Clark Bray
Adrian Butscher
Simon Rubinstein-Salzedo

Algebraic Topology

Springer

Algebraic Topology

Clark Bray · Adrian Butscher ·
Simon Rubinstein-Salzedo

Algebraic Topology

Springer

Clark Bray
Department of Mathematics
Duke University
Durham, NC, USA

Adrian Butscher
Autodesk Research
Toronto, ON, Canada

Simon Rubinstein-Salzedo
Euler Circle
Palo Alto, CA, USA

ISBN 978-3-030-70607-4 ISBN 978-3-030-70608-1 (eBook)
https://doi.org/10.1007/978-3-030-70608-1

Mathematics Subject Classification: 55-01

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Foreword

Stanford University Mathematics Camp (SUMaC) was founded in the 1994–95 academic year, when Stanford mathematics professors Rafe Mazzeo and Ralph Cohen successfully secured a four-year grant from the Howard Hughes Medical Institute to fund a mathematics summer program for high school students. I joined the founding team to help design the program and teach the course in the first summer. The students were wonderful, and the overall experience was delightfully rewarding for everyone involved, inspiring me to continue with SUMaC ever since.

From the beginning, we recognized great value in showing mathematically curious and talented high school students advanced topics from the undergraduate curriculum. Although many of these students could have developed their talents through mathematics competitions, there were few opportunities for them to explore pure mathematics in a deep way. As we put it in our first program materials, our aim was to "excite and inspire students by exposing them to the beauty of mathematics."

We recognized the value of creating a friendly environment for interaction among students with shared interest in mathematics. Along with that goal, we also sought to reach students from communities traditionally under-represented in mathematics or who did not have opportunities for advanced academics generally. Over the years, SUMaC has been successful at drawing high school students at the highest level of mathematical talent, and each year SUMaC creates a community where these students can engage in mathematics with similarly talented and curious peers. These students are immersed in a social-academic environment that shapes their educational path and leads to long-lasting friendships.

In 1995, SUMaC had just a dozen participants, primarily from the San Francisco Bay Area, and all within a two-hour drive of the Stanford campus. In this first year, the program was three weeks long, and the course was a streamlined version of an introductory course in abstract algebra at the undergraduate level that also included topics from number theory and geometry. The program consisted of lectures along with problem-solving sessions that allowed participants to engage more fully in the

course material. Additionally, the program included an opportunity to explore topics of the students' choice in greater depth, and they got practice communicating mathematics by giving presentations to their peers. These features continue to be the essential ingredients of the SUMaC program.

Building on the success of the first summer, SUMaC expanded to 28 students in the second summer, and then 35 students in the third. Although the demand could easily sustain more growth, we found having 35–42 students was an optimal size for the style of program that we had developed, and it has remained in that range over the years. From the beginning, we secured a single campus residence that we could make our own, furthering our goal of establishing a social environment where the participants and residential staff would feel like family. Starting in 1997, we had participants from outside of California, and in the following year students joined from outside the US. Now the program draws an international mix of students, representing a diversity of backgrounds and experiences, who share a common passion for mathematics.

From early on, we were interested in opportunities for students to return to SUMaC for a second summer. One of the first twelve students from 1995 returned in 1996 to explore Galois theory and other topics through guided independent study. In 1997, four students from the previous year returned for a special one-week program in Real Analysis led by Rafe Mazzeo. In 1997, we launched the first version of "Program 2," a course designed to run concurrently with the original course, which then became known as "Program 1." This allowed students the potential to return for a second summer, if they had participated in Program 1 following their sophomore year in high school.

While the Program 1 maintains a focus on abstract algebra and number theory, Program 2 has varied in topic. From 1998 through 2000, Program 2 was a course in complex analysis designed and taught by Dr. Marc Sanders, who had received his Ph.D. in Mathematics at Stanford in 1994. In 2001, Dr. Clark Bray, who had been working for SUMaC while in the Ph.D. program in mathematics at Stanford, designed a course in algebraic topology that became the program 2 course from 2001 through 2004.

