Knots and 3-dimensional computational topology

Francis Lazarus and Arnaud de Mesmay

November 28, 2017

Contents

In the next two courses, we switch our attention from 2-dimensional space (surfaces) to 3-dimensional space, focusing on *knots*, and incidentally dealing a bit with 3-manifolds. This increase in dimension has a dramatic effect from the point of view of computational topology: while most topological problems on surfaces can be solved efficiently, their generalizations in dimension 3 are much harder to understand. As an illustration, while recognizing closed surfaces can be done in linear time by just computing the Euler characteristic and orientability, all the algorithms known to detect whether two 3-dimensional spaces are homeomorphic [Jac05, Kup15] are very inefficient (requiring more than towers of exponentials) and complicated, relying on Perelman's recent proof [Per02, Per03] of Thurston's Geometrization Conjecture. Nevertheless, despite being hard, most problems in 3 dimensions are decidable, and thus form an interesting middle ground before higher dimensions where indecidability results start to kick in.

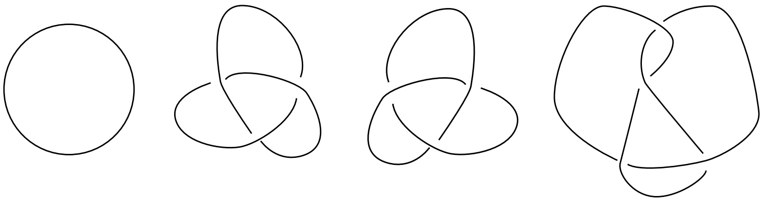

Figure 1: Example of knots: the trivial knot, the left and right trefoil knots, and the figure-eight knot. These are not polygonal but can obviously be made so.

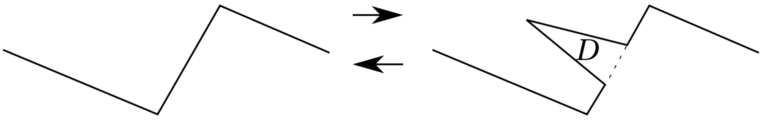

Figure 2: An elementary move on a knot: a segment is subdivided and can be moved along a triangle if the disk D does not intersect the rest of the knot.

1 Knots

A **knot** is a closed curve in $\mathbb{R}^3$, or more formally an embedding $\mathbb{S}^1 \to \mathbb{R}^3$. In contrast with the two-dimensional case, allowing arbitrary topological knots might lead to pathological objects known as **wild knots** which, while interesting in their own right, are not very relevant from an algorithmic perspective. So we restrict our attention to **tame knots**, which are polygonal[1] embeddings $\mathbb{S}^1 \to \mathbb{R}^3$. We will omit the word tame throughout these notes. Two knots are considered equivalent if they can be deformed continuously one onto the other one without introducing self-crossings in the process. More formally, the notion of equivalence considered here is through **ambient isotopy**, that is a continuous family of homeomorphisms $h_t : \mathbb{R}^3 \times [0,1] \to \mathbb{R}^3$. Two knots K_1 and K_2 are (ambient) **isotopic** if there exists an ambient isotopy h_t such that $h_0 = Id_{\mathbb{R}^3}$ and $h_1(K_1) = K_2$. Figure 1 shows some examples of knots.

Remark 1: By the *Alexander trick* (see for example [BZ85, Proposition 1.9]), any orientation-preserving homeomorphism of the 3-ball $\mathbb{B}^3 \to \mathbb{B}^3$ is actually an ambient isotopy, so homeomorphisms could simply be used to define equivalence, but the notion of ambient isotopies is more intuitive.

Remark 2: On the other hand, our notion of ambient isotopy is a bit unnatural because it breaks the polygonal structure of tame knots. The underlying idea is that we restrict our attention to tame knots to avoid pathologies, but once this restriction has been made, it is much more convenient to allow any kind of continuous deformation. Yet if one insists on only allowing polygonal deformations, one can restrict the allowed isotopies to *elementary moves* or Δ-moves which are illustrated in Figure 2. It turns out that knots are equivalent under ambient isotopies if and only if they are equivalent under elementary moves, see [BZ85, Proposition 1.10].

Remark 3: Another tempting definition of equivalence could be to use **isotopies**, i.e., to say that two knots K_1 and K_2 are isotopic if there is a continuous family of embeddings $i_t : \mathbb{S}^1 \to \mathbb{S}^3$ such that $i_0 = K_1$ and $i_1 = K_2$. However, with such a definition

[1] Considering *smooth* knots leads to the same theory.

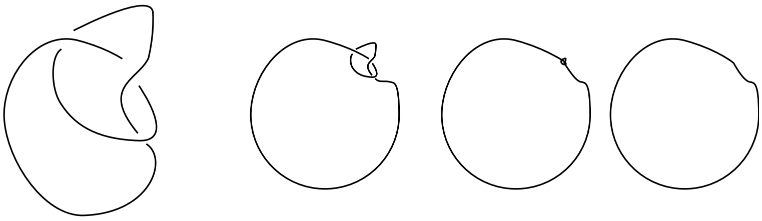

Figure 3: Equivalence of all knots using just isotopy.

all the knots would be equivalent, as illustrated in Figure 3, since it allows to pull the knotted portion progressively tighter until it disappears.

A **trivial knot** (or unknot) is a knot isotopic to the trivial embedding of the circle $\mathbb{S}^1 \to \mathbb{R}^3$. Since we restrict our attention to polygonal knots, the following problem is a well-defined algorithmic problem, which will be the focus of these two lectures.

> UNKNOT RECOGNITION
> **Input:** A knot K described as a concatenation of n segments.
> **Output:** Is K a trivial knot?

As we will see, this is a tricky problem, the current state of the art is that it lies in **NP** ∩ **co-NP** (see [HLP99, Lac16]), yet no polynomial time algorithm is known for this problem. But before delving into this, let us first observe that it is not even easy to prove that there exist non-trivial knots. In order to prove this, we introduce knot diagrams.

2 Knot diagrams

A convenient way to deal with knots is to represent them using **knot diagrams**, that is, a 2-dimensional orthogonal projection $p : \mathbb{R}^3 \to \mathbb{R}^2$ to project K on $\mathbb{R}^2$. We call such a projection regular if there are no triple (or worse) points and no vertex of K is a double point. A knot diagram is obtained from a regular projection by specifying at each crossing which strand is above the other one. By perturbing slightly either a knot or the desired projection if needed, it is easy to associate a knot diagram with any knot. Basically, this is what we did in Figure 1 without even bothering to mention it. From the point of view of graph theory, a knot diagram is a 4-regular planar graph where each vertex bears a *marking*, indicating which strand is above the other one.

The **Reidemeister moves** are the local moves relating knot diagrams represented in Figure 4. Two knot diagrams are considered equivalent if they can be related with isotopies of $\mathbb{R}^2$ and Reidemeister moves.

Theorem 2.1 (Reidemeister [Rei27])**.** *Two knots are equivalent if and only if all their diagrams are equivalent.*

PROOF. We first prove that any two regular projections of the same knot are connected by Reidemeister moves. Each projection can be identified with a point on the

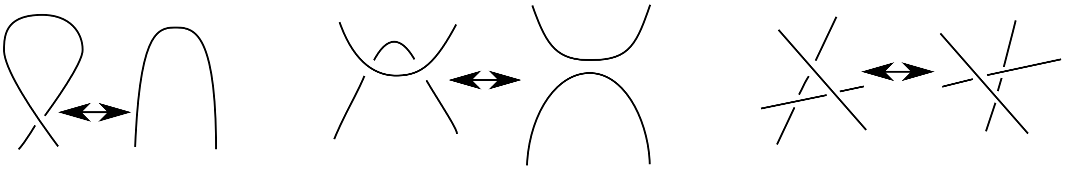

Figure 4: The three Reidemeister moves.

