An Introduction to Knot Theory

Pranjal Jain
Indian Institute of Science Education and Research,
Pune
Roll number: 20201257

Under guidance of

Dr. Tejas Kalelkar

Preface

Sections §1 to §5 of this report are based on the first five chapters [1], and §6 is based on the main theorems of [2] which are of topological interest. Most results and the ideas used in their proofs are taken from these two sources, although the wording of the proofs and the flow of the text are my own.

I have also included some results which were either left as an exercise or assumed without proof in the two references, and in such cases the proofs provided are my own. Some of these results are **proposition 1.5**, **corollary 3.5**, **lemma 6.1**, **corollary 6.3**. The proof of **proposition 6.4** was told to me by Prof. Kalelkar during one of our meetings. The proof of **theorem 1.4** is also my own, although it has been proved in [1] in a different way.

Although the contents of §1, §2 and §3 mostly match with chapters 1, 2 and 3 of [1] respectively, I decided to include the contents of chapter 4 in §5 and chapter 5 in §4. This was done in order to present results in a way which motivates the study of alternating links without making it seem like the alternating condition was brought into the picture 'out of the blue'. With this goal in mind, the flow of §4 has been altered significantly compared to chapter 5 in [1]. **Theorem 5.5** has been motivated using the contents of §4 with the goal of strengthening **corollary 4.13**.

In §6, part of the proof of **theorem 6.17** provided in [2] involving the use of various argument functions has been modified for clarity, although the core idea is the same.

Pranjal Jain

Contents

1 Introduction

A knot is a subset of oriented S^3 which is an embedding of S^1 in S^3 (in **Top**, **Diff** or **PL**). Likewise, a link is a subset of S^3 which is an embedding of finitely many copies of S^1 in S^3. Two knots or links are considered equivalent if they are ambient isotopic in S^3. Knot theory is the study of knots and links, mostly in the categories **Diff** and **PL**. In this project I studied knot theory in **PL**, although it is known that the knot theories in the categories **PL** and **Diff** are equivalent. Henceforth all discussion will be in the category **PL**, so a link can be regarded as being a union of finitely many line-segments, any two of which intersect at at most one point (which is an endpoint of both).

The condition of ambient isotopy for equivalence of links can be simplified as a result of working in **PL** in two different ways. First, every orientation-preserving homeomorphism of S^3 with itself is isotopic to the identity map on S^3. Hence, two links are equivalent if and only if there is an orientation-preserving homeomorphism $h : S^3 \to S^3$ taking one to the other. Second, two links L_1 and L_2 are ambient isotopic if and only if they are isotopic. Hence, when showing that (or using the fact that) two knots or links are equivalent we may use any one of these three definitions. Henceforth, we will not distinguish between a knot/link and its equivalence class.

The simplest knot is the **unknot**, whose equivalence class consists of all those knots which are isotopic to a planar circle. This is equivalent to the following definition.

Definition 1.1. A knot K is said to be the unknot if K bounds a disc (in S^3).

To see this equivalence, first note that clearly any round S^1 must bound a disc in S^3. Hence, any ambient isotopy taking a round S^1 to K also yields a disc bounded by K. Conversely, if K bounds a disc D, let $D' \subset D$ be a sufficiently small disc in the interior of D so that $\partial D'$ is isotopic to a round S^1. The annulus $D \backslash D'$ gives an isotopy between K and $\partial D'$.

A related concept is that of a **meridian** of a link L, which is a simple closed curve isotopic (in $S^3 \backslash L$) to a small planar circle through whose centre some component of L passes. The following definition is equivalent to this one.

Definition 1.2. A simple closed curve γ in the complement $S^3 \backslash L$ of a link L is said to be a meridian of L if γ bounds a disc which intersects L transversally and at exactly one point.

For a link L, pick a point $p \in S^3 \backslash L$. Viewing S^3 as the one-point compactification of $\mathbb{R}^3$ with p as the point at infinity, L is now a subset of $\mathbb{R}^3$ which is an embedding of finitely many copies of S^1 in $\mathbb{R}^3$. By a general position argument, we may perform an isotopy on L so that under the projection $\pi : \mathbb{R}^3 \to \mathbb{R}^2$ onto the xy-plane, line segments of L go to line segments in $\mathbb{R}^2$. Furthermore for any two segments l_1 and l_2 in the PL-decomposition of L, the segments $\pi(l_1)$ and $\pi(l_2)$ intersect at at most one point (which is not an endpoint of both if and only if $l_1 \cap l_2 = \emptyset$). Such a position for L is called a **standard position**.

An intersection point of two line segments in $\pi(L)$ which is not an endpoint is called a crossing. When drawing $\pi(L)$, we may denote which line segment is under the other at a crossing ('over/under' information) by making a break in the one which passes under the other. In a small neighbourhood of a crossing, we define the **under-strand** to be the line segment passing under and likewise define the **over-strand**. $\pi(L)$, along with the over/under information at each crossing, will be called a planar **diagram** for L. L can be reconstructed up to isotopy from a planar diagram of L in the obvious way, though L may have several diagrams which are not equivalent through isotopies of $\mathbb{R}^2$.

Now the one-point compactification of the projection plane $\mathbb{R}^2$ can be taken by reintroducing p, so that $\pi(L)$ is now a subset of a 2-sphere S^2 in S^3. In this setting, $\pi(L)$ (with the over/under information) will be called a diagram of L on S^2. Once again, L can be reconstructed upto isotopy from a diagram of L on S^2. Henceforth we will work with diagrams on S^2 throughout this report, unless explicitly mentioned otherwise.

Clearly, any homeomorphism $h : S^2 \to S^2$ can be extended to a homeomorphism $\tilde{h} : S^3 \to S^3$ so that $h \circ \pi(L) = \pi \circ \tilde{h}(L)$. Hence, an isotopy of S^2 which modifies a planar diagram of L may in effect be viewed as an isotopy of S^3 modifying L.

Reidmeister's theorem states that two diagrams (planar or on S^2) D_1

and D_2 represent the same link if and only if they are equivalent upto the so-called Reidmeister moves and an isotopy of the projection plane/sphere, i.e. there is a sequence of Reidmeister moves R-I, R-II and R-III (see fig. 1) which, together with an isotopy of the projection plane/sphere, takes D_1 to D_2.

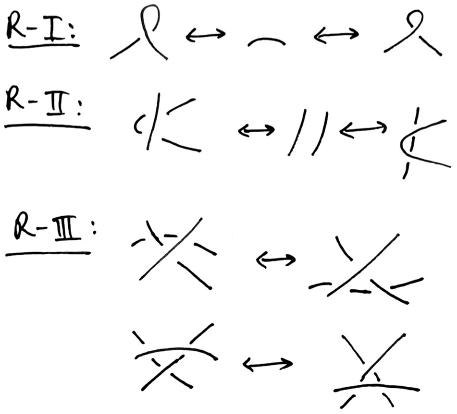


Figure 1: The three Reidmeister moves. The 'loop' occuring in the definition of R-I moves is called a **monogon**.

Hence if there is a way to associate some object (such as a number, polynomial, etc.) to every diagram, and the association is invariant under all the Reidmeister moves, then that association yields a **link invariant** (it depends only on the link represented by the diagram and not on the diagram itself). One such link invariant is the **crossing number**.

Definition 1.3. The crossing number of a link is the least number of crossings in a diagram for it.

It is easy to see that any knot diagram with 2 or fewer crossings is a diagram of the unknot, so no knot can have crossing number 1 or 2. In general the crossing number of a knot is difficult to determine, but in §4 and §5 we will see some specific cases where it can be determined.

A knot can be given an orientation by fixing a direction of travelling around it, and an orientation can be given for a link by giving an orientation for each of its components. In any diagram D of an oriented link L, we assign a sign of $+1$ to all those crossings in which the over-strand must be rotated counter-clockwise for its orientation to align with that of the under-strand. All other crossings are assigned a sign of -1. For two distinct components K_1 and K_2 of L, the **linking number** $\mathrm{lk}(K_1, K_2)$ is defined as half the sum of signs of those crossings which have one strand from K_1 and one from K_2. $\mathrm{lk}(K_1, K_2)$ is well-defined (i.e. does not depend on the specific choice of a diagram for L) since it is invariant under all three Reimeister moves. Also, it is clear that $\mathrm{lk}(K_1, K_2) = \mathrm{lk}(K_2, K_1)$. It can be seen that the linking number is an integer by using a crossing-smoothing argument[1] on one of the knots (say K_1) in a diagram for L and counting how many times the diagram for K_2 enters and exits regions bounded by the simple closed curves in the smoothened version of the diagram of K_1.

The linking number, although defined combinatorially using diagrams, has a relation with the first singular homology of the complement of a knot.

Theorem 1.4. *For K an oriented knot and $X := S^3 \backslash K$ its complement, we have $H_1(X) \cong \mathbb{Z}$. Furthermore, the map which takes the homology class of an oriented simple closed PL curve $\lambda \subset X$ to $\mathrm{lk}(K, \lambda)$ is an isomorphism from $H_1(X)$ to $\mathbb{Z}$.*

Proof. This theorem is proved in [1] using the Mayer-Vietoris sequence. Here I present my own argument. It suffices to look consider the complement of K in $\mathbb{R}^3$ and find its first homology group, so let $X = \mathbb{R}^3 \backslash K$. Using an isotopy of $\mathbb{R}^3$ we bring K into standard position and project to get a planar diagram of K on the xy-plane with n crossings. This diagram may be rotated (about some point on the plane) so that the x-coordinates of all crossings of K are distinct. Write $\mathbb{R}$ as a union of n open intervals $I_1, \ldots, I_n$ so that each contains the x-coordinate of exactly one crossing, and suppose also that the indexing of these intervals sorts them from left to right. This gives a decomposition of X as a union of the open sets $U_i := (I_i \times \mathbb{R} \times \mathbb{R}) \backslash K$ such that $U_i \cap U_j \neq \emptyset$ if and only if $|i - j| \leq 1$. Furthermore, since the projection of $K \cap (I_i \times \mathbb{R} \times \mathbb{R})$ contains only one crossing, each U_i is homeomorphic to $(D^2 \backslash P_i) \times I$, for some finite set $P_i \subset D^2$.[2] Likewise is true for the sets $U_i \cap U_{i+1}$.

---

[1] This technique is elaborated on in the proof of **theorem 2.1**.

[2] The cardinality of P_i is precisely the number of components of $K \cap (I_i \times \mathbb{R} \times \mathbb{R})$.

Hence, inductively applying the Van Kampen theorem on the decomposition $X = U_1 \cup \ldots \cup U_n$ gives a presentation for the fundamental group $\pi_1(X)$,[3] from which it is easy to see that the abelianization of $\pi_1(X)$ is isomorphic to $\mathbb{Z}$. Hence we obtain $H_1(X) \cong \mathbb{Z}$, proving the first part of the theorem. It can also be seen from here that every meridian of K generates $H_1(X)$ and an isomorphism from $H_1(X)$ to $\mathbb{Z}$ is generated by taking the homology class of any PL meridian in standard position to its linking number with K.

Now, we bring the link consisting of λ and K into standard position and once again project to get a planar diagram. Furthermore, we assume that the knot K is entirely above the projection plane. Call a crossing a λK-crossing if its over-strand belongs to λ and the under-strand to K, and likewise define $K\lambda$-crossings. For $\bullet$ being $K\lambda$ or λK, let $a_\bullet$ be the number of $\bullet$-crossings with sign 1 and $b_\bullet$ be the number of those with sign -1. Hence, the linking number of K and λ is given by

$$\operatorname{lk}(K, \lambda) = \frac{a_{K\lambda} + a_{\lambda K} - b_{K\lambda} - b_{\lambda K}}{2}$$

By smoothing all crossings of K with itself in the diagram as was done to show that the linking number is an integer, one obtains $a_{K\lambda} - b_{K\lambda} = a_{\lambda K} - b_{\lambda K}$, so that to calculate the linking number one needs only to add the signs of the crossings where λ passing over K.

Let p_0 be a point on λ which is under the plane of projection. λ can be changed upto homotopy so that it stays below the projection plane everywhere except in small nieghbourhoods of λK-crossings, where it ‘quickly’ goes over K and comes back down. Now, let c be a λK-crossing. Just before λ begins to climb before c, ‘insert’ a loop into λ which follows a straight line to p_0 and then comes back. The modified λ does not intersect K since the straight line is entirely below the projection plane (although λ intersects itself). Do the same procedure just after every c as well. Performing this procedure for every λK-crossing c, it can be seen that λ is homotopic to a composition of loops, each of which is homotopic to a meridian. Furthermore, the number of meridians in this representation with linking number 1 with K is $a_{\lambda K}$ and likewise there are $b_{\lambda K}$ meridians with linking number -1

---

[3] This presentation is essentially the same as that of Wirtinger, but I discovered it independently while attempting to prove **theorem 1.4**.

with K. Hence, the homology class of λ maps to $a_{\lambda K} - b_{\lambda K} = \mathrm{lk}(K, \lambda)$ under the above isomorphism. $\square$

Similar to the linking number, we define the **writhe** $w(D)$ of a diagram of an oriented link L as the sum of signs of all crossings in D. It can be checked easily that $w(D)$ is invariant under R-II and R-III moves, but it changes by ±1 under R-I moves depending on the orientation of the crossing added/removed. It can also be seen that the move shown in fig. 2 can be achieved (in diagrams on S^2) using only R-II and R-III moves.

