Topology, Groups, and Knots

Zach Conn
Rice University

Spring 2010

Contents

1 Preliminaries

These notes summarize the most basic definitions and results in algebra, topology, and knot theory. They are intended to be as self-contained as possible, so they depart from almost no prerequisites.

We begin by fixing notations and introducing fundamental concepts that will be in use constantly. We adopt the conventions of naive set theory and leave the term "set" undefined as well as relations such as set membership.

If A, B are arbitrary sets, a **function** from A to B is a rule that assigns to each element of A exactly one element of B. If a function is denoted by the symbol f and $a \in A$, then the element of B that corresponds to a is denoted by $f(a)$ and is called the **image** of a under f. To indicate that the symbol f refers to a function from A to B, one often writes $f : A \to B$. The set A is the **domain** of f and B is the **codomain** of f.

The concept of image can be generalized from elements of the domain A to subsets of A. If $U \subseteq A$ is any subset of A, then the symbol $f(U)$ refers to the set $\{f(x) : x \in U\}$ of images of elements of U under f. In particular, $f(A)$ is simply called the **image** of f itself. In general $f(A) \subseteq B$.

In case $f(A) = B$, the function f is called **surjective** (or **onto**). If f is surjective and $b \in B$, then there exists at least one element $a \in A$ such that $f(a) = b$; a is a **preimage** of b under f. If $V \subseteq B$ is an arbitrary subset of B, then the symbol $f^{-1}(V)$ refers to the set $\{a \in A : f(a) \in V\}$ of preimages of elements of V under f.

If $f(x) = f(y)$ implies $x = y$ for all $x, y \in A$, then f is called **injective** (or **one-to-one**). By considering the contrapositive one obtains an equivalent characterization: f is injective if $x \neq y$ implies $f(x) \neq f(y)$ for all $x, y \in A$.

A function that is at once injective and surjective is called **bijective**; such a function is a **one-to-one correspondence**. The terms injection, surjection, and bijection are occasionally used to refer to functions that are respectively injective, surjective, and bijective.

2 Topology

We proceed to a review of the notions from topology that will be useful in our study of knots and links.

2.1 Homeomorphisms

Recall that a function $f : U \to V$ from one topological space U to another topological space V is **continuous** if the preimage of an open set in V is open in U.

Definition 2.1. A **homeomorphism** from a topological space U to a space V is a function $f : U \to V$ that is continuous, bijective, and has a continuous inverse. Two spaces U and V are **homeomorphic** if there exists a homeomorphism from one to the other.

Example 2.2. Fix two positive real numbers a, b. The closed unit disc $D = \{x \in \mathbb{R}^2 : |x| \leq 1\}$ is homeomorphic to the solid ellipse $E = \{u(a\cos(v), b\sin(v)) : u \in [0,1], v \in [0, 2\pi)\}$. In fact, the function $f : D \to E$ sending a point $(x, y) \in D$ in the disc to the point $(ax, by) \in E$ in the ellipse is a homeomorphism. Its inverse is the function sending the point $(x, y) \in E$ in the solid ellipse to the point $(x/a, y/b) \in D$ in the disc, and both f and f^{-1} are clearly continuous.

Non-Example 2.3. Consider the function $f : [0, 2\pi) \to S^1 = \{x \in \mathbb{R}^2 : |x| = 1\}$ that sends $\theta \mapsto (\cos(\theta), \sin(\theta))$. If U is an open subset of $\mathbb{R}^2$ containing $(1, 0)$, then the preimage $f^{-1}(U \cap S^1)$ will contain 0 but will not contain any neighborhood of 0 and hence will not be open. Consequently, the function f is not continuous everywhere and is therefore not a homeomorphism. This is not surprising: S^1 is compact whereas $[0, 2\pi)$ is not, and compactness is preserved by continuous functions.

2.2 Manifolds

If $M \subseteq \mathbb{R}^n$ is a subset of n-dimensional Euclidean space, then M is an **m-dimensional manifold** if for each point $x \in M$ there is an open set $U \in \mathbb{R}^n$ such that

- U contains x and
- the set $U \cap M$ is homeomorphic to $\mathbb{R}^m$.

Thus, the subset M is endowed with the subspace topology it inherits from $\mathbb{R}^n$ and each point is required to admit an open neighborhood homeomorphic to $\mathbb{R}^m$.

Example 2.4. The n-sphere $S^n = \{x \in \mathbb{R}^{n+1} : |x| = 1\}$ is always an n-dimensional manifold, and we will check explicitly that $S^2 \subseteq \mathbb{R}^3$ is a two-dimensional manifold. If $x \in S^2$ is any point on the sphere besides the north pole $(0, 0, 1)$, then $S^* = S^2 - \{(0, 0, 1)\}$ is an open subset of S^2 containing x. In fact, S^* is homeomorphic to the plane $\mathbb{R}^2$, and the correspondence can be achieved explicitly via **stereographic projection** from $(0, 0, 1)$.

To carry out the projection, associate to the point $x \in S^*$ the intersection of the equatorial plane of S^2 and the line in $\mathbb{R}^3$ containing $(0, 0, 1)$ and x. If $x = (x_1, x_2, x_3)$, then this line is the image of the function $t \mapsto (1-t)(0, 0, 1)+ tx = (tx_1, tx_2, 1 - t + tx_3)$ for $t \in \mathbb{R}$. Hence this line intersects the equatorial plane when $t = 1/(1 - x_3)$, so the intersection of the line and the equatorial plane has coordinates $(x_1/(1 - x_3), x_2/(1 - x_3), 0)$ as a point in $\mathbb{R}^3$; this is the stereographic projection of x onto the equatorial plane.

With this motivation in place, we define stereographic projection from S^* to the plane $\mathbb{R}^2$ to be the function sending

$$(x_1, x_2, x_3) \in S^* \mapsto \left(\frac{x_1}{1 - x_3}, \frac{x_2}{1 - x_3} \right). \tag{1}$$

This function is continuous, but it must be verified that it is bijective if projection is to be used to show that S^* is homeomorphic to the plane. To do this, let (a, b) represent the coordinates in the equatorial plane of the projection of $x = (x_1, x_2, x_3) \in S^*$.

Suppose x_3 is fixed but x_1, x_2 vary so long as x lies on S^*. Then the points (x_1, x_2, x_3) occupy the intersection of S^* with a horizontal plane and therefore trace out a horizontal circle; the projection of this circle onto the plane is an origin-centered circle whose radius depends only on $x_1^2 + x_2^2$. This heuristic suggests the following procedure for computing x in terms of its projection (a, b).

Equation (1) implies $|(a, b)|^2 = (x_1^2 + x_2^2)/(1 - x_3)^2$. But $x_1^2 + x_2^2 = 1 - x_3^2$, so

$$|(a, b)|^2 = \frac{(1 + x_3)(1 - x_3)}{(1 - x_3)^2} = \frac{1 + x_3}{1 - x_3}$$

and

$$x_3 = \frac{|(a, b)|^2 - 1}{|(a, b)|^2 + 1}.$$

On substituting this expression for x_3 into (1) we obtain

$$x_1 = a(1 - x_3) = \frac{2a}{1 + |(a, b)|^2} \text{ and}$$
$$x_2 = b(1 - x_3) = \frac{2b}{1 + |(a, b)|^2}.$$

Therefore, each point in the plane $(a, b) \in \mathbb{R}^2$ has a unique preimage under stereographic projection, so the projection is bijective. One must finally check that the inverse of the projection is continuous, and this is clear from the three equations that determine $(x_1, x_2, x_3) \in S^*$ in terms of $(a, b) \in \mathbb{R}^2$.

To complete the proof that S^2 is a manifold, one must find a neighborhood of the north pole $(0, 0, 1)$ that is homeomorphic to $\mathbb{R}^2$. In analogy with what we have just done, we show that the open neighborhood $S_* = S^2 - \{(0, 0, -1)\}$ of $(0, 0, 1)$ is homeomorphic to $\mathbb{R}^2$ using stereographic projection from the south pole $(0, 0, -1)$. The procedure is precisely the same, so we will omit the details.

2.3 Covering spaces

The definitions here are inspired by those presented in [1]. Suppose C, X are topological spaces, U is an open subset of X, and $\phi : C \to X$ is a continuous surjective function. The subset U is **evenly covered** by the map ϕ if

- its preimage $\phi^{-1}(U)$ is the disjoint union of open subsets of X and
- the restriction of ϕ to each of these open sets is a homeomorphism onto U.

In the special case that every point $x \in X$ has an open neighborhood U evenly covered by ϕ we say that C is a **covering space** of X and that ϕ is a **covering map**; we also refer to X as the **base space**.

Example 2.5. Let C be the real line $\mathbb{R}$ with its usual topology and let X be the unit circle S^1 considered as a subspace of the plane $\mathbb{R}^2$. Define the function $\phi : C \to X$ by $\phi(t) = (\cos(2\pi t), \sin(2\pi t))$. Let us verify that C is then a covering space of X. Any point in S^1 has the form $(\cos(2\pi\theta), \sin(2\pi\theta))$ for some $\theta \in \mathbb{R}$, so select a point on S^1 by fixing a value of θ. Put $U =$

$\{(\cos(2\pi\alpha), \sin(2\pi\alpha)) : |\alpha - \theta| < 1/3\}$ and observe that the preimage $\phi^{-1}(U)$ is the disjoint union

$$\coprod_{n\in\mathbb{Z}} (n + \theta - 1/3, n + \theta + 1/3);$$

each interval in this union is mapped homeomorphically by ϕ onto U. (The constant $1/3$ is simply chosen so that the intervals in the above union are disjoint.)