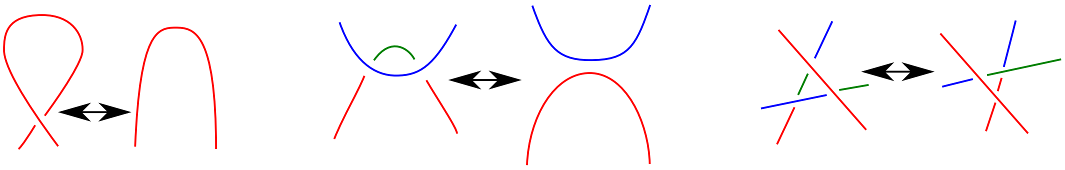

Figure 5: Tricoloring the Reidemeister moves.

sphere $\mathbb{S}^2$, and the non-regular projections are correspond to curves on this sphere. By connecting p_1 and p_2 by a path in general position with respect to these lines, it is enough to show that crossing a line of non-regular projections can be done with Reidemeister moves. The three possible situations of the knot around these lines correspond to the three Reidemeister moves. Now, since two equivalent knots can be related using elementary moves (see Remark 2 above), it is enough to show that the projection of an elementary move can be realized with Reidemeister moves, which is easily verified. □

We can leverage on this combinatorial approach to knot equivalence to provide an easy proof that the trefoil knot is non-trivial.

Proposition 2.2. *There exists a non-trivial knot.*

PROOF. A knot diagram is said to be **tricolorable** if each strand can be colored using one of three colors, with the following rule:

1. At least two colors must be used.
2. At a crossing, the three incident strands[2] are either all of the same color or all of different colors.

We claim that the tricolorability of a knot does not depend on the knot diagram. By Theorem 2.1, it suffices to prove that Reidemeister moves preserve tricolorability, which is illustrated in Figure 5.

Now, observing that the trivial knot is not tricolorable (since it can only be colored with one color), while the trefoil knot is (see Figure 6), this proves that the trefoil knot is non-trivial. □

Note that this does not detect all the non-trivial knots, since the figure-eight knot can not be colored with three colors either (left as an exercise), and it is not trivial (see Exercise 3.3). Also note that for algorithmic purposes, this approach, being based on coloring, is very inefficient.

[2] Here, the arc going "above" is considered as a single strand, while the arc going "below" is cut into two strands.

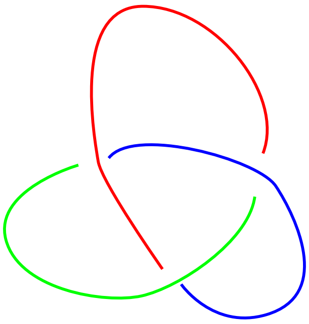

Figure 6: Tricoloring the trefoil knot.

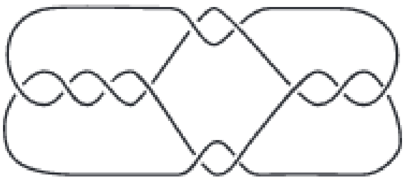

Figure 7: Goeritz's unknot.

On the other hand, Theorem 2.1 suggests a very candid approach to solve Unknot recognition: simply try combinations of Reidemeister moves until one reaches the **trivial diagram**, i.e., an embedding $\mathbb{S}^1 \to \mathbb{R}^2$. Optimistically, one might hope it is never necessary to make a knot more complicated to untangle it, i.e., maybe the Reidemeister move II increasing the number of crossings is not needed to reach the unknot. This would bound the number of combinations to try and give an exponential algorithm. However, there exist **hard unknots**, that is, knot diagrams of the unknot that require to be made more complicated before reaching the unknot. Figure 7 shows one of those, and there are infinite families of these.

That approach is not doomed to fail however, and there exist bounds on the number of Reidemeister moves needed to simplify a trivial knot. In a recent breakthrough, Lackenby obtained the following polynomial bound.

Theorem 2.3 (Lackenby [Lac15])**.** *Let D be a diagram of the trivial knot with c crossings. Then there exists a series of $(236c)^{11}$ Reidemeister moves transforming it into the trivial diagram.*

This provides an exponential time algorithm to solve Unknot recognition, and also proves that it is in **NP**: the certificate is the sequence of Reidemeister moves to be applied to reach the trivial diagram. As the constant and exponent might suggest, this theorem is very intricate and we will not prove it in this course. However, we will see another algorithm showing that the problem is in **NP**, based on *normal surface theory*, which is one of the main technical tools in the proof of Theorem 2.3. So after this course, you will be better equipped to understand the proof.

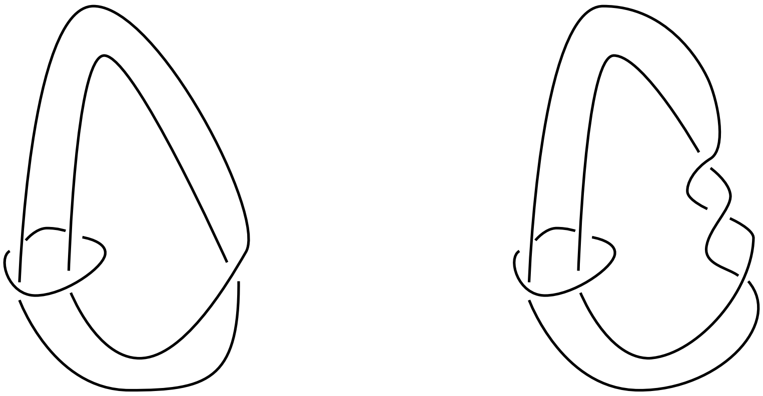

Figure 8: Links are not determined by their complements.

3 The knot complement

A central way to study knots is to study the topological properties of their complements. For this purpose, it is convenient to compactify $\mathbb{R}^3$, i.e., to add a point and identify the result to $\mathbb{S}^3$ using the stereographic projection. Knots in $\mathbb{R}^3$ and $\mathbb{S}^3$ behave identically, so we will work in this section with the latter framework. For a knot K in $\mathbb{S}^3$, drill a tubular neighborhood N around K and denote the resulting space by $M = \mathbb{S}^3 \setminus N$. This is an example of a **3-manifold with boundary**, i.e., a topological space where every point is locally homeomorphic to $\mathbb{R}^3$ or the half-space $\mathbb{R}^3_{|x\geq 0}$.

The study of knot complements as a tool to understand a knot is justified by the following theorem, which at the same time sounds very obvious and is extremely hard to prove.

Theorem 3.1 (Gordon-Luecke [GL89]). *Knots are determined by their complements, i.e., if two knots have complements that are homeomorphic with an orientation-preserving homeomorphism, then they are isotopic.*

To emphasize the strength of this theorem, let us just warn the reader that **links**, which are embeddings of disjoint copies of $\mathbb{S}^1$ into $\mathbb{R}^3$, are *not* determined by their complements: Figure 8 shows two links which are not ambient isotopic, yet their complements are homeomorphic.

So now that instead of this somewhat strange notion of ambient isotopy, we reduced the problem to determining the homeomorphism class of a space, we can try to apply the tools that we have seen in the previous lectures to solve it. As we shall see, this will not be very fruitful with respects to solving UNKNOT RECOGNITION.

3.1 Homotopy

One can define the homotopy group π_1 of the topological space $M = \mathbb{S}^3 \setminus N$ in the same way as the topological fundamental group that we introduced for surfaces. We first choose a basepoint x, consider the set of loops based at x and say that two loops are equivalent if they are homotopic (in M, that is, if they can be deformed into each other *without crossing* K). The set of loops obtained this way forms the group $\pi_1(M)$, where the identity element is the trivial loop at x and the law is the concatenation.