From 2005 through 2007, Prof. Rafe Mazzeo and his student Dr. Pierre Albin taught the Program 2 course. They kept the focus on algebraic topology while also including ideas from geometric topology, where methods from algebra and calculus have proved to be effective tools.

Starting in 2008, Adrian Butscher took over Program 2 and further developed the coursework on algebraic topology, building on the course design that had been used previously. In 2009, mathematics Ph.D. student Simon Rubinstein-Salzedo, an alumnus of SUMaC 2001, joined the instructional team of SUMaC as a TA for Adrian's algebraic topology course. Adrian continued as the SUMaC Program 2 instructor until 2013, and Simon remained one of the TAs for the course even after receiving his Ph.D. from Stanford in 2012.

In 2014, Simon took over teaching the SUMaC Program 2 course. He had been working with Adrian to refine, expand, and improve the course materials, and that collaboration continued for several years. Simon has now been teaching the

Program 2 course for six years. Given his engaging teaching style, his passion for mathematics, and his wonderful presentation of the course material, Simon has inspired his students and helped them take their mathematical talent and curiosity to a higher level. All have left the course with a deeper understanding and greater appreciation of mathematics, and many have become successful mathematicians in their own right.

Dr. Rick Sommer
Director, Stanford University Mathematics Camp
Stanford, California, USA

Introduction

This book is based on a four-week class that we have taught many times at the Stanford University Mathematics Camp (SUMaC). Students attending this camp have just finished grades 10 and 11 and are selected from among the strongest mathematics students of that age in the world. Still, we do not assume that they have seen typical material that students would be familiar with before taking an algebraic topology class, such as abstract algebra or point-set topology (or, for that matter, multivariable calculus or linear algebra). Thus we include background on these subjects as needed.

As in any mathematics book, the problems are very important. They are intended to be doable but challenging, and ideally several people will work on the problems together and share ideas. When compared with competition problems that students of this age are often familiar with, the difficulty in most of the problems in this book lies elsewhere: most of them do not require clever tricks in order to solve, but rather the challenge is in unraveling the definitions and theorems and becoming accustomed to a deeper level of abstraction.

The presentation of material in this book differs to some extent from other books on algebraic topology due to our different audience. While we aim to present the material rigorously when reasonable, there are times when we feel that the technical details of the subject are overwhelming, so we skip certain challenging steps in our arguments. This is especially true in our discussion of homology. We have chosen to work with simplicial or Δ homology, so that we can do hands-on computations. This is opposed to singular homology, where the proofs are much easier but computations are very difficult. Given our target audience, this feels like the right decision.

We also occasionally take short detours to discuss other interesting and tangentially related topics in mathematics. At least one of us feels that he would have learned many more interesting things as a student, had more authors not been so disciplined about staying on topic! Thus we have been as undisciplined as we feel we can get away with.

Each chapter of the book corresponds to one day of class at SUMaC. Each morning, the instructor presents material in the chapter in a 150-minute lecture

(with a break). In the afternoons, students work on the problems for at least 150 minutes, then possibly more in the evenings if they choose to do so. During this time, students also discuss problems from the previous chapter one-on-one with a teaching assistant. It is difficult to learn this amount of material in such a short amount of time. Some students manage to learn nearly all of it, and some students struggle more with certain topics depending on their mathematical background, geometric intuition, and other factors. But everyone who attends gets a lot out of it and learns a tremendous amount of new mathematics that they wouldn't have learned otherwise.

We believe that, at a less blistering pace, this book can also be used either for self-study or as a textbook for an introductory undergraduate topology course. For students who aren't studying this material full-time, learning a chapter or two a week is probably a more reasonable goal.

We hope you enjoy reading this book as much as we have enjoyed writing it and teaching the classes. Both of these activities have been exceptionally rewarding and exciting for us.