Figure 2: This move can be done using a sequence of R-II and R-III moves.

Proposition 1.5. *Two diagrams D_1 and D_2 (on S^2) of the same oriented link are equivalent upto R-II moves, R-III moves and isotopy of S^2 if and only if $w(D_1) = w(D_2)$.*

Proof. As noted above, the writhe of a diagram is invariant under R-II and R-III moves so the forward implication is straightforward. For the backward implication, suppose $w(D_1) = w(D_2)$. We will give the proof for the case when L is a knot, and the case when L has more than one component is almost identical. Let Σ be a sequence of R-I, R-II and R-III moves which takes D_1 to D_2 (upto isotopy of S^2). Now we will perform a modified version of Σ on D_1, with the following modifications.

- Every time Σ uses an R-I move to add a monogon of some orientation (say $+1$), we replace it with the move shown in fig. 2 which adds two monogons of opposite orientation. The monogon of $+1$ orientation is used to perform the rest of Σ, and the other is made small and left free to 'move around' along the diagram so that it does not interfere with the other moves in Σ (this movement can be done using only R-II and R-III moves).

- Every time Σ uses an R-I move to remove a monogon, we do not perform that move and instead make the monogon small and free to move around along the diagram so that it does not interfere with the other moves in Σ (see fig. 3).

The above modified sequence uses only R-II and R-III moves. Furthermore, after performing modified sequence (and a suitable isotopy of S^2) to D_1 we obtain a new diagram D_1' which differs from D_2 only in that it has some small monogons which D_2 does not (if there are no such monogons then we are done). Collect all these monogons together so that they appear consecutively on some otherwise 'straight' part of D_1' (see fig. 4).

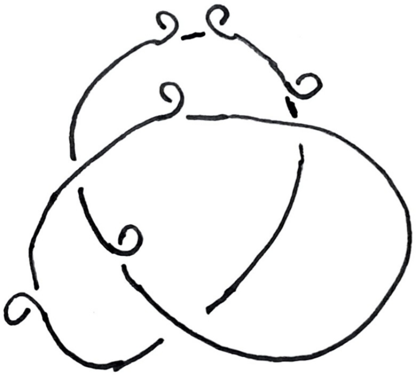


Figure 3: A diagram of the trefoil with small monogons.

Figure 4: All the monogons made consecutive.

Since writhe is invariant under R-II and R-III moves, we have $w(D_1') = w(D_2)$. Hence, there are an equal number of monogons of orientations 1 and -1. There is some pair of consecutive monogons with opposite orientations, so they can be removed using the move shown in fig. 2. This can be repeated until there are no more of these monogons left, which turns D_1' into D_2 using only R-II and R-III moves. $\square$

Two oriented knots K_1 and K_2 in different copies S_1 and S_2 of S^3 can be combined into one by taking a connected sum of S_1 and S_2 which joins K_1 and K_2 into one knot which recieves a consistent orientation from both of them. To elaborate, we choose 'small' balls $B_1 \subset S_1$ and $B_2 \subset S_2$ with $\partial B_1 \cap K_1$ and $\partial B_2 \cap K_2$ having two points each. We then glue the two balls $\overline{S_1 \backslash B_1}$ and $\overline{S_2 \backslash B_2}$ through a positively orientated homeomorphism between their boundaries ∂B_1 and ∂B_2 to get a new 3-sphere. This homeomorphism takes $\partial B_1 \cap K_1$ to $\partial B_2 \cap K_2$. Furthermore, this is done in such a way that the knot $\overline{K_1 \backslash B_1} \cup \overline{K_2 \backslash B_2}$ in the glued 3-sphere recieves consistent orientations from K_1 and K_2. Here, by saying that the balls B_1 and B_2 are 'small', we mean that the ball-arc pairs $(B_1, B_1 \cap K_1)$ and $(B_2, B_2 \cap K_2)$ are trivial. The oriented knot in the new 3-sphere is denoted by $K_1 + K_2$ and is called the **connected sum** (sum, for short) of K_1 and K_2. It can be shown using tools of PL theory that the definition of $K_1 + K_2$ does not depend on the precise choice of B_1, B_2 and the gluing homeomorphism, so that the notation $K_1 + K_2$ is well-defined. Hence, for non-oriented knots there are at most two different non-oriented knots that can be referred to as $K_1 + K_2$. If a statement is made regarding $K_1 + K_2$ with no orientation mentioned for K_1 and K_2 a priori, then it is understood that the statement is true for both interpretations of $K_1 + K_2$.

Diagrammatically, if D_1 and D_2 are diagrams of oriented knots then we obtain the diagram of the connected sum of the knots by placing D_1 and D_2 next to each other, removing small portions of each and then joining up the free ends of D_1 to those of D_2 in the unique way which is consistent with both of their orientations.

Definition 1.6. A knot P is said to be **prime** if it is not the unknot and whenever K_1 and K_2 are knots with $P = K_1 + K_2$, at least one of the K_i's must be the unknot.

The theory developed in §2 will justify the usage of the word 'prime' instead of 'irreducible' in the above definition.

2 Seifert surfaces and knot factorisation

A Seifert surface of a link L is a connected and orientable surface F with boundary $\partial F = L$. First we show that Seifert surfaces exist for all links and then define the genus $g(L)$ of L as the minimum genus of a Seifert surface for L. For knots the genus is additive under the connected sum for knots, and hence we will have an invariant for knots which plays well with connected sums. This will be useful for establishing that every knot can be uniquely 'factorised' as a connected sum of prime knots, upto reordering.

Theorem 2.1. *Every link has a Seifert surface.*

Proof. Let L be a link, and give it an arbitrary orientation. In a diagram D for L, we remove a small neighbourhood of every crossing and rejoin the strands in the unique way which is consistent with orientations and does not introduce crossings. Hence we get a new diagram D' with no crossings, and so every component of this diagram is an oriented simple closed curve. This process is called **smoothing crossings**. Let $\gamma_1, \ldots, \gamma_n$ be these curves. Viewing these curves as subsets of S^3, let $\Delta_1, \ldots, \Delta_n \subset S^3$ be discs with $\partial \Delta_i = \gamma_i$ and $\Delta_i \cap \Delta_j = \emptyset$ for $i \neq j$. Each Δ_i inherits an orientation from γ_i. Let F be the oriented surface $\bigcup\limits_i \Delta_i$.

We will now modify F so that ∂F becomes L. To do this, we reintroduce the crossings removed from D by adding half-twisted strips (rectangular copies of D^2) to F at all the places where a crossing was removed from D. The twist is done in the unique way which gives a crossing with the same strands going under and over as in D, and this way the orientations of the Δ_i's are faithfully relayed through the strips. Hence this new F is an orientable surface with $\partial F = L$. See fig. 5.

If F is not connected, thin cylindrical tubes can be used to join different components in a way which relays their orientations faithfully and does not affect the boundary. □

For a knot K with Seifert surface F, we know that ∂F is homeomorphic to S^1 and so the one-point compactification of $F \backslash \partial F$ is a closed, connected and orientable surface. Define the genus of F to be the genus of this closed surface, and define the **genus** $g(K)$ of K to be the minimum genus of F for all Seifert surfaces F of K. It is now immediate that $g(K) = 0$ if and only if K is the unknot.

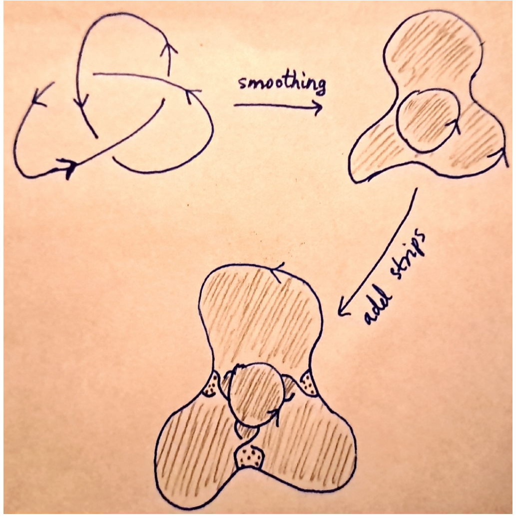


Figure 5: The construction of a Seifert surface for the trefoil. Two different shading patterns are used to represent the two different sides of the surface.

Theorem 2.2. *The genus of knots is additive with respect to the connected sum, i.e. if K_1 and K_2 are knots then*

$$g(K_1 + K_2) = g(K_1) + g(K_2)$$

For the proof of **theorem** 2.2, we will use without proof the Schönflies theorem, which is as follows.

Theorem 2.3 (Schönflies)**.** *If F is a PL embedding of S^2 in S^3, then S^3 can be written as a union of two 3-balls with common boundary F.*

Hence, if $K = K_1 + K_2$ for knots K_1 and K_2, then there exists a 2-sphere F which intersects K transversally at two points and decomposes S^3 as a union of two 3-balls B_1 and B_2 with boundary F such that the ball-arc pairs $(B_1, B_1 \cap K)$ and $(B_2, B_2 \cap K)$ are **isomorphic** to K_1 and K_2 respectively. Here, we say that a ball-arc pair (B, α) is isomorphic to a knot if the one-point compactification of $\alpha \backslash \partial\alpha$, as an embedding of S^1 in the one-point compactification of $B \backslash \partial B$, is that knot. Say that F **separates** K_1 and K_2 in K.

*Proof of **theorem** 2.2.* Let $K = K_1 + K_2$. It is clear that

$$g(K) \leq g(K_1) + g(K_2)$$

since a connected sum of Siefert surfaces gives a Seifert surface of the connected sum. For the other inequality, let F be a minimal genus Siefert surface for K. Let S be a 2-sphere intersecting K at two points p and q transversally with corresponding ball-arc pairs $(B_1, B_1 \cap K)$ and $(B_2, B_2 \cap K)$ isomorphic to K_1 and K_2 respectively. By making small adjustments to S if necessary, it may be assumed that F and S intersect transversally.

$F \cap S$ is a 1-manifold-with-boundary, and its boundary is precisely $(F \cap \partial S) \cup (\partial F \cap S) = \partial F \cap S = \{p, q\}$. Hence, one component of $F \cap S$ is an arc α between p and q and the others are simple closed curves (finitely many). If there are no components which are simple closed curves, then $F \cap B_1$ is a Seifert surface of the knot $(K \cap B_1) \cup \alpha = K_1$ and likewise $F \cap B_2$ is a Seifert surface of the knot $(K \cap B_2) \cup \alpha = K_2$. Hence, we have

$$\begin{aligned} g(K) &= \text{genus}(F) \\ &= \text{genus}(F \cap B_1) + \text{genus}(F \cap B_2) \\ &\geq g(K_1) + g(K_2) \end{aligned}$$

Now, suppose there is at least one simple closed curve in $F \cap S$. We will produce a new Seifert surface E of K with so that the number of components of $E \cap S$ is lesser than those of $F \cap S$. This will prove the claim by induction and the above case of $F \cap S = \alpha$. Let C be an innermost component of $F \cap S$ on S with respect to α being 'outside', i.e. C bounds a disc $\Delta \subset S$ with $\Delta \cap F = C$. Let N be a regular neighbourhood of C in F, so N is an annulus. Let D_1 and D_2 be discs bounded by the two components of ∂N with $D_i \cap S = \emptyset$ and $D_i \cap F = \partial D_i$. Hence, $F' = (F \backslash N) \cup D_1 \cup D_2$ is a new orientable surface with boundary K. If F' is connected, then F' is a Seifert surface of K and has genus smaller than F (a contradiction). Hence F' is disconnected, so one of the components E of F' has boundary given by K and the other has no boundary. It can now be seen that E is as desired. □

Corollary 2.4. *Knots do not have additive inverses, i.e. if K_1 and K_2 are knots with $K_1 + K_2$ the unknot, then K_1 and K_2 are both unknots.*

Corollary 2.5. *Every knot can be written as a connected sum of prime knots.*

Proof. If K is a knot with $K = K_1 + K_2$ with neither of K_1 and K_2 the unknot, then $g(K_1), g(K_2) < g(K)$. Hence, the claim follows by induction on $g(K)$. □

The following result which justifies calling prime knots 'prime' and hence gives the uniqueness of the factorisation given by **corollary 2.5**.