Non-Example 2.6. Change C to be the interval $(0, 1]$ but leave the rest of the previous example unchanged. Now $U = \{(\cos(2\pi\alpha), \sin(2\pi\alpha)) : |\alpha| < 1/3\}$ is an open neighborhood of $(1, 0) \in S^1$, but its preimage $\phi^{-1}(U) = (0, 1/3) \cup (2/3, 1]$ is not a disjoint union of open subsets of C.

Example 2.7. As a general phenomenon, products of covering spaces are covering spaces, and this can be illustrated by considering the torus $T = S^1 \times S^1$. Thus, if C, X, and ϕ are as in the first example, then the function $\phi \times \phi : C \times C \to T$ defined by

$$\phi(t_1, t_2) = (\cos(2\pi t_1), \sin(2\pi t_1), \cos(2\pi t_2), \sin(2\pi t_2))$$

is a covering map and $C \times C$ is a covering space of the torus.

Example 2.8. A comparatively exotic example is obtained by considering the complex logarithm $\log(z) = \log|z| + i \arg(z)$. Notice that if z travels around the origin in a smooth closed loop then $\log(z)$ will vary continuously with z but will never return to its original value; this occurs because the complex exponential $\exp(z)$ maps vertical lines into origin-centered concentric circles.

This situation can be remedied in the following manner. If $\mathbb{R}_+$ is the set of positive real numbers, put $C = \mathbb{R} \times \mathbb{R}_+$, $X = \mathbb{C} - \{0\}$, and define the function $\phi : C \to X$ by $\phi(\theta, r) = \log(re^{i\theta})$. Then C is a covering space of X and ϕ represents a single-valued version of the complex logarithm defined on C, which may be considered a morphed version of $\mathbb{C}$. C is called the **Riemann surface** of the complex logarithm.

3 Groups

We now make a change of pace and consider the fundamental constructions of group theory. These constructions, despite being significant in themselves,

will interact with the topological ideas already discussed in important and useful ways.

3.1 Binary operations

If S is a set, then a **binary operation** on S is just a function $S \times S \to S$.

Example 3.1. Addition is a binary operation on $\mathbb{Z}$ but division is not (because the result of dividing 1 by 2, for example, is not in $\mathbb{Z}$).

Although binary operations are functions, they are rarely written using functional notation. A symbol such as $*$ is usually used, and we denote the result of applying the operation to elements a, b (in this order) by the expression $a * b$. In certain contexts the result of applying the operation to elements a, b is written as the juxtaposition ab.

A binary operation $*$ on a set S is **associative** if $(a * b) * c = a * (b * c)$ for all $a, b, c \in S$.

Example 3.2. Addition is an associative binary operation on $\mathbb{Z}$. The cross product is a binary operation on $\mathbb{R}^3$ but is not associative.

Associativity is an example of a situation where functional notation is inconvenient. If we used the symbol f to represent the operation and used functional notation, the relation for associativity would be the clumsy equation $f(f(a, b), c) = f(a, f(b, c))$.

If $*$ is associative, then the result of applying $*$ to a finite string of elements of S is independent of how this string is parenthesized; therefore one can write without ambiguity $a_1 * a_2 * \cdots * a_n$ for any finitely many elements $a_1, \ldots, a_n \in S$. A formal proof of this proceeds by induction on n.

A binary operation $*$ on S is **commutative** if $a * b = b * a$ for all $a, b \in S$.

Example 3.3. Addition is a commutative binary operation on $\mathbb{Z}$. Subtraction is a noncommutative binary operation on $\mathbb{Z}$ since $a - b = -(b - a)$ for $a, b \in \mathbb{Z}$.

Example 3.4. The set of invertible real $n \times n$ matrices is called the **general linear group** and denoted $GL_n(\mathbb{R})$. Matrix multiplication is a noncommutative binary operation on $GL_n(\mathbb{R})$. Similarly, the set of real $n \times n$ matrices with determinant 1 is the **special linear group** $SL_n(\mathbb{R})$, and multiplication is also a noncommutative binary operation on $SL_n(\mathbb{R})$.

3.2 Groups

Group theory is an abstract stage in which it's convenient to study some of the most common themes that arise in various mathematical systems. The definition of a group, the fundamental abstraction of the theory, can be phrased nicely with binary operations having been developed separately:

Definition 3.5. A **group** is a pair $(G, *)$ where G is a set and $*$ is an associative binary operation on G such that the following two conditions hold:

(A) There exists an element $e \in G$ such that $e * a = a * e = a$ for all $a \in G$. This element e is called an **identity**.

(B) For each $a \in G$ there exists an element $b \in G$ such that $a * b = b * a = e$, where e is the identity whose existence was postulated in the previous condition. b is called an **inverse** of a.

The operation is *not* required to be commutative. A group in which this extra condition holds is called an **abelian group**. As a slight abuse of language, we will usually refer to the set G as the group when the operation is understood.

In a group G, the result of applying the operation $*$ is almost universally denoted by juxtaposition. Thus one writes ab instead of $a * b$. The addition symbol $+$ is also occasionally used, but this is traditionally reserved only for abelian groups.

We record at once two simple results:

Lemma 3.6. *If G is a group, then*

(A) *there is exactly one identity element and*

(B) *each element $a \in G$ has exactly one inverse.*

Proof.

(A) If e, e' are identities in G, then $e = ee' = e'$.

(B) (B) If $a \in G$ and $b, b' \in G$ are inverses of a, then $ab = e$. Multiply on the left by b' to obtain $b'ab = b'e$. The left member is by associativity $(b'a)b = eb = b$ and the right member is $b'e = b'$, so $b = b'$.

□

By virtue of this result one can speak of *the* identity in a group and *the* inverse of an element. In particular, the inverse of an element a in the group G is denoted by a^{-1}.

It may be tempting to denote the inverse of a by $1/a$, but this is ambiguous in nonabelian groups G: if $a, b \in G$, it is not clear whether b/a refers to $a^{-1}b$ or ba^{-1}.

Examples of groups abound.

Example 3.7. The systems $(\mathbb{Z}, +)$, $(\mathbb{Q}, +)$, $(\mathbb{R}, +)$, and $(\mathbb{C}, +)$ are all additive groups. The identity is 0 in each and inverses in the groups coincide with the ordinary additive inverses.

Example 3.8. The general linear group $GL_n(\mathbb{R})$ is in fact a multiplicative group, as is the special linear group $SL_n(\mathbb{R})$. Inverses exist by the definitions of these sets, the identity is the familiar identity matrix, and matrix multiplication is associative since it corresponds to composition of linear maps (and composition of functions is always associative).

Example 3.9. In the previous example, we noted that $SL_n(\mathbb{R})$ is a multiplicative group. In fact, $SL_n(\mathbb{Z})$, the set of $n \times n$ *integer* matrices with determinant 1, is also a multiplicative group, but here we must verify that the inverse of an integer matrix with determinant 1 is again an integer matrix with determinant 1. This verification is handled by Cramer's rule, which asserts that the inverse of a matrix can be found by dividing each entry of its adjugate (sometimes referred to as the "classical adjoint") by its determinant. The adjugate of an integer matrix is again an integer matrix, and since the determinant is 1 the division by the determinant has no effect. It remains to check that the inverse has determinant 1, but this follows directly from the formulas $\det(A) = 1$ and $\det(A)\det(A^{-1}) = 1$.

For example, we can see this explicitly in the 2×2 case using the familiar formula

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \frac{1}{ad - bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}.$$

Since $ad - bc = 1$ for any matrix in $SL_2(\mathbb{Z})$, this formula shows that the inverse of an element in $SL_2(\mathbb{Z})$ is again an element of $SL_2(\mathbb{Z})$.

Non-Example 3.10. The system $(\mathbb{Z}, \cdot)$ is not a group since $\cdot$ is not a binary operation on $\mathbb{Z}$, as remarked earlier.

Example 3.11. The subset $\{\pm 1\} \subset \mathbb{Z}$ is a multiplicative group with identity 1. The inverse of 1 is 1 and the inverse of -1 is -1. This is our first example of a *finite* group.

The **order** of a group is its cardinality. In the case of a finite group, the order is just the number of elements in the group. For instance, $\{-1, 1\}$ has order two.

Example 3.12. The nonzero rational numbers $\mathbb{Q} - \{0\}$ form a multiplicative group, as do the nonzero real numbers and the nonzero complex numbers. It is necessary to eliminate zero so that every element is invertible.

Example 3.13. If S is a set, then a **permutation** of S is a bijection from S to itself. Denote the set of all permutations of the set S by $\mathrm{Sym}(S)$. If S is nonempty, then $\mathrm{Sym}(S)$ is a group with respect to function composition. The identity of the group is the identity mapping, and the inverse of a permutation is its inverse as a function (this inverse will exist since all permutations are bijective). Moreover, the group operation is associative because function composition is *always* associative.

Example 3.14. Building on the previous example, if $S = \{1, 2, \ldots, n\}$ for some positive integer n, then we write S_n in place of $\mathrm{Sym}(S)$ and refer to S_n as the **symmetric group on n symbols**. Evidently S_n has order $n!$. To see this, consider the construction of an arbitrary permutation of $\{1, 2, \ldots, n\}$. We choose first from n options the image of 1 under the permutation. This choice being made, we choose next from $n - 1$ options the image of 2. This procedure is repeated until the images of all n elements of $\{1, 2, \ldots, n\}$ have been decided, and there are $n(n-1)(n-2)\ldots 2 \times 1 = n!$ such permutations that can thereby be constructed.

Definition 3.15. A system $(G, *)$ consisting of a set S together with an associative binary operation $*$ on S is called a **semigroup**. This terminology is occasionally useful and will appear, for instance, in the discussion of free groups later on.

Example 3.16. Put $G = \{0, 1\}$ and define $a*b = 1$ if $a = b = 1$ and $a*b = 0$ otherwise. Then $(G, *)$ is a semigroup but not a group. In fact, it is abelian and has the identity 1, but 0 fails to have an inverse.