We would like to thank many people who have read earlier versions of this book and made suggestions and corrections. These people include, but are not limited to, Porter Adams, Neil Makur, Nicholas Scoville, Lynn Sokei, Peterson Tretheway, Enrique Treviño, Nina Zubrilina, the anonymous referees, and all the TAs and students who have been part of the SUMaC community. This book also benefited from the contributions of Pierre Albin and Rafe Mazzeo, who have taught the class based on some earlier versions of this material. We would also like to thank Dahlia Fisch and the Springer production team for making this book a reality.

Contents

Chapter 1
Surface Preliminaries

1.1 Surfaces

One of the main objects of study in this book is that of a surface. We will thus spend a good deal of time in the first two chapters explaining what a surface is.

Informally, a surface is a mathematical object that "looks like a plane when we zoom in at any point." Or, just a bit more precisely, a surface is a set of points for which, around every point in the set, we can find a small neighborhood that can be deformed to a plane. Typical examples of surfaces are spheres, tori, and planes. When we refer to a sphere, we always mean just the surface of the sphere, not including the interior. We will meet these surfaces in more detail in the near future.

A very reasonable question you might have in mind at this point is why we are focusing on surfaces rather than some other sort of object. The reason is that surfaces have a number of convenient properties. For one thing, they can often be visualized so that we can use our already-existing intuition to make new concepts easier to grasp and work with. Surfaces are also nice because they aren't so trivial to understand so as be boring, but neither are they so complicated that we can't say much about them (at least, without considerably more background). The lower-dimensional version of surfaces, known as simple curves, have a very simple classification (although even here the proof of this classification still requires some work). On the other hand, higher-dimensional analogues of surfaces are extremely complicated and are not amenable to a simple description like the one we'll see for surfaces. So, surfaces are at just the right place for us along the scale from trivial to impossible.

Our informal description of a surface as something that looks like a plane when we zoom in at any point isn't mathematics yet, so let us now give a proper definition.

Definition 1.1 A *surface* S is a topological space such that for every point $p \in S$, there is an open set $U \subset S$ containing p, and a map $f : U \to V$ onto an open subset $V \subset \mathbb{R}^2$, so that f is a continuous bijection with a continuous inverse.

The above definition contains a lot of unfamiliar vocabulary that we will go through term by term in the forthcoming sections. Once we have defined all of this

C. Bray et al., *Algebraic Topology*,
https://doi.org/10.1007/978-3-030-70608-1_1

vocabulary carefully, we will revisit the definition of a surface in the next chapter with a deeper understanding.

1.2 Euclidean Space

Let us first assign some names and notation to interesting sets of numbers.

- $\mathbb{R}$ will be the set of real numbers.
- $\mathbb{N}$ will be the set of natural numbers: $\mathbb{N} = \{0, 1, 2, \ldots\}$.
- $\mathbb{Z}$ will be the set of integers:

$$\mathbb{Z} = \{\ldots, -2, -1, 0, 1, 2, \ldots\}.$$

  (We use the letter "Z" because it is the first letter of the German word "Zahlen," meaning "numbers.")
- $\mathbb{Q}$ will be the set of rational numbers, or fractions. (The "Q" stands for "quotient.")
- $\mathbb{C}$ will be the set of complex numbers.

Sometimes, we will use variants of these notations to mean rather predictable things; e.g. $\mathbb{R}_{>0}$ is the set of real numbers greater than 0.

In much of mathematics, and topology in particular, we like to think of sets of numbers in geometric terms. For instance, we think of the real numbers $\mathbb{R} = \mathbb{R}^1$ geometrically, as a line. We obtain "higher-dimensional" spaces as follows. If $S_1, S_2, \ldots, S_n$ are any sets, then we define their *Cartesian product* or *direct product* (or sometimes just *product*) to be the set

$$S_1 \times S_2 \times \cdots \times S_n = \{(s_1, s_2, \ldots, s_n) : s_i \in S_i \text{ for all } i\},$$

the set of all ordered n-tuples of elements, one from each S_i. As a special case of this, if S is any set and $n \geq 0$ is a given natural number, we write S^n for $S \times S \times \cdots \times S$, where there are n S's in the product. This is the set of n-tuples of elements of S. That is, $S^n = \{(s_1, s_2, \ldots, s_n) : \text{ each } s_i \in S\}$. If we apply this construction to $S = \mathbb{R}$, we get the higher-dimensional *Euclidean spaces*. Geometrically, we think of $\mathbb{R}^2$ as a plane, $\mathbb{R}^3$ as 3-dimensional space, and so forth.

