimum crossing number c, and this minimum number will only be achieved by alternating diagrams for L. For knots without an alternating diagram span$V_L < c$.*

Proof of theorem 9.1: Given a diagram for L with $c(D)$ crossings we calculate the bracket polynomial for that diagram as

$$< L > = \sum_{\text{states } s} \varphi_S,$$

where $\varphi_S = A^{a(S)}B^{b(S)}\delta^{|S|-1}$ is a Laurent polynomial in A with $B = A^{-1}$ and $\delta = -A^2 - A^{-2}$. Put $D_S =$ maximum degree of φ_S, and $d_S =$ minimum degree of φ_S, so that span$(\varphi_S) = D_S - d_S$. We have

$$\begin{aligned} D_S &= a(S) - b(S) + 2(|S| - 1), \\ d_S &= a(S) - b(S) - 2(|S| - 1). \end{aligned}$$

Write $S' < S$ if the state S' arises from S by changing some A markers to B markers.

Proposition 9.6 *If $S' < S$ then $D_{S'} \leq D_S$ and $d_{S'} \leq d_S$.*

Proof : It is enough to show this when one marker is altered. In this case the number of circuits will alter by one, giving $|S'| = |S| \pm 1$. Since $a(S') = a(S) - 1$ and $b(S') = b(S) + 1$ the result follows at once. □

Among all the states there is one state S_A where all the markers are A, and another, S_B where all the markers are B; then $S_B \leq S \leq S_A$ for every state S. By the proposition, $D_S \leq D_{S_A}$ and $d_S \geq d_{S_B}$ for all states S. It follows that the highest degree in $< L >$ is at most D_{S_A} and the lowest at least d_{S_B} so that span$< L > \leq D_{S_A} - d_{S_B}$.

Dual states. Every state S has a *dual* state $\hat{S}$ in which the marker at every crossing is changed.

For example, $\hat{S}_A = S_B$. We have in general that

$$D_S - d_{\hat{S}} = 2a(S) - 2b(S) + 2(|S| + |\hat{S}| - 2).$$

Theorem 9.1 now follows from the first of two 'dual state' results, which relates the numbers of circuits for a state and its dual.

Proposition 9.7 *For any dual states $S, \hat{S}$ of a connected diagram with c crossings we have $|S| + |\hat{S}| \leq c + 2$.*

Consequently, since $a(S_A) = c, b(S_A) = 0$, we have

$$D_{S_A} - d_{S_B} \leq 2a(S_A) - 2b(S_A) + 2c = 4c,$$

completing the proof of theorem 9.1. □

The remaining results can best be viewed by considering the projection of the knot diagram to lie, as a curve Γ with self-crossings, entirely in a plane, or better in a 2-sphere S^2. The diagram can be recovered by knowing which of the two choices of marker at each crossing is to be the A marker, in other words by the choice of one state from among the possible 2^c states for Γ, thought of as a selection of a marker at each crossing.

9.3 States surfaces

We construct a surface F_S in $\mathbf{R}^3$ for any given choice of a state S of Γ which meets the plane $\mathbf{R}^2 \times \{0\}$ in the singular curve Γ, having c saddle-point singularities of the height function at the crossings of Γ. The surface is completed above the plane by $|S|$ local maxima, and below by $|\hat{S}|$ local minima, and has no further singular points.

The closed surface F_S is formed by placing the $|S|$ circuits of the state S in a plane immediately above the plane of Γ and the $|\hat{S}|$ circuits of $\hat{S}$ immediately below, and joining them up by a surface with a saddle point at each crossing of Γ. (This may be done explicitly in a polygonal form, but it can just as well be viewed in terms of standard smooth saddles.) The resulting boundary components can be capped off above and below by discs, treating innermost components first, to give a surface F_S as claimed.

Since F_S lies in $\mathbf{R}^3$ it must be orientable, and its Euler characteristic is given in any case as

$$\chi(F_S) = \text{maxima} + \text{minima} - \text{saddles}.$$

Then $\chi(F_S) = |S| + |\hat{S}| - c \leq 2$, proving proposition 9.7.

9.4 The shaded states

The projection curve Γ divides the plane into regions which can be shaded alternately black and white in a checkerboard pattern. Such a shading determines a pair of dual states Bl and Wh, by choosing the marker at each crossing which joins the regions of the given shading. The only choice in the shading is to reverse the role of black and white.

It is not difficult to see that Γ comes from an alternating diagram if and only if the states S_A and S_B are the checkerboard pair {Bl,Wh}.