3.3 Modular arithmetic

There are various accounts of modular arithmetic. Two slightly different developments can be found in Herstein [5] and Artin [4].

Recall that an integer $a \in \mathbb{Z}$ is **even** if $a = 2k$ for some $k \in \mathbb{Z}$ and **odd** if $a = 2k + 1$ for some $k \in \mathbb{Z}$. In this way $\mathbb{Z}$ is partitioned into the classes of even and odd integers.

This situation is generalized by the congruence relation. One says that integers a, b are **congruent modulo** n if n divides $a - b$ or, equivalently, if $a = b + nk$ for some $k \in \mathbb{N}$. This is represented by the expression $a \equiv b \pmod{n}$. We recover the traditional idea of parity when $n = 2$: an integer a is even if $a \equiv 0 \pmod{2}$ and odd if $a \equiv 1 \pmod{2}$. The idea that modular arithmetic generalizes parity is useful in problem solving, and an example is given at the end of this section to illustrate this mode of thought.

If $a \in \mathbb{Z}$, let $\overline{a}$ denote the set of all those integers congruent to a modulo n. Thus

$$\overline{a} = \{\dots, a - 3n, a - 2n, a - n, a, a + n, a + 2n, a + 3n, \dots\}.$$

We call $\overline{a}$ the **residue class** or **congruence class** of a modulo n, and the collection of all residue classes modulo n is denoted $\mathbb{Z}_n$ or $\mathbb{Z}/n\mathbb{Z}$. (The latter notation will be explained in the section on quotient groups.)

The statement $\overline{a} = \overline{b}$ is the same as the statement $a \equiv b \pmod{n}$. The latter has the advantage that it indicates the modulus n explicitly. The former is often convenient in computations.

It seems intuitive that there should be n residue classes modulo n in the same way that parity partitions $\mathbb{Z}$ into two classes. This is true, and it is the fundamental result that makes modular arithmetic valid. Yet its proof relies on the division algorithm, which we record for a reminder:

Lemma 3.17 (Division with remainder/division algorithm)**.** *If $a, b \in \mathbb{Z}$ with $b \neq 0$, then there exist unique integers q, r such that $a = bq + r$ and $0 \leq r < |b|$.*

In this lemma, r is the remainder when a is divided by b.

Proposition 3.18. *$\mathbb{Z}_n$ has exactly n elements. In other words, congruence modulo n partitions $\mathbb{Z}$ into n distinct residue classes.*

Proof. If $a \in \mathbb{Z}$, then there exist unique $q, r \in \mathbb{Z}$ such that $a = qn + r$ and $0 \leq r < n$, so $\overline{a} = \overline{r}$; hence any congruence class is the congruence class of a nonnegative integer less than n. To show that these classes are distinct, suppose $0 \leq a < b < n$; then $0 < b - a < n$, so $n \nmid b - a$ and $\overline{a} \neq \overline{b}$. □

The proof shows that the n elements of $\mathbb{Z}_n$ are the residue classes $\overline{0}, \overline{1}, \ldots, \overline{n-1}$.

We would like to define the sum $\overline{a} + \overline{b}$ of two residue classes to be the residue class $\overline{a+b}$ of the sum of the representatives, but it is not immediately clear that this works since the same residue class will have multiple representatives. To verify that this doesn't matter, if $a \equiv a' \pmod{n}$ and $b \equiv b' \pmod{n}$ then we must show that $\overline{a+b} = \overline{a'+b'}$. There must exist integers k, k' such that $a' = a + nk$ and $b' = b + nk'$, so

$$a' + b' = a + nk + b + nk' = (a+b) + n(k+k').$$

Thus $a' + b'$ differs from $a + b$ by a multiple of n, so that $\overline{a'+b'} = \overline{a+b}$, as desired.

Similarly, we would like to define $\overline{a}\,\overline{b} = \overline{ab}$, but we must check that this is independent of the choice of representatives. If $a \equiv a' \pmod{n}$ and $b \equiv b' \pmod{n}$ then we must show that $\overline{ab} = \overline{a'b}$. As before, $a' = a + nk$ and $b' = b + nk$ for some $k, k' \in \mathbb{Z}$, so

$$a'b' = (a+nk)(b+nk') = ab + n(ak' + kb + nkk'),$$

which is congruent to ab modulo n, as desired.

Thus it is possible to perform arithmetic with the residue classes modulo n. In fact, it's easy to check that $\mathbb{Z}_n$ is an additive group with identity $\overline{0}$; notice the inverse of $\overline{a}$ is simply $\overline{-a}$. Because of the definitions of addition and multiplication of residue classes, it is possible to either work with integers in $\mathbb{Z}$ and then reduce the result modulo n or to reduce all quantities modulo n from the beginning and work formally within $\mathbb{Z}_n$.

It is possible to identify a subset of $\mathbb{Z}_n$ that is a multiplicative group as the following proposition shows. The proof provides an example of working in $\mathbb{Z}$ first and then reducing modulo n. But we will need first a simple lemma.

Lemma 3.19. *If* $a, b, c \in \mathbb{Z}$, $\gcd(a,b) = 1$, *and* $\gcd(a,c) = 1$, *then* $\gcd(a, bc) = 1$.

Proof. There exist $m, n \in \mathbb{Z}$ such that $ma + nb = 1$, so $mac + nbc = c$. If $d = \gcd(a, bc)$, then $d \mid (mac + nbc) = c$, so $d \mid a$ and $d \mid c$. Since $\gcd(a, c) = 1$, it follows that $\gcd(a, bc) = 1$. □

Proposition 3.20. *The set* $\mathbb{Z}_n^* := \{\overline{a} : \gcd(a, n) = 1\} \subseteq \mathbb{Z}_n$ *is a multiplicative group.*

Proof. $\mathbb{Z}_n^*$ is closed by the lemma. If $\gcd(a, n) = 1$, then $aq + nr = 1$ for some $q, r \in \mathbb{Z}$, so $aq \equiv 1 \pmod{n}$, i.e., q is the inverse of a modulo n. More formally, $(\overline{a})^{-1} = \overline{q}$ within $\mathbb{Z}_n^*$. □

Example 3.21. If p is prime, then $\mathbb{Z}_n^* = \{1, 2, \ldots, p-1\}$ is a multiplicative group of order $p - 1$.

The following appeared as problem 7 of part I of the 1954 Putnam examination and should illustrate the use of modular arithmetic in problem solving. (The problem statement was retrieved from [2].)

Problem 3.22. Prove that there are no integers x and y for which $x^2 + 3xy - 2y^2 = 122$.

Proof. Solving for x, we obtain by use of the quadratic formula

$$x = \frac{-3y \pm \sqrt{17y^2 + 488}}{2}.$$

If a number is a perfect square, then it is in particular a perfect square modulo 17. But the squares modulo 17 are $1, 2, 4, 8, 9, 13, 15$, and 16. (This can be found by squaring the integers from 1 to 16 and reducing the results modulo 17.) Since $17y^2 + 488 \equiv 12 \pmod{17}$, it follows that the discriminant is never a perfect square and no integer solutions exist. □

3.4 Subgroups

It may happen that a subset of a group is itself a group with respect to the same operation. This situation is important enough that it is singled out by the following definition:

Definition 3.23. If $(G, *)$ is a group and $H \subseteq G$ is a subset of G, then we say that H is a **subgroup** of G if $(H, *|_{H \times H})$ is a group, where $*|_{H \times H}$ is the restriction of the binary operation $*$ to the subset $H \times H$ of $G \times G$. This situation will be designated by writing $H < G$; the symbol $<$ in particular will not determine whether H is a proper subset of G or not.

It is best to clear up some preliminary doubts from the outset:

Lemma 3.24. *Suppose $H < G$. Then*

(A) *the groups H and G share the same identity and*

(B) *if $a \in H$, then the inverse of a with respect to H and the inverse of a with respect to G coincide.*

Proof.

(A) If $e \in G$ is the identity in G, then e will by its definition also act as an identity for H. But H has by Lemma 3.6 exactly one element that acts as inverse, so e must be it.

(B) If $a \in H$ and b is the inverse of a as an element of G, then by part (A) b will also act as an inverse of a as an element of H. Apply once more Lemma 3.6.

□

To check that a subset $H \subseteq G$ of a group G is a subgroup, it suffices by the definition to check that H is closed under the group operation and the operation of taking inverses. It is not necessary to check associativity, which is inherited from the group structure of G. With any initial fears subdued, examples follow.

Example 3.25. If $\mathbb{Z}$ is the group of integers under addition, then the even integers form a subgroup of $\mathbb{Z}$. This is because the additive inverse of an even integer is even and the sum of two even integers is even. The odd integers do **not** form a subgroup of $\mathbb{Z}$. In particular, there is no identity element, for 0 is even.

Example 3.26. More generally, the integer multiples of some fixed positive integer n form a subgroup of $(\mathbb{Z}, +)$. This is because $0 + 0 \equiv 0 \pmod{n}$.

Example 3.27. The unit circle in the complex plane is a subgroup of the multiplicative group of nonzero complex numbers. If z lies on the unit circle, we can write $z = e^{i\theta}$ for some real θ. This representation shows at once that $|z^{-1}| = |e^{-i\theta}| = 1$, i.e., the inverse of a unit complex number is a unit complex number. Similarly, $|e^{i\theta_1} e^{i\theta_2}| = |e^{i(\theta_1+\theta_2)}| = 1$ for real θ_1, θ_2, so we find that the product of two unit complex numbers is again a unit complex number.

The next section will explore in detail a useful and significant example, but before proceeding it is worth noting the following technique for showing that a subset of a group is a subgroup.

Proposition 3.28. *A nonempty subset $H \subseteq G$ is a subgroup if and only if $ab^{-1} \in H$ for all $a, b \in H$.*

Proof. If H is a subgroup, then it follows at once that $ab^{-1} \in H$ for all $a, b \in H$ by the definition of a group.