Theorem 2.6. *Let K be a knot and P be a prime knot such that $K = P+Q = K_1 + K_2$ for some knots Q, K_1, K_2. Then at least one of the following holds.*

1. *$K_1 = P + K_1'$ and $Q = K_2 + K_1'$ for some knot K_1'.*
2. *$K_2 = P + K_2'$ and $Q = K_1 + K_2'$ for some knot K_2'.*

Proof. Let B be a ball with ∂B separating P and Q in K and the ball-arc pair $(B, B \cap K)$ isomorphic to P. Let $\alpha = B \cap K$. Let F be a 2-sphere separating K_1 and K_2 in K. By making small adjustments to F, we may assume that F and ∂B intersect transversally, and so $F \cap \partial B$ is a disjoint union of simple closed curves. Furthermore, we may assume that $F \cap \partial B \cap K = \emptyset$.

The result is straighforward in the case of $F \cap \partial B = \emptyset$, so suppose $F \cap \partial B \neq \emptyset$. If C is a component of $F \cap \partial B$ bounding a disc in $F \backslash K$, then the two discs in ∂B with boundary C cannot both intersect K. Hence, such C may be picked to be innermost on ∂B with respect to $\partial B \cap K$ being outside. Hence C bounds discs $\Delta \subset F$ and $\Delta' \subset \partial B$, neither of which intersect K, and $\Delta' \cap F = C$. Such C can be removed from $F \cap \partial B$ by modifying F (while preserving the defining features of F) using the idea used in the proof of **theorem 2.2**. Hence, we may assume that no component of $F \cap \partial B$ bounds a disc in F which does not intersect K. By interchanging F and ∂B in the above argument, we may also assume that no component of $F \cap \partial B$ bounds a disc in ∂B which does not intersect K. Hence, in both F and ∂B, the components of $F \cap \partial B$ look like lines of latitude with the points of $K \cap F$ and $K \cap \partial B$ respectively being poles (**fact 1**).

Now, suppose there is some component $\Delta \subset F \cap B$ which is a disc. By the above argument, Δ must intersect K exactly once. Hence Δ separates the ball arc pair (B, α) into two ball-arc pairs. By primality of P, at least one of these ball-arc pairs is trivial. Removing a regular neighbourhood of this trivial ball-arc pair from (B, α) yields a modified (B, α) which still satisfies the defining features of B and has fewer components in $F \cap \partial B$. Hence, it may now be assumed that no component of $F \cap B$ is a disc. By fact 1, the components of $F \cap B$ are all annuli (**fact 2**).

Now if $F \cap B = \emptyset$ then $F \cap \partial B = \emptyset$ and we are done. Otherwise, a combinatorial argument shows that there exists a component $A \subset F \cap B$ (which must be an annulus, by fact 2) such that ∂A bounds an annulus $A' \subset \partial B$ which intersects F only at the boundary. Hence $\partial A = \partial A' = A \cap \partial B = A' \cap F$, so that $A \cup A'$ is a torus. Let $M \subset B$ be the region bounded by $A \cup A'$ in B (so M is a 3-manifold with boundary $A \cup A'$). Note that $M \cap K = \emptyset$.

Let $C \subset \partial A$ be a component and $\Delta \subset \partial B$ be the disc bounded by C which satisfies $\Delta \cap A' = C$. If N is a regular neighbourhood of Δ in $\overline{B \backslash M}$, then $M \cup N$ is a ball since its boundary is a 2-sphere. Let $\beta = N \cap \alpha$. If the ball-arc pair $(M \cup N, \beta)$ is trivial, then we can modify (B, α) by removing from it a (regular neighbourhood of) this trivial ball-arc pair so that B still satisfies its defining properities and ∂A is removed from $\partial B \cap F$. If $(M \cup N, \beta)$ is not trivial, then it must be isomorphic to P and the ball-arc pair $(\overline{B \backslash (M \cup N)}, \overline{\alpha \backslash \beta})$ must be trivial (by the primality of P).

Let X be the ball bounded by F which contains M and $\Delta' \subset F$ be the disc bounded by C which satisfies $\Delta' \cap A = C$. Let N' be a regular neighbourhood of Δ' in $X \backslash M$ and $\gamma = K \cap N'$. The ball-arc pair $(M \cup N', \gamma)$ is homeomorphic to $(M \cup N, \beta)$, and so is isomorphic to P. The ball-arc pair $(X, X \cap K)$ is isomorphic to one of K_1 and K_2 (say K_1), so $(M \cup N', \gamma)$ (which is entirely contained in X) allows us to write

$$K_1 = P + K_1'$$

for some knot K_1'.

Now, we will perform surgery on S^3 along ∂M to show that $Q = K_1' + K_2$. For this, we remove M from S^3 and glue to it a solid torus $T = S^1 \times D^2$ through a homeomorphism between ∂T and ∂M. This is done in such a way that $S_1 \times \{p\}$ is glued to C, for some $\{p\} \in D^2$. That this can be done follows from the decomposition of ∂M as a union of the annuli A and A'. Let the space obtained after this gluing be Y. It can be seen that $T \cup N$ is now a ball and so is $\overline{S^3 \backslash (M \cup N)} = Y \backslash (T \cup N)$, which means Y which intersect only at their common boundary. Hence Y is a 3-sphere. Also, $(T \cup N, \beta)$ is now a trivial ball-arc pair so K as a subset of Y is the knot Q. Likewise, the ball-arc pair $((X \backslash M) \cup T, X \cap K)$ is isomorphic to K_1' and so F separates K_1' and K_2 in Q. Hence,

$$Q = K_1' + K_2 \quad \square$$

Corollary 2.7. *Every knot K has a unique factorisation (upto reordering) $K = P_1 + \ldots + P_n$ for prime knots $P_1, \ldots, P_n$.*

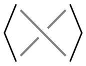
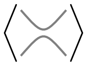
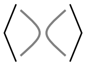
Proof. Existence is given by **corollary 2.5**. For uniqueness, use the standard algebraic argument involving induction on n, backed by **theorem 2.6**. ☐

Corollary 2.8. *If J, K, L are knots such that $J + K = J + L$, then $K = L$.*

3 The Jones polynomial of a link

The Jones polynomial $V(L)$ of an oriented link L is a Laurent polynomial in $t^{\frac{1}{2}}$ with coefficients in $\mathbb{Z}$, for some indeterminate t (so the Jones polynomial need not always be a polynomial in t). Hence, $V(L) \in \mathbb{Z}[t^{\frac{1}{2}}, t^{-\frac{1}{2}}]$. We will define it in terms of a given diagram of L and show that it is invariant under Reidmeister moves, hence establishing that it depends only on the link and not the diagram that was used for evaluating it.

For any diagram D of some link, we will first define the Kauffman bracket $\langle D \rangle$ of D as a Laurent polynomial in an indeterminate A which is invariant under R-II and R-III moves but not R-I moves. By noting precisely how $\langle D \rangle$ varies under R-I moves, it will be straightforward to see how it can be modified to be invariant under all three Reidmeister moves. We define $\langle A \rangle$ by induction on the number of crossings using the following conditions.

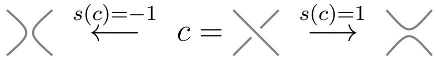
1. $\left\langle \bigcirc \right\rangle = 1$.
2. $\left\langle D \sqcup \bigcirc \right\rangle = \langle D \rangle \, (-A^2 - A^{-2})$.
3. $\Big\langle$  $\Big\rangle = A \Big\langle$  $\Big\rangle + A^{-1} \Big\langle$  $\Big\rangle$.

Here, $\bigcirc$ denotes the canonical diagram of the unknot and $D \sqcup \bigcirc$ denotes the disjoint union of D and $\bigcirc$. The third condition gives the bracket of a diagram in terms the brackets of the diagrams obtained by smoothing one of the crossings (the rest of the diagram is kept the same). The first two conditions yield

$$\Bigg\langle \underbrace{\bigcirc \sqcup \ldots \sqcup \bigcirc}_{k \text{ times}} \Bigg\rangle = (-A^2 - A^{-2})^{k-1} \tag{1}$$

Hence, it is clear that $\langle D \rangle$ will be a summation over every possible way of smoothing all the crossings of D. We will now obtain the precise expression. Let $\mathcal{C}$ be the set of crossings of D and consider all maps $s : \mathcal{C} \to \{-1, 1\}$. We define sD to be the diagram obtained by smoothing the crossings of D

as per s. By this we mean that any crossing c is decomposed as follows.



Hence sD is a disjoint union of simple closed curves, so $\langle sD\rangle = (-A^2 - A^{-2})^{|sD|-1}$ (by (1)) where $|sD|$ is the number of simple closed curves in sD. Now it can be seen by induction on the size of $\mathcal{C}$ that

$$\langle D\rangle = \sum_{s\colon \mathcal{C}\to\{\pm 1\}} A^{\sum\limits_{c\in\mathcal{C}} s(c)} (-A^2 - A^{-2})^{|sD|-1} \tag{2}$$

This expression will be useful in §4 when we consider the Jones polynomial of alternating links in particular. Direct applications of the defining conditions of the Kauffman bracket yields the following.

Lemma 3.1. *The Kauffman bracket is invariant under R-II and R-III moves.*

Proof. We will give the proof for invariance under R-II moves, and the calculation for R-III moves is similar. First, we apply the third rule for smoothing crossings as in fig. 6.

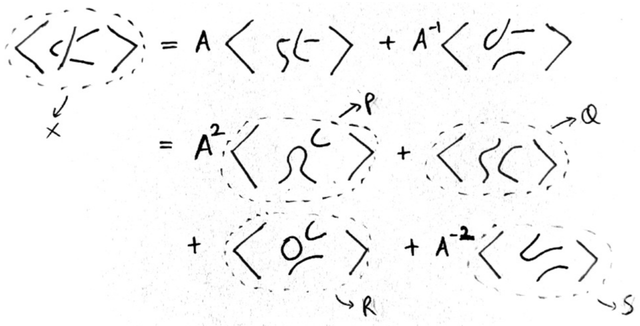


Figure 6: smoothing the crossings using the third rule.

Now, we have

$$\begin{aligned} P &= S \\ R &= (-A^2 - A^{-2})P \text{ (using the second rule)} \end{aligned}$$

Hence, we obtain

$$\begin{aligned} X &= A^2P + Q + (-A^2 - A^{-2})P + A^{-2}P \\ &= Q \end{aligned}$$

This proves that the Kauffman bracket is invariant under R-II moves. ☐

Under R-I moves, although the Kauffman bracket is not invariant, it varies in a simple way.

Lemma 3.2. *Under R-I moves, the Kauffman bracket changes according to fig. 7.*

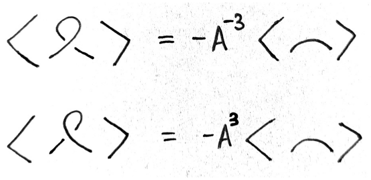


Figure 7: The Kauffman bracket under R-I moves.

We noted in §1 that for an *oriented* link L with diagram D, the writhe $w(D)$ of D is invariant under R-II and R-III moves, and under R-I moves it changes by ±1 (depending on the sign of the crossing added/removed). Hence, using **lemma 3.1** and **lemma 3.2** we see that

$$(-A)^{-3w(D)} \langle D \rangle$$

is invariant under all three Reidmeister moves. It is straightforward to see that the above expression only has terms involving even powers of A, so setting $t = A^{-4}$ (i.e. $A^2 = t^{-\frac{1}{2}}$) in the above expression yields a Laurent polynomial $V(L)$ in $t^{\frac{1}{2}}$ which depends only on the (oriented) link L and not the precise choice of diagram used for computation. $V(L)$ is called the **Jones polynomial** of the oriented link L.

Theorem 3.3. *The Jones polynomial* $V(L) \in \mathbb{Z}[t^{\frac{1}{2}}, t^{-\frac{1}{2}}]$ *defined by*

$$V(L) := \left[(-A)^{-3w(D)} \langle D \rangle\right]_{A^2 = t^{-\frac{1}{2}}}$$

is a link invariant.

$V(L)$ does not change if the orientations of all components of L are (simultaneously) reversed. In particular, if K is a knot then $V(K)$ is independent of choice of orientation on K. Also, if L has odd many components then $V(L)$ is a Laurent polynomial in t.

It is straightforward to show that for knots K_1 and K_2, we have

$$V(K_1 + K_2) = V(K_1)V(K_2)$$

Likewise, for (oriented) links L_1 and L_2 we have

$$V(L_1 \sqcup L_2) = (-A^2 - A^{-2})V(L_1)V(L_2)$$

These yield the following corollaries.