Proposition 9.8 *The surface F_{Wh} is a sphere.*

Proof : Move the sphere which contains Γ slightly so that white regions are moved up and black regions down, leaving Γ itself at the original level. This surface will have saddles, maxima and minima as required for F_{Wh}, which is then a 2-sphere. □

To complete the proof of theorem 9.4, I shall have to quote a result about graphs in S^2.

Let Γ be a 4-valent graph in S^2, which is not a sum (i.e. it cannot be broken apart non-trivially by cutting in the middle of two edges). Let Γ' be another graph in S^2 which is isomorphic to Γ by an isomorphism preserving the local pairing of edges at vertices. Then this isomorphism extends to a homeomorphism from S^2 to S^2.

Corollary 9.9 *If Γ is not a sum and $F_S \cong S^2$ then S is one of the checkerboard states.*

Proof : The isomorphism of Γ in F_S with Γ in the plane will extend to a homeomorphism carrying the circuit curves of S from the sphere F_S into the plane so that they will lie in the shaded regions of one type. The markers for S must then be the markers for that shaded state. □

Proof of theorem 9.4: Suppose first that the diagram with $c(D)$ crossings is not alternating, (and not an obvious sum). Then S_A is not a checkerboard state, so by corollary 9.9 $F_{S_A} \ncong S^2$. Now

$$\begin{aligned} \text{span}< L > \leq D_{S_A} - d_{S_B} &= 2c(D) + 2(|S_A| + |S_B| - 2) \\ &= 2c(D) + 2c(D) + 2(\chi(F_{S_A}) - 2) \\ &< 4c(D) \end{aligned}$$

since $\chi(F_{S_A}) < 2$.

Conversely, if the diagram is alternating then $D_{S_A} - d_{S_B} = 4c(D)$. For any state S' other than S_A or S_B we have $\chi(F_{S'}) \leq 0$, since $F_{S'}$ is orientable, and the diagram is not an obvious sum. Then

$$|S'| + |\hat{S}'| \leq c(D),$$

while

$$|S_A| + |S_B| = c(D) + 2.$$

Now if S' has just one B marker then $|S'| = |S_A| \pm 1$ and $|\hat{S}'| = |S_B| \pm 1$. The inequality above then shows that $|S'| = |S_A| - 1$, and consequently $D_{S'} < D_{S_A}$.

Since all other states S have $S < S'$ for some S' with a single B marker it follows that S_A is the unique state with maximum degree D_{S_A} and so the largest degree in $< L >$ is exactly D_{S_A} (occurring with coefficient ±1). Similarly the least degree in $< L >$ is d_{S_B} and thus span$< L > = D_{S_A} - d_{S_B}$, completing the proof of theorem 9.4. □

10 Tangles and algebras

Given a link $L_{(p,q)}$ which has p and q half-twists respectively in the places indicated in figure 1 it is useful, in trying to calculate $< L >$, to start by concentrating on the part of the diagram involving just the p twists.

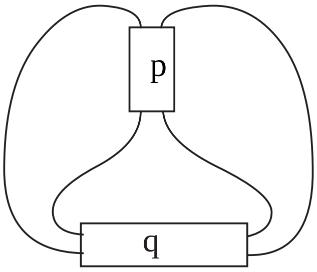


Figure 1

Begin by applying the first bracket relation at one crossing in this part of the diagram, to get

$$< L > = A< L_1 > + A^{-1}< L_2 >$$

where L_1 and L_2 differ from L only by removing the crossing in either of the two ways.

If we imagine a box around the part of the diagram with p twists, then L_1 and L_2 differ from L only within this box, and we could continue to simplify $< L_1 >$ and $< L_2 >$ by altering crossings inside the box, until $< L >$ is written as a linear combination of brackets of diagrams which are the same as L outside the box, but have no crossings at all inside the box. Taking account of the second bracket relation, which allows disjoint closed curves without crossings in a diagram to be removed, on multiplying by δ, we can simplify further to write $< L >$ as a linear combination of brackets of diagrams having no crossings or closed curves inside the box. In this case there will then be just two diagrams to account for, one where there are connections straight through the box, and one where the two top points are joined, as are the two bottom points.

Definition. We may formalise the calculations by referring to the piece of knot-diagram in the box as a *2-tangle.* More generally an *n-tangle* is a piece of knot diagram in a box where there are n points at the top and at the bottom where the strings enter and leave.

When a 2-tangle T gives rise to two 2-tangles T_1 and T_2 on cutting out a crossing in either way, as for the first bracket relation, we shall write T

formally as a linear combination

$$T = AT_1 + A^{-1}T_2.$$

Where T forms part of a link diagram the effect, when calculating $< >$, is the same as taking this linear combination of the brackets when T_1 and T_2 replace T. Further expansion of T_1 and T_2 as linear combinations of simpler tangles will then allow a simplification of the calculation of the bracket for the diagram.