If conversely the stated condition on H holds, then there is some element $a \in H$ since H is nonempty, so $e = aa^{-1} \in H$ by hypothesis. Hence, if $b \in H$, then $b^{-1} = eb^{-1} \in H$. Next, if $b, c \in H$, then we have just found that $c^{-1} \in H$, so $bc = b(c^{-1})^{-1} \in H$. □

Example 3.29. As an illustration of the technique, we will show that if G is an abelian group and $H = \{a \in G : a^2 = e\}$, then H is a subgroup of G. In fact, $e \in H$ since $e^2 = e$, so H is nonempty. If $a, b \in H$, then $(ab^{-1})^2 = a^2b^{-2} = a^2(b^2)^{-1} = e^2 = e$, so $ab^{-1} \in H$ and $H < G$ by the proposition.

3.5 Motions of the plane

It is worthwhile developing in detail this particular example, which is accessible to geometric intuition. It will provide another illustration of the utility of group-theoretic language and will also serve as a useful stage for discussing normal subgroups later.

An **isometry** or **motion** of the plane $\mathbb{R}^2$ is a distance-preserving map of the plane to itself. In sharper formulation, the function $f : \mathbb{R}^2 \to \mathbb{R}^2$ is an isometry if $|f(a) - f(b)| = |a - b|$ for all $a, b \in \mathbb{R}^2$. The isometries of the plane form a group under function composition which we will denote simply by M.

Consider now those motions of the plane that preserve the origin. Such motions also form a group, the **orthogonal group** $O(2)$, so we find that $O(2)$ is a subgroup of M. The following lemma explains the name.

Lemma 3.30. *The elements of $O(2)$ are exactly the orthogonal linear operators on $\mathbb{R}^2$.*

Proof. If T is an orthogonal linear operator on $\mathbb{R}^2$, then T preserves the dot product and hence distances. Any linear operator will preserve the origin, so $T \in O(2)$.

Conversely, if $f \in O(2)$ and $a, b \in \mathbb{R}^2$, then

$$|f(a) - f(b)|^2 = \langle f(a) - f(b), f(a) - f(b)\rangle = |f(a)|^2 - 2\langle f(a), f(b)\rangle + |f(b)|^2.$$

Since f preserves the origin and distances, $|f(a)|^2 = |f(a) - f(0)|^2 = |a|^2$ and similarly $|f(b)|^2 = |b|^2$, so we find

$$|f(a)|^2 - 2\langle f(a), f(b)\rangle + |f(b)|^2 = |a|^2 - 2\langle f(a), f(b)\rangle + |b|^2. \tag{2}$$

A similar computation yields $|a-b|^2 = |a|^2 - 2\langle a, b\rangle + |b|^2$, and this expression must coincide with (2) since f is an isometry. We have thus $\langle f(a), f(b)\rangle = \langle a, b\rangle$, so f preserves dot products.

It's now sufficient to show that f is linear. Using the preservation of the dot product we find

$$\begin{aligned}
|f(a+b) - f(a) - f(b)|^2 &= \langle f(a+b) - f(a) - f(b), f(a+b) - f(a) - f(b)\rangle \\
&= |f(a+b)|^2 - 2\langle f(a+b), f(a)\rangle - 2\langle f(a+b), f(b)\rangle \\
&\quad + 2\langle a, b\rangle + |a|^2 + |b|^2 \\
&= |a+b|^2 - 2\langle a+b, a\rangle - 2\langle a+b, b\rangle + 2\langle a, b\rangle + |a|^2 + |b|^2 \\
&= -\langle a+b, a+b\rangle + 2\langle a, b\rangle + |a|^2 + |b|^2 \\
&= -|a|^2 - 2\langle a, b\rangle - |b|^2 + 2\langle a, b\rangle + |a|^2 + |b|^2 \\
&= 0.
\end{aligned}$$

Finally, if $c \in \mathbb{R}$, then $|f(ca)| = |ca| = c|a| = c|f(a)|$ by the preservation of distance and the origin. We conclude that f is an orthogonal linear operator. □

The orthogonal operators provide a convenient representation of an arbitrary element of M.

Lemma 3.31. *If a basis for $\mathbb{R}^2$ is agreed upon, every element $f \in M$ has the form $f(x) = Ax + b$ where A is a 2×2 orthogonal real matrix and b is some vector in $\mathbb{R}^2$.*

Proof. Put $b = f(0)$ and let t_{-b} denote the translation by $-b$. Then $t_{-b} \circ f$ is an isometry preserving the origin, so it is by the preceding result an orthogonal linear operator $T : \mathbb{R}^2 \to \mathbb{R}^2$ with (orthogonal) matrix A with respect to the specified basis. Hence $f(x) = t_b \circ T(x) = Ax + b$. □

Recall that the matrix A will have always determinant ± 1 since it is orthogonal. If this determinant is $+1$, then f is called **orientation-preserving**. If this determinant is -1, then f is called **orientation-reversing**. A classification of all motions of the plane, which we will not carry out in full detail, concludes with the following result:

Theorem 3.32. *Any element $f \in M$ is exactly one of the following:*

- *Orientation-preserving motions: rotations about some point and translations by some fixed vector.*
- *Orientation-reversing motions: reflections in some line and glide reflections, which are compositions of translations and reflections.*

A direct proof can be found in Artin [4]. A slightly less direct approach uses the **Cartan-Dieudonné theorem**, which asserts in greater generality that every origin-preserving isometry of $\mathbb{R}^n$ is the composition of at most n reflections in hyperplanes containing the origin. An elegant proof of this result by induction on the dimension n can be found in [3].

The Cartan-Dieudonné theorem shows that the orientation-preserving orthogonal operators on $\mathbb{R}^2$ are the compositions of two reflections in lines intersecting the origin. But it is known from geometry that such compositions are rotations. So if we show that a rotation about the origin followed by a translation is a rotation, then the classification of orientation-preserving isometries in Theorem 3.32 will follow by use of Lemma 3.31.

Lemma 3.33. *If r_θ is the rotation of the plane about the origin through angle θ and t_a is the translation of the plane by the vector a, then the composition $t_a \circ r_\theta$ is a rotation (not necessarily about the origin).*

Proof. Working with complex coordinates, put $b = a/(1-e^{i\theta})$. If z is a point in the plane, then

$$\begin{aligned} t_b \circ r_\theta \circ t_{-b}(z) &= e^{i\theta} z - \frac{e^{i\theta} a}{1 - e^{i\theta}} + \frac{a}{1 - e^{i\theta}} \\ &= e^{i\theta} z + a \\ &= t_a \circ r_\theta(z), \end{aligned}$$

so $t_a \circ r_\theta$ is rotation about $a/(1 - e^{i\theta})$ through angle θ. □

We will omit similar calculations needed to verify the remainder of Theorem 3.30.

The group M has many interesting subgroups, and we have found already a significant one–the orientation-preserving orthogonal operators $O(2)$. We will examine more later on in connection with normal subgroups.

3.6 Cosets

We will begin now the identification of a certain class of subgroups of significant importance in group theory. These are the **normal subgroups**, and to isolate their characteristic properties we study first the somewhat peculiar constructions called **cosets**.

Definition 3.34. Suppose G is a group and $H < G$ is a subgroup of G. Fix an element $g \in G$ of G. The **left coset of H containing** g is the set

$$gH := \{gh : h \in H\}.$$

The **right coset of H containing** g is the set

$$Hg := \{hg : h \in H\}.$$

The importance of cosets will be gradually uncovered in the following development of their properties. We first note the following fundamental phenomenon. The constructions of the proof are as important as the conclusion of the proposition.

Proposition 3.35. *Given a group G and a subgroup $H < G$ of G, all left (right) cosets of H form a partition of G, i.e., the left (right) cosets of H are disjoint and their union is all of G.*

Proof. We prove the proposition for right cosets. Identify two elements of the group $a, b \in G$ if and only if $ab^{-1} \in H$. This identification is an equivalence relation, as we now verify.

Since H is a subgroup, $aa^{-1} = e \in H$, so the relation is reflexive. If $ab^{-1} \in H$, then there is an $h \in H$ such that $ab^{-1} = h$. Taking inverses gives $ba^{-1} = h^{-1} \in H$ since H is a subgroup, so the relation is symmetric. Finally, suppose $a, b, c \in G$, $ab^{-1} \in H$, and $bc^{-1} \in H$. Then there exist $h_1, h_2 \in H$ such that $ab^{-1} = h_1$ and $bc^{-1} = h_2$. On multiplying we find $(ab^{-1})(bc^{-1}) = ac^{-1} = h_1h_2 \in H$, once again because H is a subgroup. Hence the relation is transitive.

Now if $ab^{-1} \in H$, then $ab^{-1} = h$ for some $h \in H$, so $b^{-1} = a^{-1}h$; taking inverses gives $b = h^{-1}a$. But $h^{-1} \in H$ since H is a subgroup, so $b \in Ha$. Conversely, suppose $b \in Ha$, so $b = ha$ for some $h \in H$; then $ba^{-1} = h$. By the symmetry of this relation already established, we find $ab^{-1} \in H$. In other words, the equivalence class of a is the right coset Ha.

The equivalence classes automatically partition the group G in the desired manner, so the proof is complete. □

The following is a striking illustration of this result.

Example 3.36. The three-sphere S^3 can be realized as the multiplicative group of unit quaternions spanned by the basis $\{1, i, j, k\}$. Consider the circle $\{\cos\theta + i\sin\theta : \theta \in \mathbb{R}\} < S^3$. If r is any unit quaternion, then the coset rH is another circle congruent to H (since, in fact, multiplication by a unit quaternion is an isometry of the space of quaternions). But such cosets partition S^3 by the preceding result, so we find that S^3 can be partitioned into disjoint and congruent circles. This is the famous **Hopf fibration**.