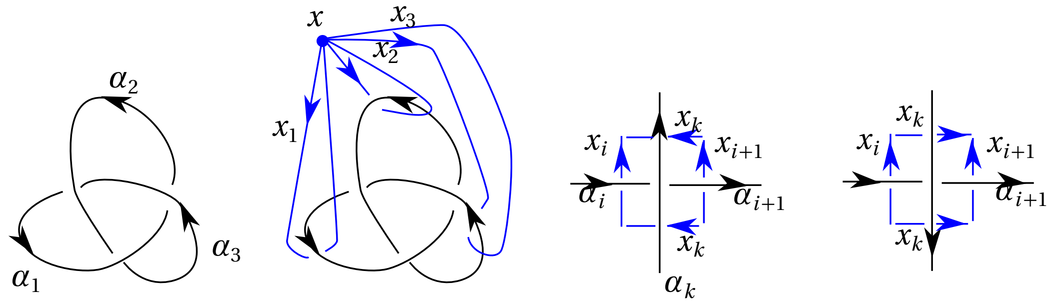

Figure 9: The Wirtinger presentation. In the last two pictures, depending on the orientation of the strands, we will have $x_k x_i = x_{i+1} x_k$ or $x_i x_k = x_k x_{i+1}$

It turns out that there is a fairly easy way to obtain a presentation of the group $\pi_1(M)$ via the **Wirtinger presentation**. Start with a knot diagram D of K. As for tricolorability, we consider a *strand* in the diagram to be and arc in the diagram between two points where it goes "below" another arc. Number the strands of D by $\alpha_1, \dots, \alpha_n$ according to the order in which they appear in D, and orient them according to an arbitrary orientation. Now, let us pick a basepoint x somewhere above the knot (for example in the eye of the reader), and define a set of loops $x_1, \dots, x_n$ by starting at x and going looping around α_i by passing under it in a right left direction, and going back to x.

Now, at each crossing involving the strands α_i, α_{i+1} and α_k, the loops x_i, x_{i+1} and x_k will verify some relation, which can be simply read depending on the orientation of the crossings, see Figure 9. In the first case, we will have $x_k x_i = x_{i+1} x_k$, while in the second case we will have $x_i x_k = x_k x_{i+1}$, we denote by r_i the corresponding relation that holds. It turns out that these relations encapsulate all the possible relations between the generators x_i.

Theorem 3.2. *The fundamental group $\pi_1(\mathbb{S}^3 \setminus N)$ admits the presentation*

$$< x_1, \dots, x_n \mid r_1, \dots r_n. >$$

PROOF. The formal way to prove this theorem is to decompose $\mathbb{S}^3$ into cells defined by the knot K and apply the *van Kampen theorem*, see Hatcher [Hat02, Section 1.2]. In order to keep things simple, we provide a somewhat vague proof that should be enough to convince the novice reader that the theorem is correct, and to convince the expert reader (who probably already knows all this) that applying the van Kampen theorem really yields this result.

Grow a family of vertical half-planes $Z = Z_1, \dots Z_n$ under each strand of the knot diagram, as in Figure 10. Some of these planes will touch (under a crossing of D). Any loop based at x not crossing any of the Z_i can be homotoped to the trivial loop, while any loop crossing the half-planes $Z_{i_1}, \dots, Z_{i_k}$ is homotopic to the concatenation of the loops $x_{i_1}, \dots, x_{i_k}$, maybe some of which are inverted depending on the orientations of the crossings. Thus the x_i generate the set of equivalence classes of loops. Furthermore, locally, a homotopy changing these generators either crosses a Z_i in two places, corresponding to a subword $x_i x_i^{-1}$, or it crosses a double intersection line of three half-planes Z_i, Z_{i+1} and Z_k, leading to modifying the word by one of the relations. □

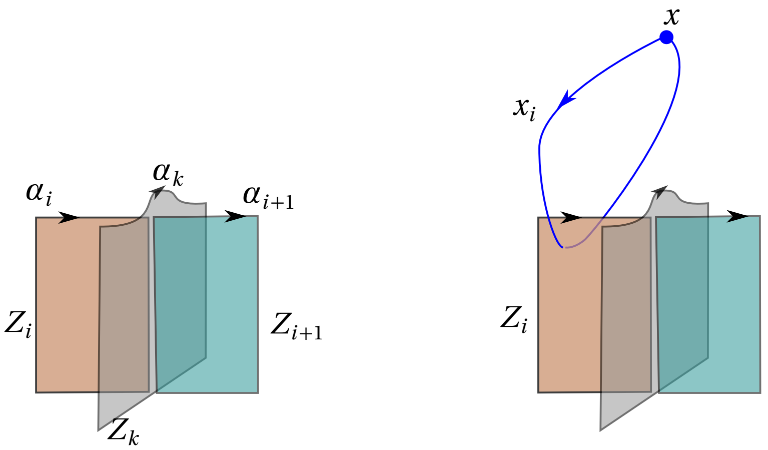

Figure 10: Half-planes under the strands

So, fundamental groups of (complements of) knots are easy to compute, but this is hard to exploit in order to distinguish knots. First, it is not true that non-equivalent knots have non-isomorphic groups. Nevertheless, it is true for the unknot (it is the only knot corresponding to group $\mathbb{Z}$), and one can add additional structure to knot groups (using *peripheral systems*, see [BZ85, Section 3.C]) so that they distinguish all the knots. However, and more importantly, we also hit the same difficulty that we met when studying surfaces: while studying the group structure of $\pi_1(M)$ leads to a very rich algebraic structure and theory, it is hard to use it to extract algorithms, since most computational problems on groups presentations are undecidable in general. The following exercise shows that even the simplest cases require some work.

Exercise 3.3. Compute a presentation of the fundamental group of the complement of the figure-eight knot, and deduce from it that it is not trivial. *Hint:* Try mapping into a non-abelian finite group.

Some positive algorithmic results can be achieved by using the special structure of these groups (as was the case for surfaces). For example, the idea of finding a non-abelian representation which underlies the exercise above can be extended to work with any non-trivial knot, ultimately providing a polynomial-sized certificate of a knot being *not* the unknot (or equivalently that UNKNOT RECOGNITION is in **co-NP**), provided the Generalized Riemann Hypothesis is true [Kup14]. But this line of work falls largely outside of the scope of this class. We refer to the survey [AFW15] for more information on this topic. Let us just mention one doomed idea in the next subsection.

3.2 Homology

If you did the exercises in the lecture notes on surfaces, you might recall Exercise 4.8 where you were asked to prove that fundamental groups of non-homeomorphic surfaces are not isomorphic. Viewed from another angle, this could be seen as a way to tell that two surfaces are not homeomorphic, similarly as what we are trying to do here, one dimension higher. One simple way to carry this out is to *abelianize* these groups (see the notes on minimum weight bases) and observe that the resulting abelian groups have different ranks. As we saw in the course on minimum weight bases, the abelianization of the fundamental group is the same thing as the first *homology group*

$H_1(M)$. Since these groups are abelian, they are much more tractable algorithmically, and maybe they are strong enough to distinguish knot complements, and thus knots. This is not the case.

Proposition 3.4. *For any knot K, we have $H_1(\mathbb{S}^3 \setminus N) = \pi_1^{ab}(\mathbb{S}^3 \setminus N) = \mathbb{Z}$.*

PROOF. The equality of the first homology group and the abelianization of the fundamental group is a general topological result known as the Hurewicz Theorem, we refer to Hatcher [Hat02, Section 4.2] for the proof. Then, starting from the Wirtinger presentation, we simply observe that all the relations are of the form $x_k x_i x_k^{-1} x_{i+1}^{-1}$ or $x_i x_k x_{i+1}^{-1} x_k^{-1}$, which after abelianization yield $x_i x_{i+1}^{-1}$. Therefore, when abelianizing, all the generators merge into a single one and we obtain the group $\mathbb{Z}$. □

So the first homology group of the knot complement will not help us much in distinguishing knots.