In our example we may also make use of a natural product on the set of 2-tangles (or equally for n-tangles where appropriate) defined by placing one on top of another. For our purposes, tangles will be considered as equal when altered by Reidemeister moves II and III inside their box. The product allows us to treat the set of linear combinations of 2-tangles as an algebra over $\mathbf{Z}[A^{\pm 1}]$, using the product as above. The tangle consisting of two straight-through strings acts as the identity in this algebra. Write h for the other 2-tangle without crossings, with strings joining top to top and bottom to bottom. Then we will have

$$\sigma = A + A^{-1}h,$$

where σ is the 2-tangle with a single crossing whose A marker runs vertically. (The single-crossing tangle with horizontal A marker is then $A^{-1} + Ah$ and is σ^{-1} in the algebra.) Under product of tangles we have

$$h^2 = \delta h, h\sigma = (A + A^{-1}\delta)h = \alpha h, \text{ say,}$$

where $\alpha = -A^{-3}$.

Now any 2-tangle can be written as a linear combination of 1 and h, as claimed above. The tangle in our example can be written as σ^p, which is then rewritten, after one crossing move as

$$\sigma^p = A\sigma^{p-1} + A^{-1}h\sigma^{p-1} = A\sigma^{p-1} + A^{-1}\alpha^{p-1}h.$$

Continuing similarly with σ^{p-1} we can eventually write σ^p in terms of 1 and h as

$$\sigma^p = A^p + \frac{\alpha^p - A^p}{\delta}h.$$

Aside. As a short-cut, note that $e = \delta^{-1}h$ is an idempotent (where suitable denominators are allowed), i.e. $e^2 = e$. Then $f = 1 - e$ is an orthogonal idempotent, i.e. $f^2 = f$, $ef = 0$. Now $\sigma = A(e+f) + \delta A^{-1} e = Af + \alpha e$, so that $\sigma^p = A^p f + \alpha^p e$, giving the formula above.

The bracket polynomial of $L_{(p,q)}$ can now be calculated, using the expressions for σ^p and σ^q in terms of 1 and h in place of the two tangles. The result is a linear combination of the bracket polynomials of four diagrams, in which tangles 1 or h are inserted in each of the two tangle boxes. These diagrams consist of either one or two curves without crossings, giving bracket polynomials either 1 or δ. Then

$$\begin{aligned} < L_{(p,q)} > &= A^p A^q + \frac{1}{\delta^2}(\alpha^p - A^p)(\alpha^q - A^q) + A^p(\alpha^q - A^q) + A^q(\alpha^p - A^p) \\ &= (\frac{1}{\delta^2} - 1)(\alpha^p - A^p)(\alpha^q - A^q) + \alpha^p \alpha^q \\ &= \alpha^p \alpha^q((1-(-t)^p)(1-(-t)^q)(\frac{1}{\delta^2} - 1) + 1), \end{aligned}$$

where $t = A^4$. This gives a calculation of V_L after correction for the writhe by multiplication by a suitable power of α. The factor depends on whether p and q are even or odd or mixed; when both are even the factor is $\alpha^{-p}\alpha^{-q}$, as is the case with both odd, when one of the possible orientations of the 2-component link is chosen. With one odd and one even the factor is $\alpha^{\pm(p-q)}$. It is generally possible to recover the pair $\{p, q\}$ from V, except for small absolute values, when a knot can be represented both with an even pair and a mixed pair of values, for example $2, q$ and $-2, q-1$.

Calculations with 2-tangles depend on being able to find a 2-tangle in the diagram. This is the next best possibility to finding a 1-tangle, which is equivalent to writing the diagram as a connected sum. Notice that the bracket of a connected sum is the product of the brackets for the two factors, as, by a similar analysis, each 1-tangle may be written as a linear combination of 1-tangles without crossings, which means simply a multiple of the single identity 1-tangle with a straight-through string. This multiple must be the bracket polynomial of the diagram given by joining the two ends of the 1-tangle, i.e. one of the factors in the sum.

Similar calculations can be done with 3-tangles or even n-tangles, although the algebras involved become more complicated. For example, in dealing with 2-tangles we reduced everything to a linear combination of two tangles, and thus a 2-dimensional algebra. For 3-tangles there are 5 basis elements for the algebra (corresponding to tangles without crossings, while in general there are $\binom{2n}{n}/(n+1)$ basis elements for the Temperley-Lieb algebra which results from n-tangles. (These algebras, in a slightly varied form, were the origin of Jones' definition of V, although the connection with knot diagrams was demonstrated some time later by Kauffman.)