Example 3.37. Modular arithmetic can be understood in the language of cosets. Consider the additive group of integers $\mathbb{Z}$ and the subgroup H consisting of all multiples of some fixed integer n. Two integers a, b are congruent modulo n if their difference is a multiple of n, i.e., if $a - b \in H$. This is precisely the equivalence relation used in the proof of the preceding proposition, so from the proof of that proposition we find that the residue class of an integer a is the coset $a + H = \{\ldots, a - 2n, a - n, a, a + n, a + 2n, \ldots\}$, as before.

3.7 Normal subgroups

Conjugation is an operation that can be performed on the elements of a group, and it plays a special role in the development of normal subgroups.

Definition 3.38. If G is a group and $g \in G$ is an element of G, the function $a \mapsto gag^{-1}$ is **conjugation by** g; it is also called the **inner automorphism induced by** g. The image gag^{-1} of a is the **conjugate of** a **by** g. Denote gag^{-1} by a^g. (This is unrelated to exponentiation.) If $H \subseteq G$ is a subset of G, write $H^g = \{a^g = gag^{-1} : a \in H\}$.

We are now in a position to define normal subgroups and to discuss their many properties.

Definition 3.39. A subgroup $H < G$ of a group G is a **normal subgroup** if it is fixed by all inner automorphisms. In symbols, this means that $H^g = H$ for all $g \in G$. If H is normal in G, we write $H \triangleleft G$. Note that this does not require that H be fixed element-wise by conjugation.

We remark without proof that it is enough to simply require $H^g \subseteq H$. This is useful in proving that a subgroup is normal, for then one only needs to show it is closed under conjugation by elements of G.

Example 3.40. If G is abelian, it's immediate that any subgroup $H < G$ is in fact normal. In fact, the conjugation maps all reduce to the identity in this case.

Example 3.41. We pause now to continue developing the example of the group M of motions of the plane $\mathbb{R}^2$. In particular, let M^+ be the group of orientation-preserving motions and let $T < M^+$ be the subgroup of M^+ consisting only of translations. T is a normal subgroup of M^+, as we now verify.

If $g \in M^+$ is an arbitrary orientation-preserving motion of the plane and $t_a \in T$ is the translation of the plane by the vector a, we must show that t_a^g is again a translation. To do this, we appeal to Lemma 9.2 so that we may write $g = t_b \circ r_\theta$ where r_θ is rotation of the plane about the origin through the angle θ and t_b is translation by the vector b. If $c = r_\theta(a)$ we find now

$$t_a^g = g \circ t_a \circ g^{-1} = t_b \circ r_\theta \circ t_a \circ r_{-\theta} \circ t_{-b} = t_b \circ t_c \circ r_\theta \circ r_{-\theta} \circ t_{-b} = t_c \in T.$$

Hence $T \triangleleft M^+$.

Consider now the group $H < T$ of *integer* translations, i.e., the group consisting of those translations of the plane by a vector whose components are integers. T is an abelian group, so it follows by the previous example that H is normal in T.

The striking observation is that H is not normal in M^+ despite the facts that H is normal in T and T is normal in M^+. To see this, suppose $t \in T$ is translation by the vector $(1, 0)$ and r is rotation of the plane about the origin through $\pi/4$ radians. The conjugate rtr^{-1} is not an integer translation.

We now expose the relationship between normal subgroups and cosets.

Proposition 3.42. *If $N \triangleleft G$ and $a, b \in G$, then $aNbN = \{an_1bn_2 : n_1, n_2 \in H\} = abN$.*

Proof. Since N is normal, $gNg^{-1} = N$ for $g \in G$. Multiplying on the right by g yields $gN = Ng$. Therefore,

$$aNbN = a(Nb)N = a(bN)N = abNN = abN$$

since $NN = N$. □

This property allows arithmetic with the cosets of a normal subgroup to be performed in the most natural way, and this is the basis of the next section.

3.8 Characteristic subgroups

It is instructive to consider a more restrictive condition that can be imposed on subgroups. In defining a normal subgroup we required that the subgroup be fixed by all inner automorphisms. By replacing "inner automorphism" by "automorphism" we obtain a special case:

Definition 3.43. A subgroup $H < G$ of a group G is a **characteristic subgroup** of G if it is fixed by all automorphisms of G.

It is immediate from the definition that a characteristic subgroup is a normal subgroup. We saw in the previous section that if A is normal in B and B is normal in C, then A is not necessarily normal in C. This is rectified by requiring that in addition A be characteristic in B.

Proposition 3.44. *If A is a characteristic subgroup of B and B is a normal subgroup of C, then A is normal in C.*

Proof. Suppose ϕ is an inner automorphism of G. Then ϕ fixes B, so its restriction to B is an automorphism (not necessarily inner) of B. But A is then fixed by this restriction and hence is fixed by ϕ. □

3.9 Homomorphisms

In this section we will be concerned with structure-preserving maps between groups.

Definition 3.45. A function $\phi : G \to H$ between groups G and H is a **homomorphism** if $\phi(ab) = \phi(a)\phi(b)$ for all $a, b \in G$. A bijective homomorphism is an **isomorphism**. If there is an isomorphism from G to H, then we say that G and H are **isomorphic** and we write $G \cong H$.

Isomorphic groups are essentially the same; one is obtained from the other by relabeling its elements.

Example 3.46. If G and H are groups, then the function $\phi : G \to H$ defined by $\phi(g) = e_H$, where e_H is the identity in H, is a homomorphism. This is sometimes called the **trivial homomorphism**.

Example 3.47. The determinant is a homomorphism from the general linear group to the multiplicative group of nonzero real numbers.

Definition 3.48. If $\phi : G \to H$ is a homomorphism of groups and e_H is the identity in H, then the preimage $\phi^{-1}(e_H) \subseteq G$ is called the **kernel** of the homomorphism ϕ and is denoted $\ker \phi$.

We now prove a series of crucial but simple properties of homomorphisms. In the following results, let G, H be groups and suppose ϕ is a homomorphism from G to H. The proofs are modeled after those given by Isaacs in [9].

Lemma 3.49. *If e_G and e_H are the identities in G and H, respectively, then ϕ preserves identities: $\phi(e_G) = e_H$. Further, ϕ preserves inverses: if $a \in G$, then $\phi(a^{-1}) = (\phi(a))^{-1}$.*

Proof. By the definition of homomorphism, $\phi(e_G) = \phi(e_G e_G) = \phi(e_G)\phi(e_G)$, so cancellation gives $e_H = \phi(e_G)$, proving the first assertion.

To see the second assertion, write $e_H = \phi(e_G) = \phi(aa^{-1}) = \phi(a)\phi(a^{-1})$, again using the definition of homomorphism. This simply means that $\phi(a^{-1}) = (\phi(a))^{-1}$, as desired. □

Lemma 3.50. *The kernel $\ker \phi$ is a normal subgroup of G.*

Proof. First we show that the kernel is even a subgroup. It's nonempty since $e_G \in \ker \phi$ (using the previous lemma). It is closed under taking inverses since, if $a \in \ker \phi$, then the previous lemma implies $\phi(a^{-1}) = (\phi(a))^{-1} = e_H^{-1} = e_H$. And it is closed under multiplication in G since, if $a, b \in \ker \phi$, then $\phi(ab) = \phi(a)\phi(b) = e_H e_H = e_H$.

Now we show that the kernel is a *normal* subgroup. Suppose $n \in \ker \phi$ and $a \in G$. Then $\phi(a^{-1}na) = \phi(a^{-1})\phi(n)\phi(a) = \phi(a)^{-1}e_H\phi(a) = e_H$.

□

Lemma 3.51. *The homomorphism ϕ is injective if and only if its kernel is trivial, i.e.,* $\ker\{\phi\} = \{e_G\}$.

Proof. First suppose ϕ is injective. If $a \in \ker \phi$, then $\phi(a) = e_H = \phi(e_G)$. Since ϕ is injective, $a = e_G$ and $\ker \phi = \{e_G\}$.

Conversely, suppose $\ker \phi = \{e_G\}$. If $\phi(a) = \phi(b)$, then $e_H = \phi(a)(\phi(b))^{-1} = \phi(ab^{-1})$ by previous results. Therefore, $ab^{-1} \in \ker \phi$, which means $ab^{-1} = e_G$, i.e., $a = b$. Hence ϕ is injective. □

3.10 Quotient groups

If $N \triangleleft G$, denote the set of cosets of N by G/N. Proposition 3.40 can be used to endow the cosets of N with a binary operation. In fact, this operation makes G/N of N into a group.

Proposition 3.52. *Suppose $N \triangleleft G$. Then the set G/N of cosets of N is a group with the binary operation defined by $(aN)(bN) = (ab)N$ for $a, b \in G$.*

Proof. That this is a binary operation on G/N is the content of Proposition 3.40. The identity is $eN = N$ where e is the identity of G. The inverse of aN is $a^{-1}N$. Associativity follows from the associativity of the group operation on G. □

If $N \triangleleft G$, then the group G/N endowed with the above operation is called a **quotient group**; the notation G/N is read aloud as "G mod N." One can imagine that G/N is obtained from G by collapsing the structure of G so that two elements $a, b \in G$ become identified if $ab^{-1} \in N$. The following examples illustrate the concept as well as this mode of thought.

Example 3.53. We can now provide a swift description of modular arithmetic. The additive group of integers modulo n is the quotient group $\mathbb{Z}/n\mathbb{Z}$; this notation is explained at last. It's important to note, however, that we have so far considered in the quotient construction only one binary operation, so the quotient group $\mathbb{Z}/n\mathbb{Z}$ does not precisely coincide with the system of modular arithmetic discussed previously. This is remedied by viewing $\mathbb{Z}$ as a **ring** instead of a group and invoking the machinery of quotient rings.