Corollary 3.4. *For a knot K, if $V(K)$ is an irreducible element in the ring $\mathbb{Z}[t^{\frac{1}{2}}, t^{-\frac{1}{2}}]$ then K is a prime knot.*

Corollary 3.5. *For an oriented link L, if $V(L)$ is not divisible by $t+1$ (in the ring $\mathbb{Z}[t^{\frac{1}{2}}, t^{-\frac{1}{2}}]$) then L is not a split link.*

Proof. $V(L)$ is divisible by $t+1$ if and only if it is divisible by $-t^{\frac{1}{2}} - t^{-\frac{1}{2}}$. Now use the above discussion. □

4 The Jones polynomial of alternating links

Throughout this section, L will denote a link with diagram D. Let the set of crossings of D be $\mathcal{C}$, with $|\mathcal{C}| = n$. Let $s_+, s_- : \mathcal{C} \to \{-1, 1\}$ be the constant states which are identically 1 and -1 respectively. The components of $S^2 \backslash D$ (the 'regions' of D) can be coloured black and white in a chessboard pattern, i.e. the intersection of the (topological) boundaries of any two regions with the same colour is a discrete set of points (which are all crossings). If D is connected (as a topological space), then it can be viewed as a connected 4-regular graph embedded in S^2 with vertex set $\mathcal{C}$. Hence, the number of regions of D can be seen to be $n+2$ using Euler's characteristic formula for S^2. When D is also alternating, all components of s_+D bound regions of one colour (say black) and those of s_-D bound the white regions. Hence, we obtain the following.

Lemma 4.1. *If D is connected and alternating, then*

$$|s_+D| + |s_-D| = n + 2$$

Something similar is true for all connected diagrams.

Lemma 4.2. *If D is connected, then*

$$|s_+D| + |s_-D| \leq n + 2$$

Proof. We will use induction on n. The base case of $n = 0$ is obvious. For $n > 0$, let c be any crossing. At least one way of smoothing c keeps D connected, say the positive way, and let D' be the diagram obtained after c is smoothened this way. Hence, $s_+D = s_+D'$ and $|s_-D| = |s_-D'| \pm 1$. This yields

$$\begin{aligned} |s_+D| + |s_-D| &= |s_+D'| + |s_-D'| \pm 1 \\ &\leq n + 1 \pm 1 \text{ (using the induction hypothesis on } D'\text{)} \\ &\leq n + 2 \end{aligned} \qquad \square$$

These observations will help us examine the highest and lowest powers of t that appear in $V(L)$ (with L given an arbitrary orientation). We give names to these quantities.

Definition 4.3. For a $\mathbb{Z}$-linear combination p of rational powers of an invariant x, let $M_x(p)$ and $m_x(p)$ be the largest and smallest powers of x that appear in p (with non-zero coefficient) respectively. The breadth $B_x(p)$ of p is the difference $M_x(p) - m_x(p)$.

Now, since we have

$$V(L) = \left[(-A)^{-3w(D)} \langle D \rangle\right]_{A^2 = t^{-\frac{1}{2}}}$$

it follows that

$$B_t(V(L)) = \frac{B_A(\langle D \rangle)}{4} \tag{3}$$

Hence the breadth of $\langle D \rangle$ depends only on the (unoriented) link L. Recall (2), the explicit formula we obtained for the Kaufmann bracket.

$$\langle D \rangle = \sum_{s\colon \mathcal{C} \to \{\pm 1\}} A^{\sum\limits_{c \in \mathcal{C}} s(c)} (-A^2 - A^{-2})^{|sD|-1}$$

Hence $M_A(\langle D \rangle)$ is at most the maximum value of the expression

$$\sum_{c \in \mathcal{C}} s(c) + 2|sD| - 2 \tag{4}$$

as s varies over all states. Likewise $m_A(\langle D \rangle)$ is at most the minimum value of the expression

$$\sum_{c \in \mathcal{C}} s(c) - 2|sD| + 2 \tag{5}$$

as s varies over all states.

The maximum possible value of $\sum\limits_{c \in \mathcal{C}} s(c)$ is n, and that is attained only when $s = s_+$. Likewise, the minimum possible is $-n$, which is attained only when $s = s_-$. Hence, it would be convenient if we could find some class of diagrams for which (4) is maximised for $s = s_+$ and (5) is minimised for $s = s_-$. We will now focus on analysing (4), and the same method can be repurposed to analyse (5) as well.

For an arbitrary state s, we can construct a sequence of states

$$s_+ = s_0, s_1, \ldots, s_{k-1}, s_k = s$$

such that for every $i \in \{0, \ldots, k-1\}$ the values of s_i and s_{i+1} differ at exactly one crossing (say c_i) and $s_i(c_i) = 1, s_{i+1}(c_i) = -1$. It is easy to see that $|s_{i+1}D| = |s_iD| \pm 1$, since changing the way one crossing is smoothened either joins two different simple closed curves to make one bigger curve or vice-versa. Hence, we see that

$$\begin{aligned} \sum_{c\in\mathcal{C}} s_i(c) + 2|s_iD| - 2 &= \left(\sum_{c\in\mathcal{C}} s_{i+1}(c) - 2\right) + 2\left(|s_{i+1}D| \pm 1\right) - 2 \\ &\geq \sum_{c\in\mathcal{C}} s_{i+1}(c) + 2|s_{i+1}D| - 2 \end{aligned} \tag{6}$$

Hence, it follows that the maximum value of (4) is attained for $s = s_+$ for *every* diagram D, without any assumptions. This, together with an analogous computation for a lower bound on (5), is summarised as follows.

Lemma 4.4. *We have*

$$\begin{aligned} M_A(\langle D\rangle) &\leq n + 2|s_+D| - 2 \\ m_A(\langle D\rangle) &\geq -n - 2|s_-D| + 2 \\ B_A(\langle D\rangle) &\leq 2n + 2|s_+D| + 2|s_-D| - 4 \end{aligned}$$

Corollary 4.5. *If D is connected, then*

$$B_A(\langle D\rangle) \leq 4n$$

Proof. Use **lemma 4.2** and **lemma 4.4**. □

The first inequality of **lemma 4.4** would be an equality if (6) is a strict inequality for at least one i whenever $s \neq s_+$. An analogous condition is obtained for the second inequality in **lemma 4.4**. In particular, if D is adequate (according to the following definition) then all the inequalities of **lemma 4.4** will be equalities.

Definition 4.6. Say that D is plus-adequate if $|sD| < |s_+D|$ whenever s is a state which takes the value -1 at exactly one crossing. Likewise, say that D is minus-adequate if $|sD| < |s_-D|$ whenever s is a state which takes the value 1 at exactly one crossing. Say that D is adequate if it is both plus- and minus-adequate.

The preceding observations yield the following.

Lemma 4.7. *If D is plus-adequate, then*

$$M_A(\langle D\rangle) = n + 2|s_+D| - 2$$

If D is minus-adequate, then

$$m_A(\langle D\rangle) = -n - 2|s_-D| + 2$$

If D is adequate, then

$$B_A(\langle D\rangle) = 2n + 2|s_+D| + 2|s_-D| - 4$$

Hence, if D is alternating, connected and adequate then we can explicitly find the link invariant $B_A(\langle D\rangle)$ using what we have observed so far.

Proposition 4.8. *If D is alternating, connected and adequate then*

$$B_A(\langle D\rangle) = 4n$$

Proof. Use **lemma 4.7** and **lemma 4.1**. ☐

It is straightforward to check whether a given diagram is alternating and connected, but adequacy seems rather contrived. To find a more natural way to frame the condition of adequacy, note that D is plus-adequate if and only if no component of s_+D abuts itself in the neighbourhood of a former-crossing. Likewise, D is minus-adequate if and only if no component of s_-D abuts itself in the neighbourhood of a former-crossing.

For D alternating, suppose we colour the regions of D using the chessboard colouring described at the beginning of this chapter so that every region bounded by a component of s_+D is black and every region bounded by a component of s_-D is white. Hence, D is plus-adequate if and only if there is no crossing $c \in \mathcal{C}$ such that the two black regions which meet c are the same. Likewise, D is minus-adequate if and only if there is no crossing $c \in \mathcal{C}$ such that the two white regions which meet c are the same. Hence, D is adequate if and only if for every crossing $c \in \mathcal{C}$, the four regions which meet at C are all distinct. Hence, D is adequate if and only if there is no crossing in D of the form shown in fig. 8 (or its reflection). Such a crossing is called a **reducible** crossing.

Figure 8: A reducible crossing. Here, A and B are parts of D which do not have any crossings in common.

More precisely, a crossing c is reducible if and only if there is a simple closed curve $\gamma \subset S^2$ which intersects D transversally at precisely two points, one of which lies on the under-strand at c and the other on the over-strand at c. D is said to be **reduced** if it has no reducible crossing. Hence, we have the following.

Lemma 4.9. *If D is alternating, then D is adequate if and only if D is reduced.*

Corollary 4.10. *If D is alternating, connected and reduced then*

$$B_A(\langle D \rangle) = 4n$$

Proof. Use **proposition 4.8** and **lemma 4.9**. □

Note that any alternating diagram for a link can be converted into a reduced and alternating diagram by removing all the reducible crossings ('reducing' the diagram) — simply twist the part of the diagram on either side of the reducible crossing. Now, comparing **corollary 4.5** and **corollary 4.10**, we see that if D is alternating, connected and reduced and D' is any connected diagram for L, then D has no more crossings than D'. Hence, we have the following.

Theorem 4.11. *If D is alternating, connected and reduced, then n is the least number of crossings any connected diagram of L can have.*

If every diagram of L is connected and D is as in the above theorem, then by the above theorem the crossing number of L is n. It is clear that every diagram of L is connected if and only if L is not a split link as per the following definition.

Definition 4.12. L is said to be **split** if there exists a 2-sphere F disjoint from L such that both the 3-balls F bounds intersect L.

Corollary 4.13. *If L is not a split link and D is alternating, connected and reduced, then n is the crossing number of L.*

In particular knots are non-split links, so the above applies when L is an alternating knot. If every pair of (distinct) components K_1 and K_2 of L have non-zero linking number then L is not split. More generally, if for every pair of components J and K of L one can find a sequence of components $J = K_0, K_1, \ldots, K_{r-1}, K_r = K$ of L such that K_i and K_{i+1} have non-zero linking number for $i = 0, \ldots, r-1$ then L is not split. Of course, not every non-split link is of this form — the Borromean rings, for instance.

If we can show that D being alternating and connected implies that L is not split,[4] then the condition of L being non-split can be dropped from **corollary 4.13** (cf. **theorem 5.7**). This motivates **theorem 5.5**, one of the two main results of §5.

Now that we have established that diagrams which are alternating, connected and reduced are special (at least among connected diagrams), one might ask just how special they are. More concretely, the question is the following.

Suppose D is alternating, connected and reduced, and D' is some non-alternating and connected diagram for L. Must D' have strictly *more crossings than D?*

We will answer this question partially, but first we need a definition.

Definition 4.14. Say that D is **strongly prime** if every simple closed curve in S^2 which intersects D transversally at exactly two points (which are not crossings) bounds, on one side of it, a disc $\Delta \subset S^2$ such that $\Delta \cap D$ is an (crossing-free) arc.

Using methods similar to those of this section, one can prove the following counterpart to **lemma 4.1**.

---

[4] This is not an entirely unreasonable conjecture — to force a disconnected diagram to be connected one needs to perform a (crossing-increasing) R-II move, which will ensure that the diagram becomes non-alternating. Hence the conjecture is that one cannot get rid of this non-alternating nature by further 'complicating' the diagram (while still keeping the diagram connected).

Lemma 4.15. *If D is non-alternating, connected and strongly prime then*

$$|s_+D| + |s_-D| < n + 2$$

The above, together with **lemma** 4.4, yields the following.

Corollary 4.16. *If D is non-alternating, connected and strongly prime, then*

$$B_A(\langle D \rangle) < 4n$$

Comparing the above with **corollary** 4.10 and using **theorem** 4.11, we obtain the following partial answer to the question from before.

Proposition 4.17. *If L has an alternating, connected and reduced diagram D with n crossings and a non-alternating, connected and strongly prime diagram D' with n' crossings, then*

$$n < n'$$

5 The geometry of alternating links

In this section, we will look at some properties (splitness and primality) which an alternating link has if and only if its alternating diagram has. Primality for a link is defined as follows.

Definition 5.1. A link L is said to be **prime** if every 2-sphere which intersects L transversally at two points bounds a ball B with $(B, B \cap L)$ a trivial ball-arc pair. If L' is a non-prime link, then we say that a 2-sphere F demonstrates the non-primality of L' if F intersects L' transversally at two points and bounds no ball B with $(B, B \cap L')$ a trivial ball-arc pair.

Note that this definition agrees with the definition of primeness for knots (cf. the discussion before the proof of **theorem 2.2**). Primality and splitness for link diagrams are defined analogous to how they are defined for links.

Definition 5.2. A link diagram D on S^2 is said to be **split** if there is a simple closed curve in S^2 not intersecting D such that both the discs it bounds (in S^2) intersect D. Such a simple closed curve is called a splitting curve for D.

Remark 5.3. D is split if and only if it is not connected.