3.3 Triangulations

Similarly as the way surfaces could be described by gluing together disks, one can cut 3-manifolds into balls and describe how these are glued together. To keep things simple (but they will still be complicated enough), we restrict our attention to the 3-dimensional analogue of triangulations using tetrahedra, which are also called triangulations instead of tetrahedrizations. A triangulation T is the topological space obtained from a disjoint set of t tetrahedra $T = (T_1, \ldots, T_t)$ by (combinatorially) gluing some pairs of two-dimensional faces of these tetrahedra; a gluing between two faces is specified by a bijection from the vertex set of the first face to the vertex set of the second face. As a result of these gluing, edges and vertices of tetrahedra are also identified; it is also allowed to glue two zero-, one-, or two-dimensional faces of the same tetrahedron.

Since in a 3-manifold with boundary, the neighborhood of every point has to be an open ball or a half ball, the following conditions are necessary for a triangulation T to be a 3-manifold (possibly with boundary):

1. Each vertex has a neighborhood homeomorphic to $\mathbb{R}^3$ or to the closed half-space;

2. After the gluings, no edge is identified to itself in the reverse orientation.

Conversely, it is known [Moi52] that any 3-manifold M is the underlying space of such a triangulation (this is the 3-dimensional case of the Kerékjártó-Radó theorem that we saw for surfaces).

It is not very hard, yet somewhat tedious to build a triangulation of a knot complement starting from the knot.

Proposition 3.5. *Given a polygonal knot K made of n segments, we can compute a triangulation of $\mathbb{S}^3 \setminus N$ in $O(n)$ time.*

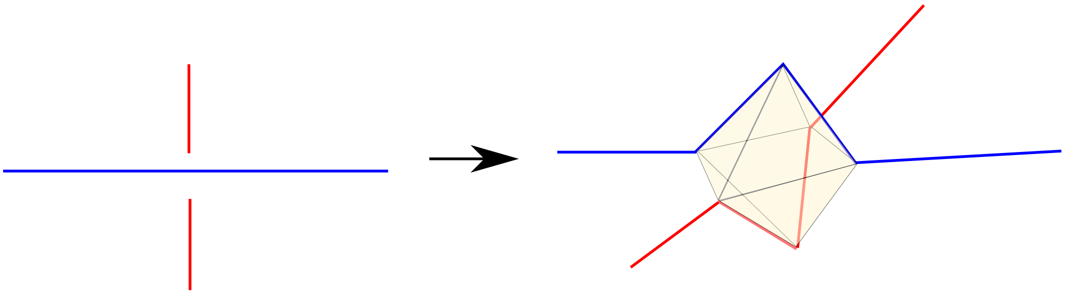

Figure 11: Putting an octahedron at every crossing.

PROOF. Starting with a polygonal knot, project it into a knot diagram, and triangulate the resulting planar graph. Then at each crossing, use the octahedron gadget of Figure 11. Now, outside the octahedra, there is a big 3-dimensional polyhedron, and after subdividing it into tetrahedra, we obtain a triangulation of $\mathbb{S}^3$ where K lies on the edges of the tetrahedra. Drilling tubes around the corresponding edges and retriangulating the space gives a triangulation of $\mathbb{S}^3 \setminus N$. □

For surfaces, it was somewhat easy to recognize which surface one would obtain by identifying the disks by just visualizing the corresponding identifications in 3 dimensions. Since our imagination is much more lacking with 4-dimensional space, the corresponding approach to recognize 3-manifolds via their triangulations is much harder, as the following mind-bending exercise dealing with a single tetrahedron (!) with a single vertex (!!) that can actually be embedded in $\mathbb{R}^3$ (!!!) showcases.

Exercise 3.6. Take a single tetrahedron and label its vertices by $0, 1, 2$ and 3. Identify the 012 face with the 130 face by sending the vertices $0, 1$ and 2 respectively to $1, 3$ and 0.

1. Prove that the resulting space is a 3-manifold with boundary.
2. Which familiar one is it?

4 An algorithm for unknot recognition

Now that we have seen many approaches that do not provide algorithms, our goal in this section is to present an algorithm for UNKNOT RECOGNITION, due to Hass, Lagarias and Pippenger [HLP99] (following ideas of Haken [Hak61]) which shows that UNKNOT RECOGNITION is in **NP**. More broadly, this algorithm is an illustration of the power of normal surface theory, which is an ubiquitous tool in the study of computational problems in low dimensions.

4.1 Normal surface theory

A **normal surface** in T is a properly embedded surface in T that meets each tetrahedron in a possibly empty collection of triangles (cutting off a vertex) and quadrilaterals (separating a pair of vertices), which are called *normal disks*. In each tetrahedron, there are 4 possible types of triangles and 3 possible types of quadrilaterals, pictured in Figure 12. The intersection of a normal surface with a face of the triangulation gives rise to a **normal arc**. There are 3 possible types of normal arcs within each face: the type of a normal arc is defined according to which vertex of the face it separates from the other two.

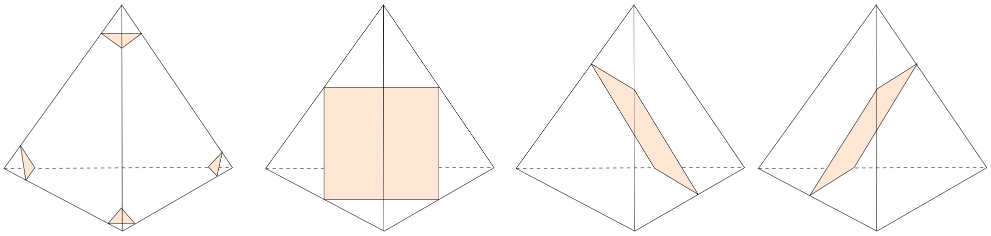

Figure 12: The seven types of normal disks within a given tetrahedron: Four triangles and three quadrilaterals.

With each normal surface S, one can associate a vector, denoted by $[S]$, in $(\mathbb{Z}_+)^{7t}$, where t is the number of tetrahedra in T, by listing the number of triangles and quadrilaterals of each type in each tetrahedron. This vector provides a very compact and elegant description of that surface.

The vector $[S]$ corresponding to a normal surface S, called its **normal coordinates**, satisfies two types of conditions:

- The first type of conditions is the **matching equations**. Consider a normal arc type in a given non-boundary face f of T. This normal arc type corresponds to exactly one triangle normal coordinate, $v_{t,1}$, and one quadrilateral normal coordinate, $v_{q,1}$, in a tetrahedron incident with f. Similarly, let $v_{t,2}$ and $v_{q,2}$ be the triangle and quadrilateral normal coordinates corresponding to that arc type in the opposite tetrahedron. The matching equation for that arc type is $v_{t,1} + v_{q,1} = v_{t,2} + v_{q,2}$. Intuitively, this means that for at a face between two tetrahedra, there are as many objects going in as objects going out. There are no matching equations for faces on the boundary of the triangulation.

- The second type of conditions, the **quadrilateral conditions**, stipulates that, within any tetrahedron, at most one of the three quadrilateral coordinates must be non-zero. Indeed, two quadrilaterals of different types within the same tetrahedron must cross, and therefore this condition is needed to ensure that the surface does not self-intersect.

Conversely, if T is a triangulation of size t and v is a vector in $(\mathbb{Z}_+)^{7t}$, then v corresponds to a normal surface if and only if the matching equations and the quadrilateral conditions are fulfilled. The reconstruction process can be depicted as follows:

- In each tetrahedron, by the quadrilateral conditions, there is at most one type of quadrilateral. One places as many parallel copies of this quadrilateral as needed in the tetrahedron, and then place the parallel triangles next to every vertex of the tetrahedron. It is straightforward to do so without having any intersection between triangles and quadrilaterals.

- One glues the faces on the triangulation together, and in the process, one needs to glue normal arcs, i.e., triangles or quadrilaterals on the one side to triangles and quadrilaterals on the other side. By the matching equations, the numbers fit, and the gluing is imposed by the order in which the normal disks are placed in the tetrahedra.