Example 3.54. Suppose $G = \mathbb{C}$ is the multiplicative group of nonzero complex numbers and N is the multiplicative group of unit complex numbers (the unit circle in the complex plane). Two complex numbers are to be identified if their quotient has unit magnitude, and this occurs only if the two numbers have the same magnitude. Thus the elements of G/N can be imagined as the concentric origin-centered circles in the complex plane with the product of two such circles being the circle whose radius is the product of the radii of the two factor circles. Evidently this is the same as the multiplicative group of positive real numbers (by identifying such a circle with its radius, a positive real number).

Definition 3.55. Suppose N is a normal subgroup of the group G. Consider the function $\pi : G \to G/N$ defined by $\pi(g) = Ng$. Since $NaNb = Nab$ for

all $a, b \in G$, this function is a surjective homomorphism. It is called the **canonical homomorphism** from G onto G/N.

Example 3.56. The canonical homomorphism $\mathbb{Z} \to \mathbb{Z}/n\mathbb{Z}$ provides a homomorphism from the additive group of integers to the group of residues modulo n by sending each integer to its residue class.

We now mention but do not prove a crucial theorem that somewhat solves the problem of identifying quotient groups. It is sometimes called the **first isomorphism theorem**.

Theorem 3.57. *([9], Theorem 3.3) Let $\phi : G \to H$ be a surjective homomorphism from a group G onto a group H. Then there is a unique isomorphism $\theta : G/\ker\phi \to H$ such that $\theta \circ \pi = \phi$, where π is the canonical homomorphism $G \to G/\ker\phi$. In particular, $H \cong G/\ker\phi$.*

This theorem can be applied to the previous example to avoid the use of intuitive reasoning. We will use it again in formulating the mapping properties of free groups.

3.11 Free groups

The presentation of the free group given here is inspired by that given by Artin in [4]. Free groups provide the important theoretical underpinnings for the Wirtinger presentation, introduced in the section on knots.

Given an alphabet $\mathcal{A} = \{a, b, c, \dots\}$, we consider the set W of all *words* over $\mathcal{A}$–finite strings of letters from the alphabet $\mathcal{A}$ with repetition allowed. We introduce a product or binary operation on W via concatenation or juxtaposition. For instance, the concatenation of b and ab is bab. The set W endowed with this product is a semigroup, sometimes called the free semigroup on the alphabet $\mathcal{A}$.

For each symbol $a \in \mathcal{A}$ we introduce the "inverse" symbol a^{-1}. Let $\mathcal{A}'$ denote the collection $\{a, a^{-1}, b, b^{-1}, \dots\}$ of letters from $\mathcal{A}$ together with their inverses. Let W' be the set of words over the alphabet $\mathcal{A}'$.

Introduce the following cancellation rule in W'. If any string of the form aa^{-1} or $a^{-1}a$, where $a \in \mathcal{A}$, occurs as a substring of a word $w \in W'$, then w is to be identified with the word obtained from w by deleting the substring aa^{-1} or $a^{-1}a$. Thus $bacc^{-1}b$ is to be identified with bab. A word is called **reduced** if no such cancellations are possible. The following lemma, reproduced here

together with its proof from Artin [4], shows that when multiple cancellations are possible the same reduced word is always obtained if all cancellations are performed, no matter the order.

Lemma 3.58 (Artin, Pg. 218.)**.** *Any word $w \in W'$ has precisely one reduced form.*

Proof. We proceed by induction on the length of the word. Suppose the unreduced word w has the form $\ldots aa^{-1}\ldots$. It is enough to show that all reduced forms of w can be obtained by deleting aa^{-1} first, for the result then follows by induction on the smaller word that results from this cancellation.

Suppose ω is a reduced form of w. ω is obtained from w by a sequence of cancellations. If aa^{-1} occurs in this sequence of cancellations, then the sequence of cancellations can be rearranged so that aa^1 is cancelled first. This case is hence settled.

So suppose aa^{-1} does not occur in the sequence of cancellations yielding ω from w. The pair aa^{-1} cannot be present in ω since it is reduced, so w must contain a substring of the form $a^{-1}aa^{-1}$ or $aa^{-1}a$. Since aa^{-1} is not cancelled, in both possibilities it must be $a^{-1}a$ that is cancelled. But the result of this cancellation is precisely the same as the result of cancelling aa^{-1}. So we may replace this cancellation by the cancellation of the original pair aa^{-1}, reducing to the previously settled case. □

First, it should be verified that if $w_1 \simeq w_2$ and $v_1 \simeq v_2$, then $w_1v_1 \simeq w_2v_2$ (where the juxtaposition of two words indicates their product). Write w_1' for the reduced form of w_1 and v_1' for the reduced form of v_1, so that w_1v_1 has the form $w_1'v_1'$.

Since w_2 is equivalent to w_1 and v_2 is equivalent to v_1, w_2 can be reduced to w_1' and v_2 can be reduced to v_1'. Hence w_2v_2 has also the form $w_1'v_1'$.

Now further cancellations can potentially be performed in $w_1'v_1'$ to obtain the common reduced form of the two products w_1v_1 and w_2v_2.

Definition 3.59. The set of equivalence classes of words W' over the alphabet $\mathcal{A}'$ forms a group with respect to concatenation. (Associativity and the existence of an identity follow from the properties of W'. The inverse of the concatenation $a_1a_2\ldots a_n$ is the concatenation $a_n^{-1}a_{n-1}^{-1}\ldots a_1^{-1}$.) This is the **free group** on the alphabet $\mathcal{A}$.

3.12 Group presentations

Suppose that G is a group. A subset S of G is said to **generate** G if every element of G can be written as the product of finitely many of the elements of S along with their inverses.

Suppose $S \subseteq G$ is a generating set of G. There is then a homomorphism ϕ from the free group on S to G. This homomorphism may fail to be injective if G has any significant structure beyond that implied by the group axioms. More precisely, G is isomorphic to $F/\ker\phi$, where F is the free group on S. (Apply the first isomorphism theorem.)

Definition 3.60. The elements of $\ker\phi$ are called the **relations** satisfied by the generators of the group G.

Given the generators and the relations they satisfy, the group $F/\ker\phi$ is determined, and hence G is determined, too. But in general not all the relations are necessary. In fact, it suffices to consider any subset of $\ker\phi$ that generates the group $\ker\phi$.

Definition 3.61. A subset R of $\ker\phi$ that generates $\ker\phi$ (i.e., $\ker\phi$ is the smallest normal subgroup of G containing R) is called a set of **defining relations** for G.

A list of generators together with a set of defining relations will determine the group G. These data constitute a **presentation** of the group. If the generators are the elements $g_1, g_2, \ldots, g_n$ and the relations are the elements $r_1, r_2, \ldots, r_m$, then the presentation is written $\langle g_1, g_2, \ldots, g_n \,|\, r_1, r_2, \ldots, r_m\rangle$.

To summarize, the presentation $\langle g_1, g_2, \ldots, g_n \,|\, r_1, r_2, \ldots, r_m\rangle$ determines the group which is the quotient of the free group on the set of generators by the smallest normal subgroup containing the words $r_1, r_2, \ldots, r_m$.

4 Knots

We have built enough infrastructure to now consider our main topic: knots. While knots are inspired by those that we can make out of shoelaces, the study of these knots, starting with rigorous definitions, quickly develops into something of its own.

4.1 What is a knot?

We recall that S^3 is the 3-sphere: the locus of all points in $\mathbb{R}^4$ that are unit distance from the origin. Alternatively, S^3 is the one-point compactification of $\mathbb{R}^3$: it is the result of adjoining to $\mathbb{R}^3$ a point at infinity. This is in many ways the more useful viewpoint for our purposes.

We must be precise about what we mean by a knot. The following definition is simultaneously rigorous and intuitive.

Definition 4.1. A **knot** is a homeomorphic image of S^1 in S^3.

It is often convenient–and sometimes even crucial–to assign an orientation to a knot, which can be visualized as a flow of arrows along the knot.

The fundamental problem of knot theory is to classify knots. Before this can be done it is necessary to agree on a notion of equivalence of knots. Let's agree to use the following notion, which is standard (see [6]):

Definition 4.2. Two knots A and B, considered as subsets of S^3, are regarded as being the same knot if they are related by an **ambient isotopy**: there exists a continuous function $\phi : S^3 \times [0,1] \to S^3$ such that (i) for each $t \in [0,1]$ the function $\phi(\cdot, t)$ is a homeomorphicm from S^3 onto S^3 and (ii) $\phi(A, 1) = B$.

Notice that an ambient isotopy is a certain continuous deformation of the ambient space containing the knots, not just of the knots themselves.

As we proceed we will develop a small but useful library of knots. The first item in this library is something we may not ordinarily call a knot at all.

Example 4.3. The **unknot** is any knot equivalent up to ambient isotopy to the standard embedding of S^1 in S^3 as a geometrically round circle. The unknot is depicted in Figure 1. The **trefoil knot** is depicted in Figure 2.

4.2 Knot diagrams

Knots more complicated than the unknot are difficult to visualize in three-dimensional space. It is often much more convenient to find a two-dimensional representation that contains information about the three-dimensional structure of the knot. Such a representation is obtained by selecting a plane and

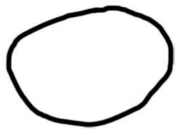

Figure 1: The unknot, a knot that is not much of a knot.

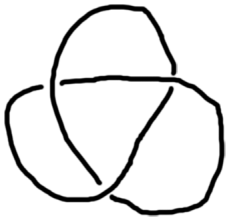

Figure 2: The (left-handed) trefoil knot.

projecting the knot onto the plane. We require that the projection is injective everywhere except at finitely many double points, which are often called crossings, and that each crossing is distinguished as either an overcrossing or undercrossing, thereby facilitating the reconstruction of the original knot from the projection. Such a projection is a **knot diagram**. For instance, Figure 2 is a knot diagram of the left-handed trefoil.