Definition 5.4. A link diagram D on S^2 is said to be **prime** if every simple closed curve which intersects D transversally at two points (which are not crossings) bounds a disc $\Delta \subset S^2$ such that $(\Delta, \Delta \cap D)$ is a diagram of a trivial ball-arc pair, i.e. the one-point compactification of $D \cap (\Delta \backslash \partial \Delta)$ as a subset of the one-point compactification of $\Delta \backslash \partial \Delta$ is a diagram (with under/over information inherited from D in the obvious way) for the unknot.

The theorems we will prove are the following.

Theorem 5.5. *An alternating link L with alternating diagram D is split if and only if D is split.*

Theorem 5.6. *An alternating link L with alternating diagram D is prime if and only if D is prime.*

As noted in §4, **corollary 4.13** and **theorem 5.5** yield the following corollary, which is important enough to be called a theorem.

Theorem 5.7. *If a link L has an alternating, connected and reduced diagram D with n crossings, then n is the crossing number of L.*

The similarities between **theorem** 5.5 and **theorem** 5.6 are more apparent if we rephrase **theorem** 5.6 as follows.

Theorem 5.8. *An alternating link L with alternating diagram D is non-prime if and only if D is non-prime.*

This is because for demonstrating both splitness and non-primality (of a link or a diagram) we need to produce a 'separation' of some kind — a 2-sphere or simple closed curve which separates components in the case of splitness, and likewise for non-primality. Now, suppose we are given a 2-sphere F for an alternating link L (with alternating diagram D) which demonstrates its splitness or non-primality. To then conclude something about D, we would of course need to bring L and F to 'nice' positions with respect to the projection sphere S^2 so that the intersection of S^2 with F gives us the requisite simple closed curve to demonstrate the splitness or non-primality of D.

Bringing L to a nice position is done in a straightforward manner. At each crossing in D place a bubble, which is a small (round) ball centred at the crossing. Fix one side of S^2 (the projection sphere) in S^3 as the 'outside' and one as the 'inside'. The part of a bubble which is outside S^2 will be the north and the part inside S^2 will be the south. Modify D so that at every crossing (with bubble b placed on it), the over-strand goes along the northern hemisphere of ∂b and the under-strand goes along the southern hemisphere of ∂b. This modified D is now a union of disjoint simple closed curves in S^3 (i.e. a link), and it is precisely L.

By removing from S^2 its intersection with the bubbles and replacing the holes with the northern hemispheres of the boundaries of the corresponding bubbles, we obtain a new 2-sphere S_+. Likewise, using the southern hemispheres instead will yield a 2-sphere S_-. We have $S_+ \cap S_- \subset S^2$ and $L \subset S_+ \cup S_-$. The spheres S_+ and S_- respectively bound balls B_+ and B_- with $B_+ \cap B_- = S_+ \cap S_-$. The union $B_+ \cup B_-$ covers all of S^3 save the interiors of all bubbles.

Now for any surface F (not necessarily a 2-sphere) which intersects L transversally we state precisely is meant by the nice position previously alluded to. We will later elaborate on what we mean by 'saddles' in the following conditions. F will be said to be in standard position if it satisfies the following.

(A) F intersects S_+ and S_- transversally in a disjoint union of simple closed curves.

(B) F intersects each bubble in disjoint saddles.

(C) F intersects the north-south axis of each bubble transversally.

(D) $F \cap B_+$ and $F \cap B_-$ are unions of disjoint discs.

(E) Every component of $F \cap S_+$ and $F \cap S_-$ intersects every bubble at most once.

(F) Every component of $F \cap S_+$ and $F \cap S_-$ intersects L or some bubble.

Here, a saddle in a bubble b is a disc Δ such that

(i) $\partial\Delta$ separates the two components of $L \cap \partial b$ in ∂b.

(ii) $\partial\Delta$ intersects each hemisphere (north and south) in two disjoint arcs parallel to the component of $L \cap \partial b$ in that hemisphere.

Remark 5.9. As a result of (i), the two arcs of $\partial\Delta$ in each hemisphere must be on different sides of the arc of $L \cap \partial b$ in that hemisphere.

Here condition (i) is purely topological whereas (ii) is more geometric in nature. As the name suggests, saddles may be visualised like the classical saddle-points one sees in 3-dimensional plots of smooth functions on two real variables. Note that the conditions for standard position do not require that F intersect the projection sphere S^2 transversally.

Lemma 5.10. *For L a split link with non-split diagram D, a splitting 2-sphere in standard position exists.*

Proof. Let F be any splitting 2-sphere for L, and we will modify F so that it satisfies all the conditions for standard position.

(A) Small changes can be made to F through isotopy so that it intersects S_+ and S_- transversally.

(B) It suffices to show that F can be modified so that it intersects each bubble b in a disjoint union of discs, with each one's boundary separate (in ∂b) the two components of $L \cap \partial b$. First, suppose there is some component of

$F \cap \partial b$ which does not bound a disc component of $F \cap b$. Among all such components of $F \cap \partial b$, let C be innermost on ∂b. Let $\Delta \subset \partial b$ be the disc it bounds in ∂b with respect to which it is innermost. Hence F may intersect the interior of Δ, but every component in that intersection bounds a disc component of $F \cap b$. Shift Δ into b (while shifting its boundary along F) to obtain a new disc Δ' with $F \cap \Delta' = \partial \Delta'$, which is a simple closed curve parallel to C in F. If Δ_1 and Δ_2 are the two discs in F bounded by $\partial \Delta'$, then at least one of the two spheres $\Delta' \cup \Delta_i$ is a splitting 2-sphere for L. Hence let F now be this new splitting 2-sphere. It is clear that F now has fewer components in $F \cap \partial b$ which do not bound disc components of $F \cap b$ (since the component of $F \cap b$ which C bounded before had disconnected boundary). Furthermore, the intersection of F with any bubble has not increased. Hence we may perform the above procedure finitely many times and modify F so that $F \cap b$ is a disjoint union of discs.

Now, suppose $F \cap \partial b$ has a component C which does not separate the two components of $L \cap \partial b$. Among all such components let C be innermost on ∂b, and so C is in fact innermost among all components of $F \cap \partial b$. By the previous paragraph, C bounds a disc component $\Delta \subset F \cap b$. Now F can be modified by simply 'pulling' Δ out of b by isotopy so that C is removed from $F \cap \partial b$. As before, this procedure can be performed finitely many times to obtain a modified F which is as desired.

(C) Follows by making small changes to F by isotopy.

(D) The procedure used to modify F is identical to the first paragraph for condition (A), with B_+ and B_- respectively playing the same role as b.

(E) Let b be a bubble with northern hemisphere N, and suppose some component C of $F \cap S_+$ intersects N more than once, i.e. $C \cap N$ has distinct arc components λ_1 and λ_2. Note that $\partial N = \partial b \cap S^2$ and λ_1 and λ_2 have endpoints in ∂N. We may choose C, λ_1 and λ_2 so that λ_1 and λ_2 are consecutive on N, in the sense that the component of $N \backslash (\lambda_1 \cup \lambda_2)$ whose closure contains $\lambda_1 \cup \lambda_2$ does not intersect with F. Now consider two cases.

Case I: λ_1 and λ_2 are on opposite sides of $L \cap N$ in N.
In this case λ_1 and λ_2 are in the boundary of the same saddle, say σ. Pick points $x \in \lambda_1$ and $y \in \lambda_2$ and let γ be the simple closed curve which first

follows one ‘half’ of C from x to y and then returns to x along σ. On one hand, γ is a meridian of L. On the other hand, γ is contained entirely in F and so is null-homotopic in $S^3\backslash L$. This is a contradiction, so the hypothesis of this case cannot be true.

Case II: λ_1 and λ_2 are on the same side of $L \cap N$ in N.
Let σ_1 and σ_2 be the saddles in b whose boundaries contain λ_1 and λ_2 respectively (so $\sigma_1 \neq \sigma_2$). Let p_1 and p_2 be the points where σ_1 and σ_2 intersect the north-south axis of b. Let γ be the simple closed curve which follows the following sequence of arcs.

(a) An arc from p to some (arbitrary) point $x \in \lambda_1$ along σ_1.

(b) One ‘half’ of C to some (arbitrary) point $y \in \lambda_2$.

(c) An arc from y to q along σ_2.

(d) The segment of the north-south axis of b from q to p.

For simplicity, x and y can be chosen as ‘midpoints’ of λ_1 and λ_2 respectively. Now γ is a union of two arcs with common endpoints, one of which is contained in F and another which intersects F at only its endpoints. Using condition (D), one can also find a disc bounded by γ which does not intersect L and intersects F only in the arc $\gamma \cap F$. Hence, F can be ‘pushed through’ this disc using an isotopy which removes the intersection points p and q of F with the north-south axis of b. This modification can be done in such a way that

- all of F except for some regular neighbourhood of $\gamma \cap F$ in $(B_+ \cup b) \cap F$ is not changed.
- In place of the two components σ_1 and σ_2 of $F \cap b$ we now have a single component.
- F is still a splitting 2-sphere for L.

Now the procedure for (B) and (D) can be repeated on F so that those conditions are satisfied, and then the above procedure can be performed finitely many times so that F satisfies (E) with respect to S_+. Then the same procedure is repeated with respect to S_-.

(F) If some component $C \subset F \cap S_+$ does not intersect any bubble, then it must be contained in S^2. It bounds discs in both B_+ and B_-, and so F is precisely the union of these discs. However now it can be seen that C is a splitting S^1 for D, a contradiction. ☐

Lemma 5.10 concludes the topological part of the proof of **theorem** 5.5. The combinatorial part is as follows.

Proof of ***theorem 5.5****.* It is clear that if D is split, then so L. If D is not split but L is split, then by **lemma** 5.10 there is a splitting 2-sphere F for L in standard position. $F \cap S_+$ cannot be empty, so let $C \subset F \cap S_+$ be a component innermost on S_+. Give an arbitrary orientation to C. In a regular neighbourhood of S^2 which contains all the bubbles, let π be the projection map onto S^2. Hence we recover D from L as $D = \pi(L)$.

C does not intersect L, so $\pi(C)$ can intersect D only near crossings (this will correspond to C going through the northern hemisphere of some bubble). We also have $\pi(C) \cap \pi(L \cap S_+) = \emptyset$. Now $\pi(L \cap S_+)$ is precisely what we get if we break the under-strands at all the crossings in D, which means that intersections of $\pi(C)$ with D correspond to C going through the gaps made by these breaks.

We also know that C intersects some bubble (by condition (F)), so $\pi(C)$ intersects D at least once. By considering the regions of $S^2 \backslash D$, we see that $\pi(C)$ intersects D an even number of times (in particular, at least twice). Since $\pi(C)$ has to pass through the gaps made by the breaks in the under-strands at these intersections, for any two consecutive intersections we see that the over-strand is to the left of $\pi(C)$ at one of them whereas at the other it is to the right. This means that there is at least one saddle σ_1 to the left of C and at least one saddle σ_2 to its right. By condition (E), we have

$$\partial\sigma_i \cap S_+ \neq C \cap \sigma_i \text{ for } i = 1, 2$$

Hence, by considering the components of $\partial\sigma_1 \cap S_+$ and $\partial\sigma_2 \cap S_+$ which are not in C we see that there are components of $S_+ \cap F$ on both the discs bounded by C in S_+. This contradicts the fact that C was innermost on S_+. ☐

The proof of **theorem** 5.8 proceeds similarly.

Lemma 5.11. *For L a non-split and non-prime link with diagram D, a 2-sphere in standard position demonstrating the non-primality of L exists.*

Proof. Let F be any 2-sphere demonstrating the non-primality of L. Using a straightforward isotopy if necessary, we may assume that the two points of $F \cap L$ are not contained in (the boundary of) any bubble. Now, the modifications to F required for satisfying the conditions of standard position are as follows.

(A) Same as in the proof of **lemma 5.10**.

(B) First we make an observation. Suppose Δ' is a disc which does not intersect L and whose boundary is in F. Let Δ_1 and Δ_2 be the two discs in F which share boundary with Δ'. Hence both points of $F \cap L$ are in one of the Δ_i's (say Δ_1). Since L is not split, one of the balls bounded by $\Delta' \cup \Delta_2$ (in particular, that ball which does not contain Δ_1) does not intersect L. Hence $\Delta' \cup \Delta_1$ also demonstrates the non-primality of L. Using this observation, F can be modified so that it satisfies condition (B) as was done in the proof of **lemma 5.10**.

(C) Same as in the proof of **lemma 5.10**.

(D) Use the observation made above for condition (B) and mimick the procedure used in the proof of **lemma 5.10**.

(E) Let $b, N, C, \lambda_1, \lambda_2$ be as in the proof of **lemma 5.10**. Here too λ_1 and λ_2 can be chosen to be consecutive on N. Hence we have the same two cases.