The case $n = 0$ deserves special attention. By the construction above, we have $S^0 = \{()\}$. Writing it in that way is a bit unwieldy, so we prefer to think of S^0 as a set consisting of just one element, without necessarily giving that element the name $()$. Hence, for any S, S^0 is a single point.

Thus we have now explained the first, and most basic, unfamiliar term in Definition 1.1, namely $\mathbb{R}^2$. This is just the set of ordered pairs of real numbers. The second unfamiliar term is *topological space*. For now, a topological space is simply a *subset* of a Euclidean space of some dimension. That is our preliminary definition until we're ready for the correct definition of this concept, in Chapter 3.

1.3 Open Sets

The next unfamiliar term appearing in Definition 1.1 is an *open set*. We will discuss open sets in three stages: we start with open balls in Euclidean space, then open sets in Euclidean space, and finally open sets in topological spaces as previously defined.

Open Balls. An *open interval* in $\mathbb{R}$ with endpoints a and b is simply the set of numbers denoted (a, b) and defined by $(a, b) = \{x \in \mathbb{R} : a < x < b\}$. If the open interval takes the form $(p - r, p + r)$ for some $p \in \mathbb{R}$ and $r \in \mathbb{R}_{>0}$, then this interval has width $2r$ and is centered at p. An alternative characterization of $(p - r, p + r)$ is thus as the set of points whose distance to p is less than r. Mathematically speaking, we write $(p - r, p + r) = \{x \in \mathbb{R} : |x - p| < r\}$, since the inequality $|x - p| < r$ is equivalent to $p - r < x < p + r$. (Note: we will use the notation (a, b) for both the open interval and for a point in $\mathbb{R}^2$. It will always be possible from context to determine which one we mean!)

An *open ball* in the Euclidean space $\mathbb{R}^n$ is a generalization of the notion we have just described. We will need a notion of *distance* in Euclidean space. This is given by the Pythagorean formula: if $x = (x_1, x_2, \ldots, x_n)$ and $y = (y_1, y_2, \ldots, y_n)$ are two points in $\mathbb{R}^n$, then the distance between them is defined as

$$d(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2}.$$

We can now state our definition.

Definition 1.2 Let $r > 0$ be a positive real number, and let $p \in \mathbb{R}^n$ be a point. We let $B_r(p)$ be the set of points in $\mathbb{R}^n$ whose distance from p is less than r. That is,

$$B_r(p) = \{x \in \mathbb{R}^n : d(x, p) < r\}.$$

We call $B_r(p)$ the *open ball* of radius r centered at p.

Let us look at some examples of open balls in low dimensions.

Example Let $p \in \mathbb{R}$ be a point. Since $d(x, p) = |x - p|$ in this case, then indeed $B_r(p) = (p - r, p + r)$ as described above.

Example Let $p = (p_1, p_2) \in \mathbb{R}^2$ be a point in the plane. Then $B_r(p)$ is the set of points $(x_1, x_2) \in \mathbb{R}^2$ with $\sqrt{(x_1 - p_1)^2 + (x_2 - p_2)^2} < r$, or

$$(x_1 - p_1)^2 + (x_2 - p_2)^2 < r^2.$$

Hence, $B_r(p)$ consists of all the points on the inside of a circle of radius r centered at p.