The first question to be investigated is how the diagrams of a given knot are related. The two diagrams in Figure 3 are both projections of the unknot; this illustrates that this is an important and significant question.

4.3 The Reidemeister moves

The answer was stumbled upon by German mathematician Kurt Reidemeister in his 1926 paper "Elementare Begründung der Knotentheorie." Reidemeister showed in this paper that if two knot diagrams represent the same knot, then one diagram is related to the other via finitely many applications of the three **Reidemeister moves**.

Figure 3: These two knot diagrams both depict the unknot. How are the diagrams related?

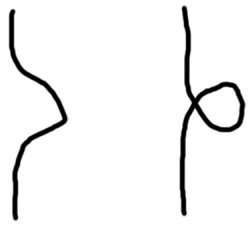

Figure 4: The first Reidemeister move allows either of the displayed strands to be transformed into the other.

Figure 5: The second Reidemeister move allows one strand to be pulled just over another or for one strand to be pulled off another.

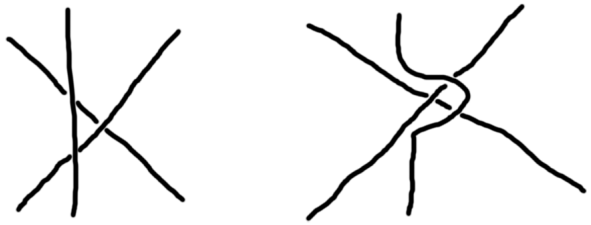

Figure 6: The third Reidemeister moves an arc to the other side of a crossing.

Definition 4.4. 1. The **first Reidemeister move** adds or removes a crossing in the form of a simple loop.

2. The **second Reidemeister move** adds or removes two crossings simultaneously.

3. The **third Reidemeister move** move slides a strand from one side of a crossing to the other.

These moves are depicted in Figures 4-6. It is important to notice that these figures are *local* zoomed-in pictures of larger knot diagrams.

For emphasis, we record Reidemeister's result as a theorem, which we will not prove:

Theorem 4.5 (Reidemeister)**.** *If two knot diagrams represent the same knot, then either of the diagrams can be transformed into the other via the application of finitely many of the three Reidemeister moves.*

4.4 Homotopy

One of the primary methods of identifying knots is to assign a construct to each knot in such a way that two knots are equivalent precisely when the associated constructs are equivalent (the notion of equivalence of the associated constructs must be decided upon as well). Such a construct is a **knot invariant**. A partial knot invariant is obtained by forming the fundamental group of the exterior of a knot. We now define these ideas in detail. The treatment given here is inspired by that in [1].

It is necessary first to provide an exposition of homotopy.

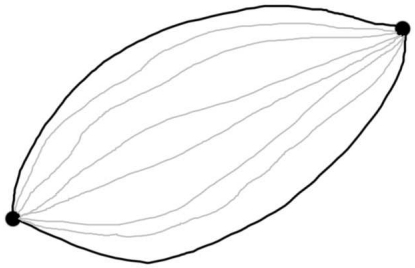

Figure 7: A homotopy with fixed endpoints between two paths.

Definition 4.6. If X is a topological space and $a, b \in X$ are points in X, then a **path from** a **to** b **in** X is a continuous function $\gamma : [0,1] \to X$ satisfying $\gamma(0) = a$, $\gamma(1) = b$.

It may be that for two given paths from a to b in the space X there exists a continuous family of intermediate paths, as in Figure 7.

This notion is made precise by the following definition:

Definition 4.7. Two paths α and β from a to b in the space X are **homotopic with fixed endpoints** if there is a continuous function $F : [0,1] \times [0,1] \to X$ such that

$$\begin{aligned} F(u,0) &= \alpha(u), && u \in [0,1], \\ F(u,1) &= \beta(u), && u \in [0,1], \\ F(0,v) &= a, && v \in [0,1], \\ F(1,v) &= b, && v \in [0,1]. \end{aligned}$$

We write $\alpha \simeq \beta \ \mathrm{rel}\{0,1\}$ and refer to F as a homotopy from α to β.

For each $v \in [0,1]$ the function $F(\cdot, v) : [0,1] \to X$ is a path in X from a to b. The first two conditions mean that $v = 0, v = 1$ correspond to the paths α, β, respectively. The final two conditions mean that the paths corresponding to all values of $v \in [0,1]$ go from a to b.

Example 4.8. Perhaps the simplest example of a homotopy is achieved by linear interpolation of paths in $\mathbb{R}^n$. If α and β are paths in $\mathbb{R}^n$ with the same fixed endpoints $a, b \in \mathbb{R}^n$, then $F(u,v) = \alpha(u) + v[\beta(u) - \alpha(u)]$ is a homotopy from α to β. Continuity of $F : [0,1] \times [0,1] \to \mathbb{R}^n$ follows from the continuity of α, β since continuity is preserved by vector addition and scalar multiplication.

Proposition 4.9. *The homotopy relation on paths with the fixed endpoints a, b in the topological space X is an equivalence relation.*

Proof. We check reflexivity, symmetry and transitivity in sequence. Fix three paths α, β, γ in X going from a to b.

Reflexivity. $\alpha \simeq \alpha$ holds because of the constant or identity homotopy $F(u, v) = \alpha(u)$.

Symmetry. If $\alpha \simeq \beta$ with the homotopy $F(u, v)$, then $\beta \simeq \alpha$ using the homotopy $F(u, 1 - v)$.

Transitivity. This case must be handled with greater care. Suppose $\alpha \simeq \beta$ with the homotopy $F(u, v)$ and $\beta \simeq \gamma$ with the homotopy $G(u, v)$. Then the function from $[0, 1] \times [0, 1] \to X$ that equals $F(\cdot, 2v)$ for $v \in [0, 1/2]$ and $G(\cdot, 2v - 1)$ for $v \in [1/2, 1]$ is a homotopy from α to γ because this function's domain can be split into the union of two disjoint closed sets with its restriction to each being continuous.

□

4.5 The fundamental group

Consider the set of paths with common starting and ending point a in the space X. We would like to introduce a group structure on this set, and to do so we introduce the following product or composition on such paths.

Definition 4.10. Given paths $\alpha, \beta : [0, 1] \to X$ starting and ending at a, the **product path** $\gamma = \alpha \cdot \beta$ traverses α and then β at twice the speed. More precisely, $\gamma(t) = \alpha(2t)$ for $t \in [0, 1/2]$ and $\gamma(t) = \beta(2t - 1)$ for $t \in [1, 2, 1]$.

Lemma 4.11. *Let α_i, β_i for $i \in \{1, 2\}$ be paths in X with starting and ending points a. This product is well-defined in the sense that if $\alpha_1 \simeq \alpha_2$ and $\beta_1 \simeq \beta_2$ using the homotopies $F, G : [0, 1]^2 \to X$, respectively, then $\alpha_1 \cdot \beta_1 \simeq \alpha_2 \cdot \beta_2$.*

Proof. The path $F(\cdot, v) \circ G(\cdot, v)$ is defined for $v \in [0, 1]$ and facilitates a homotopy from $\alpha_1 \cdot \beta_1$ to $\alpha_2 \cdot \beta_2$. □

Proposition 4.12. *The set of paths starting and ending at the fixed point a (where paths are considered equivalent up to based homotopy) in the topological space X forms a group under this product.*

Proof Sketch. Fix three such paths α, β, γ.

If $\phi : [0,1] \to [0,1]$ is any continuous function satisfying $\phi(0) = 0$ and $\phi(1) = 1$, then $\alpha \circ \phi \simeq \alpha$ via linear interpolation. In words, homotopy class is preserved by continuous reparameterization.

The product $\alpha \cdot (\beta \cdot \gamma)$ is a reparameterization of the product $(\alpha \cdot \beta) \cdot \gamma$ where the reparameterization map ϕ is piecewise linear. It is thus shown that the product is associative.

The path e that is identically equal to a serves as the identity. More precisely, $\alpha \circ e$ is a reparameterization of α by a piecewise linear map, and the same is true of $e \circ \alpha$.

Finally, we must check the existence of inverses. In fact, the inverse of $\alpha(t)$ is the path $\alpha^{-1}(t) = \alpha(1-t)$. This must be verified by showing that $\alpha \circ \alpha^{-1}$ and $\alpha^{-1} \circ \alpha$ both reduce to the identity e, defined above. We omit these verifications.

□

Definition 4.13. Given a topological space X and a fixed point $a \in X$, the group of all paths in X starting and ending at a, considered up to based homotopy, with respect to the above product is written $\pi_1(X, a)$ and called the **fundamental group** of X with basepoint a.

4.6 The knot group

We are now in the position to define a useful partial knot invariant. The construct that this invariant associates to a knot is a group.

Definition 4.14. Given a knot K in S^3, the **exterior** of K is the complement of a small open tubular neighborhood of K. The **knot group** of K is the fundamental group of the exterior of K.

Two knots equivalent up to ambient isotopy have homeomorphic exteriors and hence isomorphic knot groups. In other words, two equivalent knots have isomorphic knot groups. As a partial converse, Gorden and Luecke showed in [7] that two knots with homeomorphic complements are equivalent up to ambient isotopy, but there exist knots with isomorphic knot groups whose complements that are nonetheless not homeomorphic.

4.7 The Wirtinger presentation

We now introduce a simple algorithm to compute the fundamental group of the exterior of a knot. More precisely, this algorithm provides a presentation, called the **Wirtinger presentation**, of the knot group. We do not prove the algorithm's correctness here but instead refer the reader to [8].