A **normal isotopy** is an ambient isotopy of M that fixes globally each vertex, edge and face of T. Therefore, a normal surface is represented up to a normal isotopy by a vector in $(\mathbb{Z}_+)^{7t}$ satisfying the matching equations and the quadrilateral conditions. Moreover, given a triangulation and normal coordinates, checking that the matching equations or the quadrilateral conditions hold can trivially be done in linear time.

From this construction, one sees moreover that every vector of normal coordinates corresponds to a unique normal surface, up to a normal isotopy. Therefore, we will often abuse the notation and call both S and $[S]$ a normal surface.

Finally, if one is given normal coordinates, how can one recognize which surface it corresponds to? For usual surfaces it is enough to compute the Euler characteristic and check for orientability, but there is an issue of compression here: a vector of complexity n can correspond to a normal surface with 2^n normal disks, and thus computing the Euler characteristic "by hand" would take an exponential time. But one can do better:

Lemma 4.1. *There exists a linear form e on $\mathbb{Z}_+^{7t}$ such that if $[S]$ is the coordinate vector of a normal surface S, $e([S]) = \chi(S)$ where χ is the Euler characteristic.*

Proof. A normal coordinate describes the number of normal triangles or quadrilaterals of some type in a tetrahedron. One can estimate how much such a triangle or quadrilateral contributes to the Euler characteristic of the entire surface. Let us pick a triangular normal disk t adjacent to edges e_1, e_2 and e_3 of T, and let us denote by v_i the valence of edge e_i, i.e., the number of tetrahedra around it. Then the contribution of t to the Euler characteristic of S is $1/v_1 + 1/v_2 + 1/v_3 - 3/2 - e_\partial/2 + 1$, where e_∂ is the number of edges of t on the boundary of the manifold. Indeed, the vertices of t (on the edges e_1, e_2 and e_3) contribute 1 but are "shared" between all the normal disks adjacent to them. Similarly, the edges contribute -1 but are shared between two normal disks (except for the boundary ones), and the face contribution is exactly once. For a quadrilateral, we get $1/v_1+1/v_2+1/v_3+1/v_4-4/2-e_\partial/2+1$. Summing all of these contributions provides a linear form e which matches with the Euler characteristic. □

4.2 Trivial knot and spanning disks

The apparatus of normal surface theory is designed to study the surfaces embedded in a 3-manifold. In the case of knot complements, this is very relevant to unknot recognition because of the following easy lemma.

Lemma 4.2. *A knot K is trivial if and only if it is the boundary of an embedded disk.*

PROOF. One direction is immediate: if K is trivial, then it is ambient isotopic to the usual embedding of $\mathbb{S}^1$, which obviously bounds a disk. An ambient isotopy is a homeomorphism, thus it preserves this disk, so the knot K also bounds a disk.

The other direction is not much harder: if a knot K bounds a disk, then there is an ambient isotopy that contracts K progressively along this disk until it lies in an arbitrarily small neighborhood of a point, where it will be ambient isotopic to a usual embedding of $\mathbb{S}^1$. □

The starting idea of the algorithm for UNKNOT RECOGNITION is to simply find this disk if it exists. But the triangulation we will work with $\mathbb{S}^3 \setminus N(K)$ instead of just $\mathbb{S}^3$ with K in its 1-skeleton. This is very much needed in order to apply normal surface theory: indeed, we will want to find this disk as a normal surface, and normal surfaces are by construction *transverse* to the 1-skeleton of the triangulation. Therefore, one needs to drill a small tube around K in order to use normal surfaces, we denote this small tube by T_K. The corresponding lemma that we will need is the following one.

Lemma 4.3 ([HLP99, Lemma 4.1])**.** *A knot K is trivial if and only if there exists a disk D in $\mathbb{S}^3 \setminus N(K)$ such that the boundary of D is non-trivial in ∂T_K.*

PROOF. The proof works similarly: If K is trivial, there is an ambient isotopy carrying it to the standard embedding of $\mathbb{S}^1$ into $\mathbb{R}^3$. This ambient isotopy preserves the trivializing disk and the homotopy class of its boundary on ∂T_k, which is therefore non-trivial.

The other direction requires some more work. A **meridian** of a knot is a simple closed curve on T_K bounding a disk in T_K. A **longitude** of a knot is a simple closed curve on T_K crossing the meridian exactly once, and inducing a null-homologous curve in $\mathbb{S}^3 \setminus N(K)$. The existence of these curves follows, for example from the computation of $H_1(\mathbb{S}^3 \setminus N(K))$ in Section 3.2.

If K is non-trivial, let us assume by contradiction that there exists a disk D having boundary γ non-trivial on ∂T_K. The homology of γ on ∂T_K is $a_1[m] + a_2[\ell]$ with $a_1, a_2 \neq (0,0)$ where $[m]$ and $[\ell]$ denote the homology classes induced on ∂T_K by the meridian and the longitude, respectively. Via the inclusion map $\partial T_k \hookrightarrow \mathbb{S}^3 \setminus N(K)$, γ has a homology class in $H_1(\mathbb{S}^3 \setminus N(K))$, which is a_1 by definition of the meridian and the longitude. Since γ bounds a disk, we thus have $a_1 = 0$, and thus $a_2 = \pm 1$ by simplicity of γ. Therefore K cobounds an annulus with γ within T_K. Gluing this annulus with the disk D, we obtain a disk bounded by K and we can apply the previous lemma. □

Let us call a disk satisfying the above properties a **spanning disk**. The key theorem behind the **NP** algorithm is the following one.

Theorem 4.4. *Let K be a trivial knot and T be a triangulation of $\mathbb{S}^3 \setminus N(K)$ obtained by the process of Section 3.3. Then there exists a spanning disk of K that is normal with respect to T and of which normal coordinates are bounded by $2^{O(t)}$.*

Assuming Theorem 4.4 for now, we can (almost) provide the advertised algorithm. Naively, this should be very simple: now that we have established that there exists a spanning disk that is normal and has coordinates of bounded size, one can simply use this spanning disk (or rather its normal coordinates) as a certificate. The size of the coordinates might be exponential, but exponential numbers can be encoded using only a polynomial number of bits (this trivial observation is what makes everything work!), thus the certificate has polynomial size. But there is a hidden issue lurking here: one also needs to verify in polynomial time that the certificate is indeed a spanning disk. This should be easy in principle, but as we have already mentioned, the certificate is very **compressed** due to this integer encoding, and it turns out to be non-trivial.

Corollary 4.5. UNKNOT RECOGNITION *is in* ***NP****.*

PROOF. Starting with a knot K, described by a diagram with n crossings, one first builds a triangulation of its complement following Section 3.3. This triangulation has $O(n)$ tetrahedra. The certificate is then the normal spanning disk promised by Theorem 4.4. It is a vector in $O(n)$ dimensions and the size of the coordinates is bounded by $2^{O(n)}$, which can be encoded with a number of bits polynomial in n.

Once one is given the certificate, one can verify that is it indeed a spanning disk in the following way:

1. One first verifies that this is indeed a normal surface S, by checking that the matching and quadrilateral constraints are satisfied.
2. One checks whether S is connected.
3. One checks whether S is a disk.
4. One checks whether the boundary of S is non-contractible.

If all the answers are positive, then we have a spanning disk and K is unknotted. Step 1 can be easily done in polynomial time since the matching equations are linear, and the quadrilateral constraints are easy to verify by looking at each coordinates. Let us assume that Step 2 has been done for now. In order to do step 3, it is enough to compute the Euler characteristic of S, which can be done in polynomial time using the linear form of Lemma 4.1. For step 4, from the normal coordinates of S one can obtain normal coordinates[3] of ∂S. Since the fundamental group of the torus is $\mathbb{Z}^2$, testing whether ∂S is contractible can be done in polynomial time from these normal coordinates.