Case I: λ_1 and λ_2 are on opposite sides of $L \cap N$ in N.
As before, λ_1 and λ_2 are in the boundary of the same saddle, say σ. Let x, y, γ be as in the proof of **lemma 5.10**. We lift $\gamma \cap S_+$ slightly into B_+ to ensure that γ does not intersect L. As before, γ is a meridian of L — let Δ be a disc bounded by γ which intersects L transversally in one point and

- $\Delta \cap F = \gamma$.
- $\Delta \cap S_+$ equals $\Delta \cap \partial b$, which is an arc on ∂b from x to y.

Let Δ_1 and Δ_2 be the discs in F bounded by γ, so that $\Delta \cup \Delta_1$ and $\Delta \cup \Delta_2$ are 2-spheres. Both of them intersect L at least once and at most thrice, so they both intersect L exactly twice. At least one of them demonstrates the non-primality of L (say $\Delta \cup \Delta_1$), so we may let $\Delta \cup \Delta_1$ be the new (modified) F. Now every saddle in F except σ was either removed or left untouched by this modification of F, and no new saddles were added. Furthermore, what remains of σ is part of a disc in $F \cap b$ which does not separate the two components of $L \cap \partial b$, so the procedure from condition (B) can be applied to obtain a new F with one saddle lesser.

Case II: λ_1 and λ_2 are on the same side of $L \cap N$ in N.
The procedure is the same as that of case II in the proof **lemma 5.10** (condition (E)), but as in case I we may lift $\gamma \cap S_+$ slightly into B_+ to ensure that γ does not intersect L.

(F) If some component $C \subset F \cap S_+$ does not intersect any bubble, then it must be contained in S^2. It bounds discs in both B_+ and B_-, and so F is precisely the union of these discs. However now it can be seen that C must intersect L, since F does. □

*Proof of **theorem 5.8**.* It is clear that if D is non-prime, then so is L. Now suppose L is non-prime. If L is split, then so is D (by **theorem 5.5**) and so it is straightforward to see that D is non-prime. Hence suppose L is not split, and so D is also not split (by **theorem 5.5**). By **lemma 5.11**, there is a 2-sphere F in standard position which demonstrates the non-primality of L. Let $C \subset F \cap S_+$ be a component innermost on S_+ and give it an arbitrary orientation. Recall the projection π onto S^2 from the proof of **theorem 5.5**.

$\pi(C)$ must intersect D at least once (since C intersects L or some bubble). Also, they intersect an even number of times as noted in the proof of **theorem 5.5**. No two consecutive intersections of $\pi(C)$ with D can correspond to intersections of C with some bubbles, since that would lead to a contradiction exactly as in the proof of **theorem 5.5**. Hence it cannot be that C does not intersect L (in particular, every component of $F \cap S_+$ which is innermost on S_+ must intersect L). We now take two cases.

Case I: C intersects L at two points.
Every component of $F \cap S_+$ other than C cannot intersect L. Hence, if $F \cap S_+ \neq C$ then we can pick a component of $\subset F \cap S_+$ which is innermost on S_+ and does not intersect L (a contradiction). Hence we have $F \cap S_+ = C$.

Now, if C does not intersect any bubble then it follows that $C = \pi(C)$ is a simple closed curve which demonstrates the non-primality of D, and so we are done. On the other hand, if C intersects some bubble then let σ be a saddle which C intersects. Hence (by condition (E)) there is some component $C' \subset F \cap S_+$ different from C which also intersects σ (a contradiction).

Case II: C intersects L at only one point.
By the paragraph before case I, we see that C intersects exactly one bubble (and hence exactly one saddle, say σ). Hence the arc $C \backslash \sigma$ is contained in S^2. Hence there is a component $C' \subset F \cap S_-$ which contains the arc $C \backslash \sigma$. We now see that C' contains an endpoint from both the arc-components of $\sigma \cap S_-$, so C' contains both the arc-components of $\sigma \cap S_-$. This contradicts condition (E). ☐

Definition 5.12. A surface F, other than a 2-sphere, contained in a 3-manifold M is said to be incompressible if any disc $\Delta \subset M$ which spans F (i.e. $\Delta \cap F = \partial\Delta$) has the property that $\partial\Delta$ bounds a disc in F. A 2-sphere is incompressible in M if it does not bound a ball.

The method of the above proofs — first bringing a surface F into standard position and then examining the intersections of $\pi(F \cap S_+)$ and $\pi(F \cap S_-)$ with D — can also be used to prove the following.

Theorem 5.13. *Let L be a non-split, prime, alternating link and F be a closed incompressible surface in $S^3 \backslash L$. There exists a disc $\Delta \subset S^3$ spanning F which intersects L transversally at exactly one point.*

Corollary 5.14. *Let L be as above. If T is an incompressible torus in $S^3 \backslash L$, then it is parallel to a solid torus neighbourhood of some component of L.*

Corollary 5.14, together with the theory of existence of hyperbolic structures on 3-manifolds developed by W.P. Thruston, shows that the complement of 'almost' every non-split, prime, alternating link in S^3 has a complete hyperbolic structure of finite volume.[5]

---

[5] The non-split, prime, alternating links for which this result is not true are all twist links.

6 Knot tabulation and Dowker-Thistlethwaite notation

When tabulating knots using their diagrams, the very first challenge one encounters is listing all possible knot diagrams with a given number of crossings. One tool that can be used for this is the Dowker-Thistlethwaite notation for a diagram, which is defined as follows. Start with a knot diagram D with n crossings. Give an arbitrary orientation on D and follow along it, starting at some chosen basepoint $p_0 \in D$ (not a crossing). Label the crossings $1, 2, \ldots, 2n$ in the order that they are reached. Hence, every crossing is assigned two numbers, one for the under-strand and one for the over-strand. This gives an involution $a : [2n] \to [2n]$ (with no fixed points) such that for every $i \in [2n]$, the two numbers assigned to the crossing labelled i are i and a_i. It turns out that $i \equiv a_i + 1 \pmod 2$ for every $i \in [2n]$ (this will be shown later, cf. **lemma 6.1**), and so a is completely determined by the sequence

$$[a_1, a_3, \ldots, a_{2n-1}]$$

which is simply some permutation of $\{2, 4, \ldots, 2n\}$. To further encode the over/under information at each crossing, every number in the above sequence which corresponds with an under-strand (at the respective crossing) is decorated with a minus sign. The Dowker-Thistlethwaite notation for D, then, is the sign-decorated sequence

$$[\pm a_1, \pm a_3, \ldots, \pm a_{2n-1}]$$

which is obtained by choosing p_0 and the orientation on D so that the sequence

$$[a_1, a_3, \ldots, a_{2n-1}]$$

is lexicographically minimum. Of course, this notation helps in the goal of listing all diagrams with a given number of crossings only if it is known that no two diagrams can have the same notation, and also how the diagram can be reconstructed using the notation. This is done in [2], and here we provide a sketch of the argument.

Lemma 6.1. *With the involution $a : [2n] \to [2n]$ as above (not necessarily satisfying the lexicographic minimality condition), we have*

$$i \equiv a_i + 1 \pmod 2$$

for every $i \in [2n]$.

Proof. This result is stated without proof in [1] and [2]; the following proof is my own. Fix $i \in [2n]$ and let c be the crossing with labels i and a_i. Without loss of generality, assume $i < a_i$. Following D along the chosen orientation starting at the i-th crossing, we get a (possibly self-intersecting) loop $\gamma : [0,1] \to S^2$ at c which passes through the crossings with labels $i, i+1, \ldots, a_i - 1, a_i$ (in that order). Likewise, following D starting at the a_i-th crossing yields a (possibly self-intersecting) loop $\lambda : [0,1] \to S^2$ at c which passes through the crossings with labels $a_i, a_i+1, \ldots, 2n, 1, 2 \ldots, i$ (in that order).

We wish to show that γ passes through an even number of crossings (counting multiplicity). Every self-intersection of γ is a crossing which γ passes through twice, so it suffices to show that $\gamma|_{(0,1)}$ and $\lambda|_{(0,1)}$ intersect an even number of times. For this, consider γ and λ as a knot diagrams (in which c is no longer a crossing). By moving γ and λ away from each other near c, we see that $\lambda \cup \gamma$ now becomes a diagram for a link with two components — γ and λ are diagrams of the two components respectively.[6] Now we know that there are an even number of crossings in the diagram $\gamma \cup \lambda$ in which one strand is from γ and the other from λ (this is precisely the same as saying that $\mathrm{lk}(\gamma, \lambda)$ is an integer, which we saw in §1 using a crossing-smoothing argument). ☐

Definition 6.2. Say that two link diagrams D_1 and D_2 are **related** if they differ at most in the over/under information they carry. In this case, they are called **relatives** of each other.

Corollary 6.3. *Any knot diagram (with at least one crossing) has precisely two relatives which are alternating.*

Proof. By **lemma 6.1**, any relative of a knot diagram D is alternating if and only if either all of its odd-labelled crossings are over-crossings or under-crossings. ☐

A similar but unrelated result that can be proved using the involution a is as follows.

Proposition 6.4. *Every knot diagram has some relative which is a diagram of the unknot.*

---

[6] We have essentially smoothened the crossing c.

Proof. For any knot diagram D, the relative with the property that whenever $i < a_i$, the i-th crossing is an over-crossing is the desired relative. $\square$

For the rest of this section, let S be a given (arbitrary) involution of $[2n]$ (for some $n \geq 1$) with no fixed points which we will denote in one-line notation as

$$S = [a_1\, a_2\, \ldots\, a_{2n}]$$

For the knot tabulation problem mentioned above, we will try to answer the question of when is it that S 'comes from' a knot diagram in the way described at the beginning of this section, and if so whether the corresponding diagram is unique upto homeomorphism of S^2. First we give some definitions to state this question precisely.

Definition 6.5. Say that S is **realisable** if there is a PL map $\rho : [0, 2n] \to S^2$ such that

- $\rho(0) = \rho(2n)$.
- $\rho(i) = \rho(a_i)$ for every $i \in [2n]$.
- ρ restricted to $[0, 2n] \backslash \{0, \ldots, 2n\}$ is an embedding.

ρ is called a **realisation** of S. Two realisations ρ_1 and ρ_2 of S are said to be equivalent if there is a homeomorphism $h : S^2 \to S^2$ such that $h \circ \rho_1 = \rho_2$.

Hence, the question we wish to answer is that of conditions required on S which make it realisable, and when this realisation is unique. Note that if S comes from a diagram D (i.e. D as a subset of S^2 is the image of some realisation of S), then S also comes from every relative of D. Hence, for the purposes of this section we will talk of knot diagram with no under/over information.

Easier than questions about realisability of S are questions about realisability of S with an **orientation**, which we will define now.

Definition 6.6. An orientation for S is a map $f : [2n] \to \{1, -1\}$ such that $f(i) = -f(a_i)$ for every $i \in [2n]$.

Definition 6.7. For an orientation f on S, say that the pair (S, f) is **relisable** if there is a realisation ρ of S with the property that $f(i) = 1$ if and only if the arc $\rho((i-1, i+1))$ crosses the arc $\rho((a_i - 1, a_i + 1))$ from left to right (here $i+1$ and $a_i + 1$ are interpreted modulo $2n$). See fig. 9. Then ρ is said to be a **realisation** (S, f), and two realisations ρ_1 and ρ_2 of (S, f) are said to be equivalent if there is an orientation-preserving homeomorphism $h : S^2 \to S^2$ such that $h \circ \rho_1 = \rho_2$.

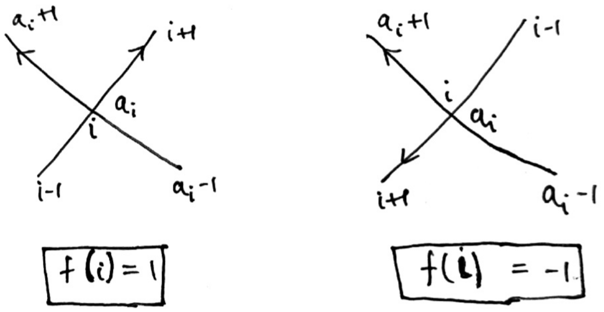


Figure 9: The orientation at the i-th crossing in terms of f. Clearly, the condition that $f(i) = -f(a_i)$ is necessary for realisability of (S, f).

For distinct $i, j \in [2n]$, let $[i, j]$ be the interval $\{i, i+1, \ldots, j\}$, where the elements are interpreted modulo $2n$. Likewise, $[i, j)$ will denote the set $[i, j] \backslash \{j\}$ and so on. These notations are not to be confused with the notation for intervals of real numbers — it should be clear from context which one is being referred to. *The letters i, s, t will henceforth always be used to denote integers modulo $2n$.*

When tabulating knots, one only cares about diagrams insofar as they represent different knots and contain as few crossings as possible. Hence, if there is some i such that $a_i = i + 1$ then any diagram that S comes from has a monogon and can be immediately simplified. Likewise, knot tabulation usually only considers prime knots and hence it suffices for us to only ask about when S comes from a prime diagram (although a prime diagram does not imply primality of the knot). It is straightforward to see that if S comes from a prime and has no monogons, it must satisfy the following condition.