Given a knot K, we begin with a diagram of K contained in the plane. We construct from this diagram a graph whose vertices are the crossings of the diagram and whose edges are the connected arcs of the diagram that terminate at the crossings. Let us label the edges $a_0, \ldots, a_{n-1}$ such that a_i is connected to a_{i+1} and a_{i-1} (with subscripts computed modulo n). The ordering of the subscripts induces an orientation on the knot (and on each arc a_i individually as well). We work with this orientation.

Using this orientation, under each arc a_i draw an arrow x_i going from "right to left." We imagine that each x_i is a *loop* in the exterior of K with some fixed basepoint a (imagined as the tip of the nose of the reader). The loop proceeds from a to the tail of x_i, thence along x_i to its head, and finally from its head back to a.

We examine each crossing of the diagram. Consider, for instance, the crossing of the three arcs a_i, a_{i+1}, and a_k (where the union of a_i and a_{i+1} is a strand of the knot passing under the transverse arc a_k). Depending on the relative orientations, one of two relations will hold: either $x_k x_i = x_{i+1} x_k$ or $x_i x_k = x_k x_{i+1}$. We let r_i designate the relation that is true, and we obtain a collection of n relations $\{r_1, \ldots, r_n\}$ satisfied by the n elements $\{x_1, \ldots, x_n\}$.

Theorem 4.15 ([8], Pg. 57)**.** *The fundamental group of the exterior of the knot K has the* ***Wirtinger presentation***

$$\langle x_1, x_2, \ldots, x_n | r_1, r_2, \ldots, r_n \rangle.$$

Example 4.16. We find that the fundamental group of the exterior of the unknot has one generator and no relations. It is hence isomorphic to $\mathbb{Z}$.

Example 4.17. We now give a nontrivial example that also illustrates that the knot group isn't a complete knot invariant. This outstanding example comes from [10]. Figures 8 and 9 show knots A and B. These knots are not equivalent, but we will now see that they have isomorphic knot groups.

Knot A has the Wirtinger presentation

$$\langle a, b, c, d, e, f | aba^{-1} = c, faf^{-1} = b, bfb^{-1} = a, cdc^{-1} = e, ece^{-1} = e, ded^{-1} = f \rangle$$
$$\cong \langle b, c, d | bcb = cbc,\ cdc = dcd \rangle.$$

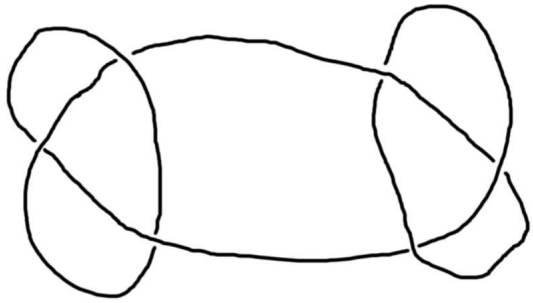

Figure 8: Knot A.

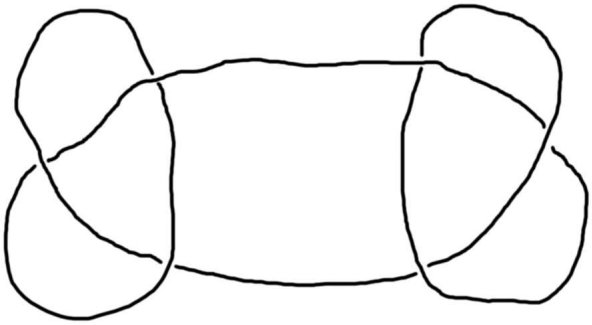

Figure 9: Knot B.

Knot B has the Wirtinger presentation

$$\langle a,b,c,d,e,f | aba^{-1}=c, faf^{-1}=b, bfb^{-1}=a, ede^{-1}=c, dfd^{-1}=e, fef^{-1}=d\rangle$$
$$\cong \langle a,e,f | afa=faf,\ efe=fef\rangle.$$

The simplified presentations are evidently describing the same group since they differ only in the names of the generators.

4.8 Linking numbers

Definition 4.18. A **link** is a finite collection of disjoint knots that may loop around one another but do not intersect.

Let K_0 and K_1 be two oriented knots in S^3. Choose a diagram of the link (containing the projections of both knots onto the same plane) and assign

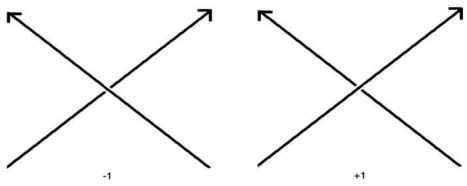

Figure 10: Left-handed and right-handed crossings.

to each crossing ± 1 according to whether the crossing is right-handed or left-handed. Examples of these crossings are shown in Figure 10.

Definition 4.19. The **linking number** of K_0 and K_1, denoted $\mathrm{lk}(K_0, K_1)$, is the sum of the signs associated to each crossing of the knot diagram.

There is an immediate issue: if the linking number depends on the choice of knot diagram, then it is not well-defined. However, it can be shown that the linking number is preserved by each of the three Reidemeister moves and is hence, in fact, independent of the choice of projection.

4.9 Branched covers

Definition 4.20. A **branched cover** is a continuous function $p : X \to M$ between compact manifolds X, M such that for a submanifold $A \subset M$, $M - A$ is exactly the collection of points that are evenly covered by p. We call A the **branch set** of p.

Example 4.21. The map

$$z \mapsto \frac{z^k}{|z|^{k-1}}$$

is a branched cover of the unit disc. The branch set consists only of the origin.

Let J be a fixed knot in S^3. Suppose K is a knot in the exterior of J that lifts in the 2-fold branched cover of S^3, branched over J, to a link of two components K_0 and K_1. We now examine the linking number $\mathrm{lk}(K_0, K_1)$.

Suppose J is the unknot, in which case the 2-fold branched over is S^3. We will say that K is in **standard position** if in the projection of $K \coprod J$ to a plane all crossings where J passes over K are adjacent and all those where J passes under K are adjacent.

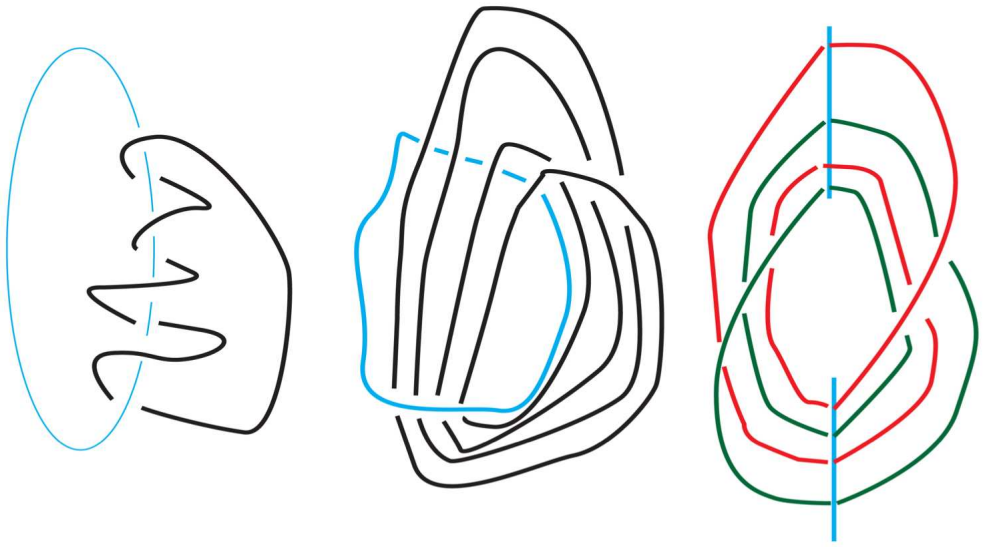

Figure 11: From left to right: a knot K (black) in S^3 linked with the unknkot (blue); K (black) in standard position linked with the unknot (blue); the two-component lift of K (green and red) to the 2-fold branched cover of S^3 branched over the unknot (blue).

Let us assume K is in standard position–any knot is equivalent to a knot in standard position. The lift of K to the 2-fold branched cover of S^3, branched over J, is the link $K_0 \coprod K_1$ of two components, and we compute $\mathrm{lk}(K_0, K_1)$ in this cover.

Let K' be the tangle obtained by cutting the strands of K as they pass through the disc bounded by the unknot J. We thereby obtain a projection of $K_0 \coprod K_1$ in the plane by gluing two copies of K' together in an orientation-preserving way.

Example 4.22. This method yields the following result when K takes on a particular simple form. Let K be a knot that links the unknot J exactly $2n$ times in the same direction without crossing iteslf. Then $\mathrm{lk}(K_0, K_1) = n$. The case $n = 4$ is depicted in the Figure 11.

References

[1] Gamelin, T. W., and Greene, R. E., "Introduction to Topology", Second Edition, Dover, 1999.

[2] Larson, L., "Problem-Solving Through Problems", Springer Science + Business Media, 2006.

[3] Stillwell, J., "Naive Lie Theory", Springer Science + Business Media, 2008.

[4] Artin, M., "Algebra", Prentice-Hall, 1991.

[5] Herstein, I. N., "Abstract Algebra", Third Edition, John Wiley & Sons, 1999.

[6] Amenta, N., Peters, T. J., and Russell, A., "Computational Topology: Ambient Isotopy Approximation of 2-Manifolds", 2001.

[7] Gordon, C. and Luecke, J., "Knots are determined by their complements", J. Amer. Math. Soc. **2** (1989) 371-415.

[8] Rolfsen, D., "Knots and Links", American Mathematical Society, 2003.

[9] Isaacs, I. M., "Algebra: A Graduate Course", Brooks Cole, 1993.

[10] Birrell, E., "The Knot Quandle", The Harvard College Mathematics Review, 2007, 44.