However, Step 2 is hard. The naive algorithm, which would just follow a starting normal disk and count the resulting connected components could require *exponential* time, since the size of the normal coordinates could be exponential. A polynomial algorithm to check connectivity of normal surfaces was designed by Agol, Hass and Thurston [AHT06] and can be used here, but we will not delve into this. □

The rest of this section is devoted to the proof of Theorem 4.4. We first show how to normalize a spanning disk, and then how to prove the existence of one with suitably bounded coordinates.

[3] More formally, ∂S is a curve on a torus, and the coordinates we obtain describe this curve with respect to the triangulation of the torus inherited from T. The underlying theory of **normal curves** is similar (but much simpler) than for normal surfaces.

4.3 Normalization of spanning disks

A key observation, initially due to Haken [Hak61], is that there exists a spanning disk that is a normal surface:

Lemma 4.6. *Let K be a trivial knot and T be a triangulation of $\mathbb{S}^3 \setminus N(K)$ obtained by the process of Section 3.3. Then there exists a spaning disk that is normal with respect to T.*

Proof. The proof proceeds by starting with a spanning disk D (which exists by Lemma 4.3) and *normalizing* it. We can first, by using a small perturbation, put D in general position with respect to the triangulation T: i.e., all the intersections between D and T can be assumed to be transverse. Now, we look at what can go wrong, i.e., which pieces in D are not normal with respect to T. Let us write $c(D) = (\mathrm{wt}_1(D), \mathrm{wt}_2(D))$, where wt_1 and wt_2 respectively denote the number of connected components of the intersection of D with the edges of T and the faces of T, and order the pairs $c(D)$ with the lexicographic order. There are many different occurences of non-normality which we deal with distinct moves, but each time the complexity $c(D)$ will only go down. Since this complexity is finite to begin with, after a finite number of steps, the process will finish and we will have a normal surface.

1. If D intersects an inner face F of T in an arc that hits twice the same interior edge e of T, one can push D locally to reduce its number of intersections with e, see Figure 13.

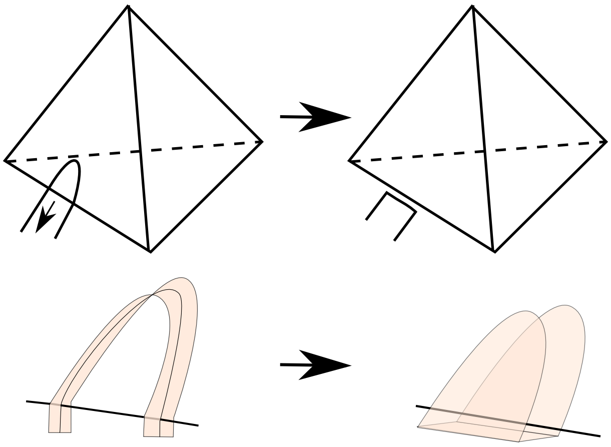

Figure 13: If there are excess intersections with an interior face and an interior edge, one can push the disk to reduce its intersections with the edges of T. The first picture shows what the move induces on the face F and the second one is the corresponding 3-dimensional move.

2. If D intersects a boundary face F of T in an arc that hits twice the same boundary edge e of T, one can push D similarly, reducing its number of intersections with e, see Figure 14. The boundary of D is moved by this operation, but only by a homotopy, hence the resulting disk is also a spanning disk.

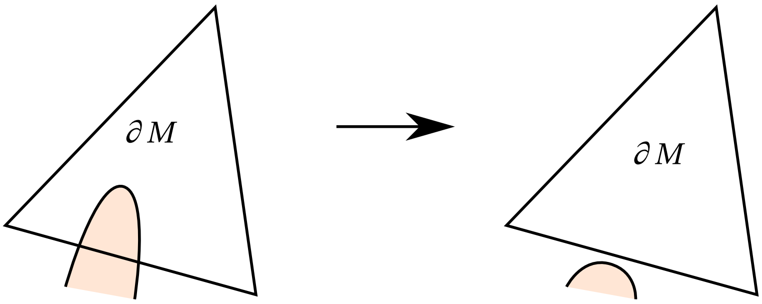

Figure 14: If there are excess intersections with a boundary face, one can push the disk to reduce its intersections with the edges of T

3. If D intersects a face F in a cycle disjoint from its edges (forming locally a tube), one can "cut" (or **compress**) this tube, which will reduce the number of intersections of D with F. Such a compression cuts the disk into a sphere and a disk, and one can continue the normalizing process with the disk, which will have smaller complexity, see Figure 15.

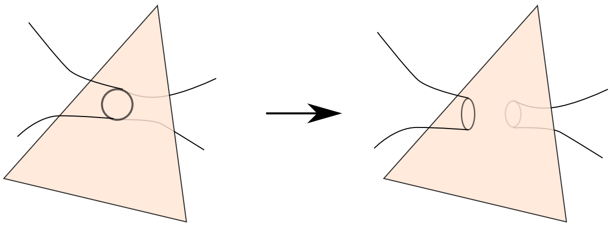

Figure 15: If the disk forms a tube crossing a face of T, one can cut this tube and keep one of the components to reduce the number of intersections with the faces of T.

4. If D intersects a tetrahedron t in a tube (i.e., locally, D intersects ∂t in two components), then one can compress this tube as well. For the same reasons, by picking the remaining disk, the complexity will have decreased, see Figure 16.

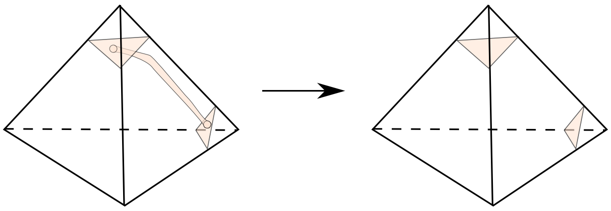

Figure 16: If the disk forms a tube inside a tetrahedron, one can cut this tube and keep one of the components to reduce the number of intersections with the edges of T.

5. If D intersects an inner face F of T in an arc a that hits twice the same boundary edge e of T, one can "cut" (or **boundary compress**) along the disk that is bounded by a and e. This has the effect of cutting the disk D into two subdisks, one of which must be spanning. Taking this one reduces the complexity, see Figure 17.

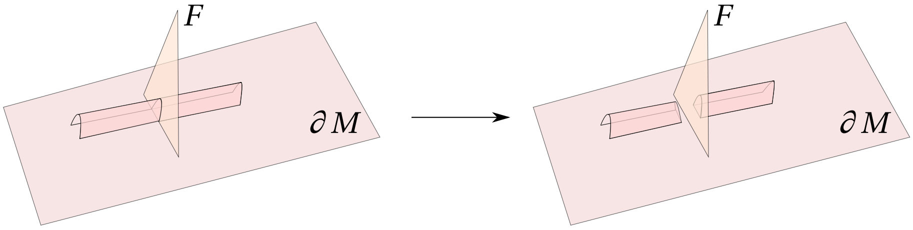

Figure 17: If there are excess intersections with an interior face and a boundary edge, one can push the cut the disk and keep one of the components to reduce its intersections with the edges of T.

6. Finally, D might be locally too complicated inside a tetrahedron t, i.e., $D \cap t$ can have more than 4 arcs. In this case one can show that there are at least 8 arcs, and some edge is hit at least twice. Then, if that edge is not on the boundary of the triangulation, one can reduce the complexity by pushing D towards that edge, see Figure 18. Otherwise, one can do a boundary compression as in the previous step.