Condition C1. $n \geq 3$ and for any distinct $i, j \in [2n]$ the involution a does not map $[i, j]$ to itself, i.e. there exists some $k \in [i, j]$ such that $a_k \notin [i, j]$.

Hence, we will now look at the realisability of pairs (S, f) such that S satisfies condition C1. For this, we will think of S as representing a graph with vertex set $\{\{i, a_i\} \mid i \in [2n]\}$. An edge is made between $\{i, a_i\}$ and $\{j, a_j\}$ whenever $|i - j| = 1$ or $|i - a_j| = 1$ or $|a_i - j| = 1$ or $|a_i - a_j| = 1$. For this we will now define realisability for graphs, orientations for graphs and realisability of oriented graph. For this section, a graph is a simple graph with no self-loops and no repeated edges.

Definition 6.8. A graph G is said to be **realisable** if there is a PL map $\rho : G \to S^2$ such that

- No two vertices have the same image.
- The images of all edges are arcs between the images of the corresponding vertices.
- The images of any two edges intersect only at a the image of a shared vertex (if there is no shared vertex then they do not intersect).

ρ is called a **realisation** of G. Two realisations ρ_1 and ρ_2 of G are said to be equivalent if there is a homeomorphism $h : S^2 \to S^2$ such that $h \circ \rho_1 = \rho_2$.

Definition 6.9. For a graph G and a vertex $v \in G$, a **cyclic orientation** at v is a cyclic ordering of the edges at v. G is said to be oriented if a cyclic orientation has been specified at every vertex.

Definition 6.10. An oriented graph G is said to be **strongly realisable** if it has a realisation ρ such that for every vertex $v \in G$, the counter-clockwise order in which ρ places the edges at v matches with the cyclic orientation at v. ρ is said to be a **strong realisation** for G. Two strong realisations ρ_1 and ρ_2 of G are said to be equivalent if there is an orientation-preserving homeomorphism $h : S^2 \to S^2$ such that $h \circ \rho_1 = \rho_2$.

Definition 6.11. An oriented graph G is said to be **weakly realisable** if it has a realisation ρ such that for every vertex $v \in G$, at least one of the counter-clockwise and clockwise orders in which ρ places the edges at v matches with the cyclic orientation at v. ρ is said to be a **weak realisation** for G. Two weak realisations ρ_1 and ρ_2 of G are said to be equivalent if they are equivalent as realisations of G as an unoriented graph.

Hence, using an orientation f for S, the graph corresponding to S can be given a natural orientation so that (S, f) is realisable if and only if the corresponding oriented graph is strongly realisable. Also, S is realisable if and only if the corresponding graph is weakly realisable. Henceforth, we will use the graph corresponding to (S, f) and its (weak and strong) realsability interchangeably with (S, f) and the realisability of S and (S, f). Finally, we now come to the result for which viewing S as a graph was necessary in the first place.

Theorem 6.12. *If an oriented 2-connected and 3-edge-connected graph G is strongly realisable, then it is uniquely strongly realisable (upto equivalence).*

Together with the following lemma, we will obtain a result regarding uniqueness of realisation for (S, f).

Lemma 6.13. *If S satisfies condition C1, then the corresponding graph is 2-connected and 3-edge-connected.*

Corollary 6.14. *If S satisfies condition C1 and (S, f) is realisable, then (S, f) is uniquely realisable (upto equivalence).*

Remark 6.15. In [2], a stronger version of **theorem 6.12** is proved — it is shown that if G is weakly realisable then it is uniquely weakly realisable (upto equivalence). Hence, the corresponding stronger version of **corollary 6.14** is that if S satisfies condition C1 and is realisable, then it is uniquely realisable (upto equivalence). Hence, there are precisely two orientations f of S such that (S, f) is realisable.
Here we will only prove the weaker statement that is **theorem 6.12**, since its proof is slightly cleaner.

The proof of **lemma 6.13** is straightforward. We now prove **theorem 6.12**.

*Proof of **theorem 6.12**.* Let ρ be a strong realisation of G. We may assume that G has at least one edge, since otherwise the claim is trivial. We will construct a sequence of subgraphs $G_1, G_2, \dots$ of G with strictly increasing edge sets, such that each G_i is uniquely strongly realisable (with orientation inherited from G) and every vertex in G_k has degree at least two (in G_k). Hence the sequence $(G_k)_{k\geq 1}$ must terminate with G, proving the claim.

Let $e = uv$ be an edge. Since $G - e$ is connected, so there is some path from u to v which does not include e. Hence, we obtain a cycle C containing e.

Let $f \neq e$ be an edge incident on u in C. Now $G - \{e, f\}$ is also connected, so there is a path P from u to some vertex in $C - u$ which does not contain any vertex from C other than its endpoints. Let $G_1 = C \cup P$. Hence, G_1 is a subgraph of G which topologically is a union of three arcs with common end-points and no other common points (see fig. 10).

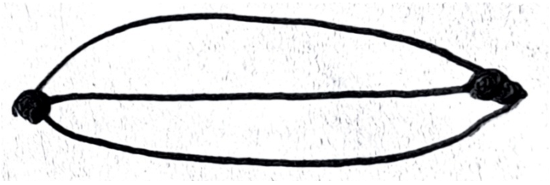


Figure 10: The subgraph G_1.

Clearly, G_1 is uniquely strongly realisable (with strong realisation $\rho|_{G_1}$) and every vertex in G_1 has degree two or three in G_1. Now, suppose G_k has been constructed as desired for some k and $\rho_k = \rho|_{G_k}$ is its unique strong realisation. If $G_k = G$ then we are done. Otherwise, there is some edge $e \notin G_k$ with one endpoint $u \in G_k$. Let the other endpoint of e be v. In a small disc neighbourhood Δ of u, the edges of G_k incident to u are radial lines which divide Δ into several sectors. Using the orientation of G at u, there is only one sector among these which e can intersect (see fig. 11).

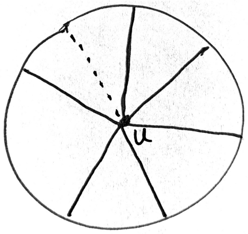


Figure 11: The small neighbourhood Δ of u. The solid lines are edges of u which are in G_k and the dotted line is e.

Hence, there is a component $F \subset S^2 \backslash \rho_k(G_k)$ such that any strong realisation ρ' of $G_k \cup e$ satisfies $\rho'(e) \subset \overline{F}$.[7] Hence, $G_k \cup e$ is uniquely strongly realisable.

---

[7] Here we are using the fact that $\rho_k(G_k)$, as a 1-dimensional CW complex, has empty boundary. This in turn is a consequence of the fact that every vertex in G_k has degree at least two in G_k.

Now consider two cases.

Case I: $v \in G_k$.
$G_{k+1} = G_k \cup e$ works.

Case II: $v \notin G_k$.
Since $G - u$ is connected, there is some path Q in $G - u$ from v to some vertex $w \in G_k - u$ such that Q does not intersect G_k except at w. Hence, $Q \cup e$ is a path between two distinct vertices $u, w \in G_k$ whose interior does not intersect G_k. Hence, $G_{k+1} = G_k \cup Q$ works (for the same reason that $G_k \cup e$ worked in case I). $\square$

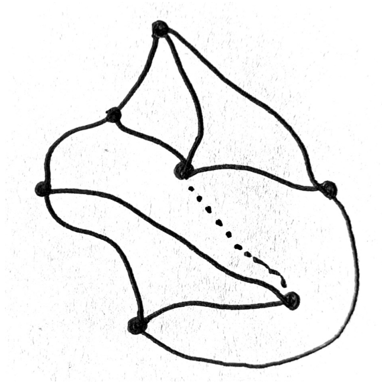


Figure 12: A schematic for G_{k+1}. The solid lines represents G_k and the dotted line represents e in case I and Q in case II.

Now we come to the question of when (S, f) is realisable in the first place. Suppose we naïvely start trying to construct a realisation $\rho : [0, 2n] \to S^2$ for (S, f) by first mapping $[1, 2]$ to some arbitrary arc and then inductively mapping $[i, i+1]$ ($i+1$ is interpreted modulo $2n$) to some arc which is picked in a way that is consistent with f. If we are able to do this all the way through and obtain a realisation ρ, we are done. If not, we can be certain that (S, f) is not realisable (since anything that was done in this construction was strictly necessary for realisability). One such situation is shown in fig. 13.

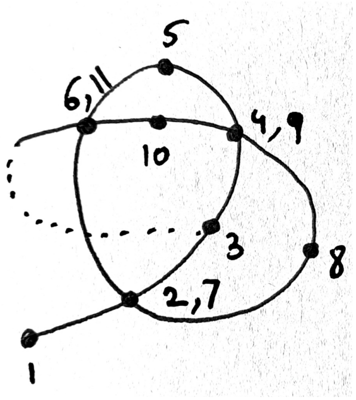


Figure 13: A failed attempt at constructing a realisation for the sequence given by the involution $1 \leftrightarrow 16, 2 \leftrightarrow 7, 3 \leftrightarrow 12, 4 \leftrightarrow 9, 5 \leftrightarrow 14, 6 \leftrightarrow 11$, $8 \leftrightarrow 13, 10 \leftrightarrow 15$. An arc from the 11-th crossing to the 12-th crossing (which is also the 3-rd crossing) cannot be made.

As can be seen in fig. 13, one possible obstruction for the construction ρ could be that for some i satisfying $a_i < i$, the a_i-th crossing and a_{i+1}-th crossings simply cannot be joined by an arc in any way, let alone an arc which is consistent with f (in fig. 13, this value of i is 11). Another obstruction that could occur is if for some i with $a_i < i$, the a_i-th and a_{i+1}-th crossings can only be joined by some arc which is not consistent with f. For instance, consider $i = 11$ in a modified version of fig. 13 which has $a_{12} = 5$ (instead of $a_{12} = 3$) and the given orientation required the arc from the 11-th crossing to the 12-th crossing to approach from inside the $4-5-6-10-4$ region.

Hence, to combinatorially capture these obstructions from S and f, we need to find a way of capturing some sort of 'inside-outside' information at every stage of the naïve construction described above. For this, we define maps $\phi_i : [2n] \to \{-1, 1\}$ for every $i \in [2n]$ defined inductively as follows. Here, all integers are interpreted modulo $2n$.

$$\phi_i(r) = \begin{cases} 1 & r = i \\ \phi_i(r-1) & a_r \notin [i, a_i] \\ -\phi_i(r-1) & a_r \in [i, a_i] \end{cases}$$

We will soon see that the inductive relation works for $r = i$ as well, i.e.

$\phi_i(i) = -\phi_i(i-1)$. First, we note that for $r \in [i+1, i-1]$ we have

$$\phi_i(r) = (-1)^{\#\{t \in (i,r] \,|\, a_t \in [i,a_i]\}} \tag{7}$$

If we interpret $(i, i]$ to be the empty set, then clearly the above definition works for $r = i$ as well. The above claim that the inductive definition works for $r = i$ is essentially saying that (7) works for i when $(i, i]$ is interpreted to be $\{i+1, i+2, \ldots, i-1, i\} = [2n]$ as well. To see that this is true, note that $\#[i, a_i] \equiv a_i - i + 1 \equiv 0 \pmod 2$ (by **lemma 6.1**), so we have

$$\begin{aligned} (-1)^{\#\{t \in [2n] \,|\, a_t \in [i,a_i]\}} &= (-1)^{\#[i,a_i]} \\ &= 1 \\ &= \phi_i(i) \end{aligned}$$

Hence, (7) is in fact true for all $r \in [2n]$, where $(i, i]$ can be interpreted as both $\emptyset$ and $[2n]$. In particular, this yields

$$\phi_i(i-1) = -1 \tag{8}$$

The cardinality of $\{t \in [i, a_i] \mid a_t \in [i, a_i]\}$ is even (since if t is in this set then so is a_t), so eq. (7) also yields

$$\phi_i(a_i) = -1 \tag{9}$$

Now, suppose (S, f) were realisable with realisation ρ. Fix $i \in [2n]$ and let 'the loop' refer to the closed loop $\rho([i, a_i])$ at the i-th crossing $\rho(i) = \rho(a_i)$. We will now demonstrate two necessary conditions for realisability ((10) and (11)) in terms of the functions ϕ_i and f which will correspond with the two possible obstructions for realisability that were discussed using fig. 13. Naturally, the next step will be to prove the sufficiency of these two conditions.

In the complement $S^2 \backslash \rho([i, a_i])$ of the loop, say that the component containing the arcs $\rho((i-1, i))$ and $\rho((a_i, a_i + 1)$ is **outside**. Any point p in the complement of the loop is also said to be outside if it can be reached from the abovementioned component using an arc in S^2 which intersects $\rho([i, a_i] \backslash \{i, i+1, \ldots, a_i\})$ transversally an even number of times. Likewise, any point p in the complement of the loop is said to be **inside** if it can be reached from some outside point q using an arc in S^2 which intersects $\rho([i, a_i] \backslash \{i, i+1, \ldots, a_i\})$ an odd number of times. Using the fact that the linking number is always an integer, it can be seen that any point in the complement of the loop cannot be both inside and outside.