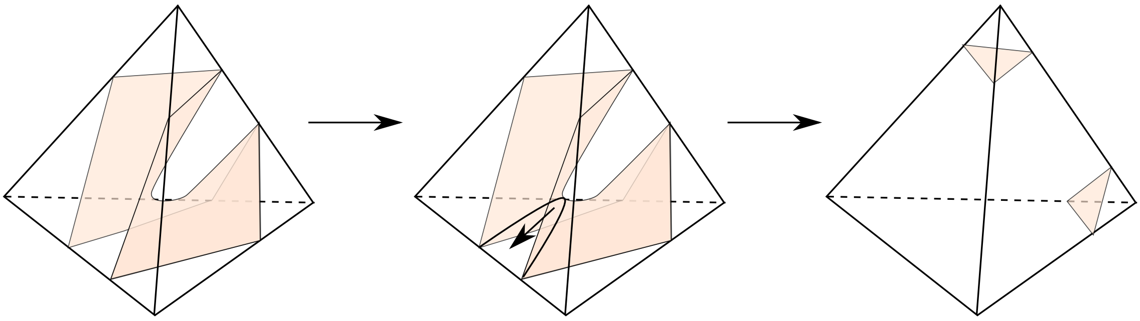

Figure 18: If there is a piece inside a tetrahedron with more than 4 arcs, one can push it to simplify it.

Note: For all these moves, there might be other pieces of the disk in the way of the indicated move. This is resolved by always applying first the moves corresponding to an innermost disk, i.e., first normalizing the non-normal behavior closest the the boundary of the tetrahedron.

Once none of these non-normal cases happen, the disk D intersects the triangulation T in a normal way, hence we have found a normal spanning disk. □

We have thus established that it suffices to find a normal spanning disk to certify that a knot is trivial. Since normal surfaces can be described by a vector in $\mathbb{Z}_+^{7t}$, this would give an algorithm if one could bound the size of these coordinates. Such a bound will be established by exploiting the additive structure on normal surfaces provided by these vectors.

4.4 Haken sum, fundamental and vertex normal surfaces

The set of vectors of $\mathbb{R}_+^{7t}$ verifying the matching equations is called the **Haken cone** $\mathscr{C}$. The normal surfaces are the integral points in this cone that also satisfy the quadrilateral constraints. It they have no conflicting quadrilaterals, two normal surfaces can be added by adding their vectors, and the result will still be a normal surface since the matching equations are linear. This operation is called the **Haken sum** of normal surfaces. Note that by Lemma 4.1, if S is the Haken sum of two normal surfaces S_1 and S_2, $\chi(S) = \chi(S_1) + \chi(S_2)$.

A normal surface $[S]$ is called **fundamental** if it can not be written as a sum $[S] = [S_1] + [S_2]$ with $[S_1]$ and $[S_2]$ two non-empty normal surfaces. A fundamental normal surface $[S]$ is a **vertex normal surface** if $c[S] = c_1[S_1] + c_2[S_2]$ for positive integers c, c_1 and c_2 implies that $[S_1]$ and $[S_2]$ are multiples of $[S]$. Fundamental and vertex normal surfaces are the building blocks for normal surfaces, and crucially, one can bound their complexity:

Lemma 4.7.

- *Let $[S]$ be a vertex normal surface in a triangulation T with t tetrahedra. Then the normal coordinates of S have size bounded by $2^{O(t)}$.*
- *Let $[S]$ be a fundamental normal surface in a triangulation T with t tetrahedra. Then the normal coordinates of S have size bounded by $2^{O(t)}$.*

PROOF.

- Let us intersect the cone $\mathscr{C}$ with the hyperplane $H = \sum_i x_i = 1$. This forms a polyhedron $\mathscr{P}$ and the vertex normal surfaces will be obtained as the first integral multiples of some of the vertices of $\mathscr{P}$. Now, the vertices of $\mathscr{P}$ are obtained as a solution of $7t$ equations, which either come from the matching equations, from the hyperplane H or from a hyperplane of the form $x_i = 0$. Thus, such a vertex v verifies $Mv = (0, \dots, 0, 1)^T$ for some matrix M with entries in $\{-1, 0, 1\}$. By Cramer's rule, the coordinates v_i are obtained by the quotient $\det M_i / \det M$, where M_i is the matrix M where the ith column has been replaced by $(0, \dots 0, 1)^T$. One can bound these determinants using Hadamard's inequality $(\det M)^2 \leq \prod_i ||r_i||^2$ where r_i are the rows of M. We obtain $|\det M_i| = 2^{O(t)}$ and $|\det M| = 2^{O(t)}$, and thus $v_i = 2^{O(t)}$. Then, the size of the coordinates of the vertex normal surfaces is bounded by $v_i |\det M| = 2^{O(t)}$ as well.
- Let $[S]$ be a fundamental normal surface, then multiples of $[S]$ can be decomposed on the vertex normal surfaces: $c[S] = \sum c_i[S_i]$, or equivalently $[S] = \sum c_i/c[S_i]$. Note that $c_i/c \leq 1$, otherwise one would have $[S] = ([S] - [S_i]) + [S_i]$ which would be a non-trivial integral decomposition of S, a contradiction. Thus any coordinate of $[S]$ is at most the sum of the coordinates of the vertex normal surfaces $[S_i]$, of which there are at most $2^{O(t)}$ (one can for example bound the number of matrices M involved in the previous item). This concludes the proof.

□

A common principle in normal surface theory is that "interesting" surfaces in a 3-manifold can be found among the fundamental normal surfaces, and even sometimes among the vertex normal surfaces of the triangulation. This turns out to be true for spanning disks.

Proposition 4.8. *Let K be a trivial knot and T be a triangulation of $\mathbb{S}^3 \setminus N(K)$ obtained by the process of Section 3.3. Then there exists a spanning disk that is a fundamental normal surface with respect to T.*

Combining Lemma 4.7 and Proposition 4.8 directly proves Theorem 4.4, and thus the **NP** algorithm (modulo the connectivity issue already mentioned).

Remark: The main issue in the **NP** algorithm is to check connectivity in polynomial time. One way to circumvent it could be to verify that the certificate describes a fundamental normal surface, since fundamental normal surfaces are connected. But there is no easy way to do that either. On the other hand, one can certify that a normal surface is a vertex normal surface, by exhibiting the family of $7t$ equations it satisfies (see the proof of Lemma 4.7). And it is also true, but harder to prove, that there exists a spanning disk that is a vertex normal surface: we refer to Jaco and Tollefson [JT95] for a proof. Thus, if one admits this, it gives an alternative way to provide a **NP** certificate.

There remains to prove Proposition 4.8.

PROOF OF PROPOSITION 4.8. The main idea of the proof is to use the Euler characteristic as an accounting device. Indeed, since the Euler characteristic is linear on the normal coordinates, one can use it to discard the vast majority of the bad cases, and the last ones will be handled by hand. More precisely, let D be a normal spanning disk. We can assume that its boundary crosses at most once every triangle of ∂M and that it is of minimal complexity subject to this. If it is not fundamental, it can be written as a sum $[D] = [S_1] + [S_2]$ where the $[S_i]$ are non-trivial normal surfaces, and thus $\chi(S_1) + \chi(S_2) = 1$. Among all the decompositions into S_1 and S_2, let us pick the one that minimizes the number of connected components of $S_1 \cap S_2$.

We first claim that that S_1 and S_2 are connected. Indeed, if S_1 is not connected and consists of two disjoint connected components A and B, since D is connected, S_2 intersects both A and B. But then $[D] = [S_2 + A] + [B]$ and this decomposition has less intersections than S_1 and S_2, contradicting our assumption.

Then, since the Euler characteristic of a connected surface is at most 2, there are only a few cases to deal with. The favorable one is when S_1 is a disk and S_2 is a torus, then S_1 has the same boundary as D, and thus is a spanning disk of smaller complexity, a contradiction. There remains to discard the bad cases (note that since we are in $\mathbb{R}^3$, unpunctured projective planes and Klein bottles can be ruled out straight away) :

1. S_1 is a punctured torus or Klein bottle and S_2 is a sphere.

2. S_1 is a Möbius band or an annulus and S_2 is a disk.

In order to do so, we must pause a bit and figure out what a Haken sum means geometrically: if S_1 and S_2 are two intersecting normal surfaces, their Haken sum is obtained by taking the normal disks of S_1 and S_2 and "reconnecting" them differently. This amounts geometrically to looking at the intersection curves between S_1 and

S_2, cutting along them and doing a **switch**, as pictured in Figure 19. There are two possibilities for a switch, and the one carried out depends on the situation of the intersecting normal disks inside a tetrahedron : both switches gives surfaces, but only one gives a normal surface. But if one performs the “bad” switch, one can still normalize the resulting surface to obtain a normal surface.

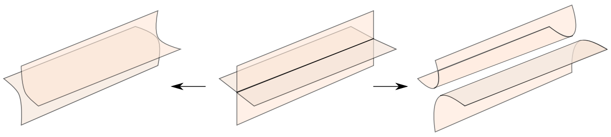

Figure 19: The two switches at an intersection curve.

For the first case, let α be a curve of intersection between S_1 and S_2. We first claim that α is not separating in the punctured torus S_1. Indeed, otherwise, in the case where S_1 is a punctured torus, by cutting S_1 along α and performing the switch that patches a disk of S_2 bounded by α on this cut, and normalizing if necessary, one would obtain either a spanning disk of less complexity, or a decomposition of D into two normal surfaces which intersect less than S_1 and S_2, contradicting our assumptions. If S_1 is a punctured Klein bottle, there is a third case if α cuts the Klein bottle into a Möbius band and a Möbius band with an additional boundary, but pasting the Möbius band on the disk of S_2 yields an immersion of the projective plane in $\mathbb{R}^3$ without triple points, which is impossible [Ban74].

Now, since S_1 intersects S_2 along at least two non-contractible curves, and we pick the two outermost such curves, i.e., the pair of non-contractible curves closest to ∂S_1 on S_1, see Figure 20. When one cuts along these curves and glue disks coming from S_2, one obtains a disk with boundary ∂S_1. Since this cut-and-pasting corresponds to a (good or bad) switch, after normalizing if needed, we obtain a spanning disk of lower complexity than D, a contradiction.

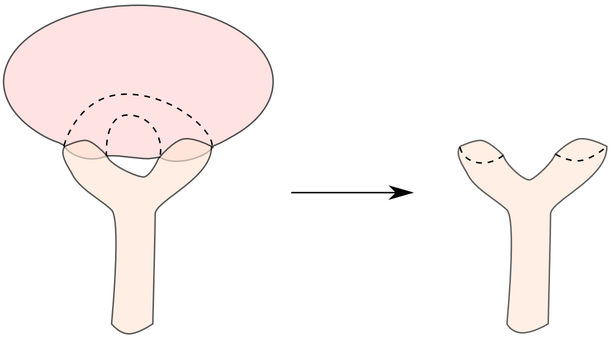

Figure 20: One can cut the torus S_1 along non-contractible cycles and patch it with disks of S_2 to obtain a spanning disk of lower complexity.

In the second case, ∂S_2 and ∂S_1 cross each other, which cannot happen since ∂D crosses at most once each triangle of ∂M. Thus case 2 is ruled out, and this concludes the proof. □

5 Knotless graphs

To conclude this chapter with a striking open problem, let us discuss a bit about knotless graphs. A polygonal embedding of a graph G into $\mathbb{R}^3$ is **knotless** if every simple cycle of G is mapped to a trivial knot by this embedding. A graph G is **knotless** if it admits a knotless embedding, and **intrinsically knotted** otherwise. It is not obvious that there exist intrinsically knotted graphs at all (remember that is was not obvious that there existed non-trivial knots either), but one can prove, for example using the *Arf invariant* that K_7 is intrinsically knotted [CG83].

As we discussed in this chapter, no polynomial-time algorithm for UNKNOT RECOGNITION is known. Therefore, recognizing whether a given embedding of graph is knotless in polynomial time is also out of reach for current techniques. One could expect recognizing knotless graphs to be even harder since naively, it amounts to making this test for every possible embedding of G into $\mathbb{R}^3$. Therefore the following proposition might come as a shock.

Proposition 5.1. *There exists an algorithm to recognize knotless graphs in polynomial time.*

PROOF. If H is a minor of G and G is knotless, H is knotless as well: if i is a knotless embedding of G, every simple cycle of H corresponds to a simple cycle of G and is thus mapped by i to a trivial knot. Thus knotless graphs form a minor-closed family, and it follows from Robertson-Seymour theory (see for example [Lov05] for an introduction) that they can be recognized in polynomial time. □

The proof of this proposition is shockingly unsatisfying: not only is the algorithm, as most algorithms coming from Robertson-Seymour theory, extremely inefficient, but we actually *do not know* what it is: what the theory proves is that minor-closed families are characterized by a finite family of forbidden minors, and testing for forbidden minors can be done in polynomial time – but we do not know what these are. It is an open problem to find an *explicit* polynomial-time algorithm to recognize knotless graphs.

References

[AFW15] Matthias Aschenbrenner, Stefan Friedl, and Henry Wilton. Decision problems for 3–manifolds and their fundamental groups. *Geometry & Topology Monographs*, 19(1):201–236, 2015.

[AHT06] Ian Agol, Joel Hass, and William Thurston. The computational complexity of knot genus and spanning area. *Transactions of the American Mathematical Society*, 358:3821–3850, 2006.

[Ban74] Thomas Banchoff. Triple points and singularities of projections of smoothly immersed surfaces. *Proceedings of the American Mathematical Society*, 46(3):402–406, 1974.

[BZ85] Gerhard Burde and Heiner Zieschang. *Knots.* De Gruyter, 1985.

[CG83] John H. Conway and Cameron McA. Gordon. Knots and links in spatial graphs. *Journal of Graph Theory*, 7(4):445–453, 1983.

[GL89] Cameron Gordon and John Luecke. Knots are determined by their complements. *J. Amer. Math. Soc.*, 2(2):371–415, 1989.

[Hak61] Wolfgang Haken. Theorie der Normalflachen, ein Isotopiekriterium für den Kreisnoten. *Acta Mathematica*, 105:245–375, 1961.

[Hat02] Allen Hatcher. *Algebraic topology.* Cambridge University Press, 2002. Available at `http://www.math.cornell.edu/~hatcher/`.

[HLP99] Joel Hass, Jeffrey C. Lagarias, and Nicholas Pippenger. The computational complexity of knot and link problems. *Journal of the ACM*, 46(2):185–211, 1999.

[Jac05] William Jaco. Peking summer school, 2005. Lecture Notes available at `https://www.math.oakstate.edu/~jaco/pekinglectures.htm`.

[JT95] William Jaco and Jeffrey L. Tollefson. Algorithms for the complete decomposition of a closed 3-manifold. *Illinois Journal of Mathematics*, 39(3):358–406, 1995.

[Kup14] Greg Kuperberg. Knottedness is in NP, modulo GRH. *Advances in Mathematics*, 256:493–506, 2014.

[Kup15] Greg Kuperberg. Algorithmic homeomorphism of 3-manifolds as a corollary of geometrization. *arXiv:1508.06720*, 2015.

[Lac15] Marc Lackenby. A polynomial upper bound on Reidemeister moves. *Ann. Math. (2)*, 182(2):491–564, 2015.

[Lac16] Marc Lackenby. The efficient certification of knottedness and thurston norm. arXiv:1604.00290, 2016.

[Lov05] László Lovász. Graph minor theory. *Bulletin of the AMS*, 43(1):75–86, 2005.

[Moi52] Edwin E. Moise. Affine structures in 3-manifolds. V. The triangulation theorem and Hauptvermutung. *Ann. of Math. (2)*, 56:96–114, 1952.

[Per02] Grisha Perelman. The entropy formula for the Ricci flow and its geometric application. arXiv:math/0211159, 2002.

[Per03] Grisha Perelman. Finite extinction time for the solutions to the Ricci flow on certain three-manifolds. arXiv:math/0307245, 2003.

[Rei27] Kurt Reidemeister. Elementare begründung der knotentheorie. *Abhandlungen aus dem Mathematischen Seminar der Universität Hamburg*, 5(1):24–32, 1927.