Remark 6.16. What we are calling inside and outside here are simply the black and white regions one would obtain using the colouring described in §4 (by interpreting the loop as a knot diagram). However, to remain consistent with [2] we will stick with the inside/outside terminology.

Now, note that $\phi_i(a_i) = -1$ (by (9)) and the arc $\rho((a_i, a_i+1))$ just beyond a_i is outside. By taking this as a base case and using induction, we see that

$$\phi_i(s) = \begin{cases} -1 & \rho((s, s+1)) \text{ is outside} \\ 1 & \rho((s, s+1)) \text{ is inside} \end{cases} \quad \forall\, s \in [a_i, i)$$

In particular, if $s \in (a_i, i)$ such that $a_s \in (a_i, i)$, then the point $\rho(s) = \rho(a_s)$ is not on the loop and hence inside or outside. Hence we have $\phi_i(s) = \phi(a_s)$, or in other words

$$\phi_i(s)\phi_i(a_s) = 1 \text{ whenever } s, a_s \in (a_i, i) \tag{10}$$

Now note that $\phi_i(i) = 1$, and outside is to the left of $\rho((i, i+1))$ if and only if $f(i) = 1$ (see fig. 9). Using this as a base case and performing induction we see that for any $s \in [i, a_i)$, outside is to the left of $\rho((s, s+1))$ if and only if $\phi_i(s)f(i) = 1$. From fig. 9 it can also be seen that the arc $\rho((a_s, a_s+1))$ is to the left of $\rho((s-1, s+1))$ if and only if $f(s) = 1$. Combining these two obsevations for the case of $a_s \notin [i, a_i]$, we see that $\phi_i(a_s) = -1$ if and only if $\phi_i(s)f(i) = f(s)$. In other words,

$$\phi_i(a_s) = -\phi_i(s)f(i)f(s) \text{ whenever } s \in (i, a_i) \text{ and } a_s \in (a_i, i)$$

By interchanging the role of s and a_s in the above and using $f(s) = -f(a_s)$, we obtain

$$\phi_i(s)\phi_i(a_s) = f(i)f(s) \text{ whenever } s \in (a_i, i) \text{ and } a_s \in (i, a_i) \tag{11}$$

To conclude, satisfying the following is necessary for realisability.

Condition C2. For every $i \in [2n]$ and $s \in (a_i, i)$, we have

$$\phi_i(s)\phi_i(a_s) = \begin{cases} 1 & a_s \in (a_i, i) \\ f(i)f(s) & a_s \in (i, a_i) \end{cases}$$

Note that ϕ_i depends only on S, so this is in fact a completely combinatorial condition on (S, f) which does not require any attempts to be made at constructing a realisation.

Theorem 6.17. *If S satisfies condition C1, then condition C2 is necessary and sufficient for realisability of (S, f).*

Proof. Necessity has been demonstrated. The proof of sufficiency is simply a formal demonstration of the idea that if we naïvely begin constructing a realisation ρ for (S, f), the only obstructions in extending ρ that we can encounter at any point are those two which were discussed using fig. 13, and that these are indeed captured by (10) and (11). Suppose (S, f) is not realisable, and we will find some $i \in [2n]$ and $s \in (a_i, i)$ which violate condition C2.

Let G be the oriented graph corresponding to (S, f) obtained from $[0, 2n]$ by glueing 0 to $2n$ and i to a_i for every $i \in [2n]$. For each $t \in [2n]$, let H_t be the oriented subgraph induced by vertices with labels $1, \ldots, t$. Let $r \in [2n]$ be maximal such that H_r is strongly realisable, and let $\rho : [1, r] \to S^2$ be a strong realisation. If $a_{r+1} \notin [1, r]$, then clearly ρ can be extended to a strong realisation of H_{r+1} (contradicting the maximality of r). Hence, $a_{r+1} \in [1, r]$.

Let α and β be points on the edge from r to $r+1$ in G with $r < \alpha < \beta < r+1$. Extend ρ to a strong realisation of the image of $[1, \alpha]$ in G by adding a small arc from $\rho(r)$ to some point $\rho(\alpha) \notin \rho([1, r])$. Likewise, extend ρ further to a strong realisation of the image of $[1, \alpha] \cup [\beta, r + 1]$ in G by adding a small arc from some point $\rho(\beta) \notin \rho([1, r])$ to $\rho(a_{r+1})$. Since this extension is a strong realisation, the arc $\rho([\beta, r + 1])$ is placed to the right or left of the arc $\rho((a_{r+1} - 1, a_{r+1} + 1))$ as per the value of $f(a_{r+1})$.

For convenience, view S^2 as the one-point compactification of the complex plane $\mathbb{C}$ with α at the origin and β at ∞. Since ρ cannot be extended to a strong realisation of H_{r+1}, the points α and β must be in different components of $S^2 \backslash \rho(H_r)$. Let C be the cycle in the graph H_r whose image under ρ is the boundary of the component of $S^2 \backslash \rho(H_r)$ containing 0. Orient $\rho(C)$ (as a simple closed curve) so that it has it has winding number 1 (about the origin). Let $v_1, v_2, \ldots, v_k \in \mathbb{C}$ be the successive vertices of $\rho(C)$ along its orientation, and let $e_1 = v_1v_2, e_2 = v_2v_3, \ldots, v_k = v_kv_1$ be the edges. By virtue of how C was defined, for each $j \in \{1, \ldots, k\}$ we can choose $t_j \in [1, r]$ so that the following are satisfied (see fig. 14).

- $a_{t_j} \in [1, r]$.
- $\rho(t_j) = \rho(a_{t_j}) = v_j$.

- $|a_{t_{j+1}} - t_j| = 1$.
- If $[t_j, a_{t_{j+1}}]$ denotes the interval of real numbers between t_j and $a_{t_{j+1}}$, then the edge e_j is the arc $\rho([t_j, a_{t_{j+1}}])$.

Figure 14: A depiction of $\rho(H_r)$ around the origin, with C thickened. C 'changes strands' at every vertex.

Here, $j + 1$ is reduced modulo k. Now, for $z \in \rho(C)$ we define $\arg_C(z)$ to be the unique branch of the argument function which satisfies

- $\arg_C(v_1) \in [0, 2\pi)$.
- $\arg_C$ is continuous on e_1.
- $\arg_C$ is continuous on $\rho(C)\backslash\{v_1\}$.

Hence, as we follow $\rho(C)$ along its orientation starting just before v_1, $\arg_C$ makes a 'jump' of -2π at v_1. More formally, if $\sigma_j = a_{t_{j+1}} - t_j \in \{-1, 1\}$ and $c_j \in \mathbb{Z}$ satisfies

$$\lim_{\epsilon \to 0^+} \arg_C(\rho(a_{t_j} - \sigma_j \epsilon)) - \arg_C(v_j) = 2\pi c_j \tag{12}$$

then $2\pi c_j$ is the 'jump' in $\arg_C$ which happens at v_j as we pass from e_{j-1} to e_j. Hence, $c_1 = -2\pi$ and $c_j = 0$ for $j \neq 1$. In particular,

$$\sum_{j=1}^{k} c_j = -1 \tag{13}$$

Next, we define $\arg_H : [1, r] \to \mathbb{R}$ to be the unique continuous map satisfying $\arg_H(1) \in [0, 2\pi)$ and

$$e^{i \arg_H(x)} = \rho(x) \ \forall\, x \in [1, r]$$

Let $d_j \in \mathbb{Z}$ satisfy

$$\arg_H(a_{t_j}) - \arg_H(t_j) = 2\pi d_j \tag{14}$$

Hence, $2\pi d_j$ is the 'jump' in $\arg_H$ which happens at v_j as we pass from e_{j-1} to e_j.

For every j, $\arg_C$ is continuous on the arc $e_j \backslash \{v_{j+1}\} = \rho([t_j, a_{t_{j+1}}))$. Hence, on $[t_j, a_{t_{j+1}})$ we have that $\arg_C \circ \rho$ and $\arg_H$ are both continuous. Hence they differ by some constant integer multiple of 2π, say

$$(\arg_C \circ \rho - \arg_H)|_{[t_j, a_{t_{j+1}})} \equiv 2\pi l_j \tag{15}$$

Putting together all the equations we have obtained so far yields

$$\begin{aligned}
2\pi \sum_{j=1}^{k} d_j &= \sum_{j=1}^{k} \arg_H(a_{t_j}) - \sum_{j=1}^{k} \arg_H(t_j) \text{ (by (14))} \\
&= \lim_{\epsilon \to 0^+} \sum_{j=1}^{k} \arg_H(a_{t_j} - \sigma_j \epsilon) - \sum_{j=1}^{k} \arg_H(t_j) \text{ (} \arg_H \text{ is continuous)} \\
&= \lim_{\epsilon \to 0^+} \sum_{j=1}^{k} \left[\arg_C(a_{t_j} - \sigma_j \epsilon) - 2\pi l_{j-1}\right] - \sum_{j=1}^{k} \left[\arg_C(t_j) - 2\pi l_j\right] \text{ (by (15))} \\
&= \sum_{j=1}^{k} 2\pi c_j \text{ (by (12))} \\
&= -2\pi \text{ (by (13))}
\end{aligned}$$

In particular, there is some $j_0 \in \{1, \ldots, k\}$ such that d_{j_0} is odd — let $i \in [1, r]$ such that $\rho(i) = \rho(a_i) = v_{j_0}$ and $i < a_i$. Hence, we have $1 \leq i < a_i \leq r$.

As before, let 'the loop' refer to $\rho([i, a_i])$. Note that condition C2 is satisfied by (S, f) for $s \in [1, r]$ whenever $a_s \in [1, r]$, since H_r is strongly realisable (and so the geometric arguments used to derive condition C2 work on H_r).

We now claim that for i as above and $s = r + 1$, (S, f) does not satisfy condition C2.

Since d_{j_0} is odd, the loop has an odd winding number around the origin, i.e. any arc from the origin to ∞ which intersects $\rho([i, a_i]\backslash\{i, \ldots, a_i\})$ transversally does so an odd number of times, i.e. if the origin $\rho(\alpha)$ is inside then $\infty = \rho(\beta)$ is outside and vice-versa. Hence if $a_{r+1} \notin [i, a_i]$ then we have

$$\phi_i(r)\phi_i(a_{r+1}) = -1$$

Also, $\phi_i(r + 1) = \phi_i(r)$ (since $a_{r+1} \notin [i, a_i]$), so

$$\phi_i(r + 1)\phi_i(a_{r+1}) = -1$$

This violates condition C2 for $s = r + 1$. Next, suppose $a_{r+1} \in [i, a_i]$. We claim that the following are equivalent.

1. $\phi_i(a_{r+1})f(i) = f(a_{r+1})$.
2. $\rho(\beta)$ is inside.
3. $\rho(\alpha)$ is outside.
4. $\phi_i(r) = -1$.

First, we prove $1 \iff 2$. We know that outside is to the left of the arc $\rho((a_{r+1}, a_{r+1} + 1))$ if and only if $\phi_i(a_{r+1})f(i) = 1$. Since the arc $\rho([\beta, r + 1])$ was placed in accordance with $f(a_{r+1})$, we also know that $\rho(\beta)$ is to the left of $\rho([a_{r+1}, a_{r+1} + 1))$ if and only if $f(a_{r+1}) = -1$. Hence $1 \iff 2$ follows.

$2 \iff 3$ follows from the fact that the loop has odd winding number around the origin, as noted before. $3 \iff 4$ follows from the geometric deductions regarding ϕ_i made earlier. Now, the equivalence of 1 and 4 yields

$$\begin{aligned} \phi_i(a_{r+1})f(i)f(a_{r+1}) &= -\phi_i(r) \\ \implies \phi_i(r)\phi_i(a_{r+1}) &= f(i)f(a_{r+1}) \\ &= f(i)f(r + 1) \end{aligned}$$

Finally, we note that $\phi_i(r+1) = -\phi_i(r)$ since $a_{r+1} \in [i, a_i]$. Hence, we obtain

$$-\phi_i(r + 1)\phi_i(a_{r+1}) = f(i)f(r + 1)$$

Once again, this violates condition C2 for $s = r + 1$. □

Corollary 6.18. *If (S, f) satisfies condition C1, then it is uniquely realisable (upto equivalence) if and only if it satisfies condition C2.*

References

[1] Lickorish, W. R. (2012). *An introduction to knot theory* (Vol. 175). Springer Science & Business Media.

[2] Dowker, C. H., & Thistlethwaite, M. B. (1983). Classification of knot projections. *Topology and its Applications, 16*(1), 19-31.