Open Sets in Euclidean Spaces. We are ready to move on to open sets in $\mathbb{R}^n$. For the one-dimensional case, open sets can be characterized quite simply as unions of

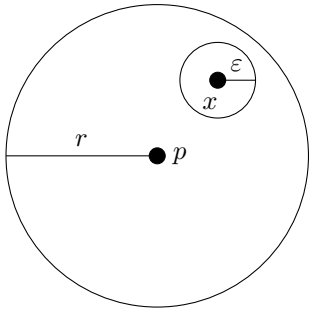


Figure 1.1 Open balls are open.

collections of disjoint open intervals. Hence, we may describe all open sets in $\mathbb{R}$ as being of the form $(a_1, b_1) \cup (a_2, b_2) \cup \cdots$, where each $a_i < b_i$, and all the intervals are disjoint. (We allow the endpoints to be at $\pm\infty$.) This union may contain finitely many or infinitely many terms; for example, the set

$$\bigcup_{n\in\mathbb{Z}} \left(n - \frac{1}{4}, n + \frac{1}{4}\right)$$

is an open set.

In higher dimensions, no such simple characterization exists. Instead, we capture the property of "openness" in a somewhat indirect way.

Definition 1.3 A set $S \subset \mathbb{R}^n$ is said to be *open* if, for every point $p \in S$, we can find some positive number r (which may depend on p) so that $B_r(p) \subset S$.

Example Open intervals inside $\mathbb{R}$ are open sets according to the definition above. To see this, let (a, b) be an open interval. (We can have a or b being equal to $\pm\infty$, and we ought to consider those cases separately. Let us assume, however, that $a, b \neq \pm\infty$.) For any $x \in (a, b)$, we have $a < x < b$, so let $r = \min(x - a, b - x) > 0$. Then for every y with $d(x, y) < r$, we have $y \in (a, b)$. Hence, $B_r(x) \subset (a, b)$. Thus (a, b) is open.

Example Open balls inside $\mathbb{R}^2$ are open sets according to the definition above. To see this, let $B_r(p)$ be an open ball and choose $x \in B_r(p)$. We must show that there exists a radius $\varepsilon > 0$ so that $B_\varepsilon(x) \subset B_r(p)$. An ε that will work is some number smaller than the distance between x and the edge of the circle of radius r centered at p; namely $\varepsilon = \frac{1}{2}(r - d(x, p))$. (Here, the fraction $\frac{1}{2}$ is arbitrary—the point is that it is less than one!) Now it is "pictorially" obvious that $B_\varepsilon(x) \subseteq B_r(p)$, based on Figure 1.1. But we can prove this rigorously using the *triangle inequality* (namely: $d(x, y) \leq d(x, z) + d(z, y)$ for any choices of $x, y, z \in \mathbb{R}^n$) as follows. Pick any $y \in B_\varepsilon(x)$. Then by definition $d(y, x) < \varepsilon$. Consequently,

$$\begin{aligned} d(y, p) &\leq d(y, x) + d(x, p) \\ &< \tfrac{1}{2}(r - d(x, p)) + d(x, p) \\ &= \tfrac{1}{2}r + \tfrac{1}{2}d(x, p) \\ &< \tfrac{1}{2}r + \tfrac{1}{2}r \\ &= r. \end{aligned}$$

Therefore $y \in B_r(p)$ because we have just shown that its distance to p is less than r. Since y was arbitrarily chosen inside $B_\varepsilon(x)$, we can say that $B_\varepsilon(x) \subset B_r(p)$.

Remark 1.4 Observe that we can actually replace the $\frac{1}{2}$ with 1 in the above example. However, it might not be so clear in advance that all the inequalities will work out if we do that. There is no prize for bravery here: no extra points are awarded for finding the best ε in town! So, it's better to be safe and choose something that you *know* is going to work.

Proposition 1.5 *The following are true of open sets in* $\mathbb{R}^n$*:*

(1) The union of an arbitrary number of open sets is open.
(2) The intersection of finitely many open sets is open.
(3) The empty set is open.
(4) The entire space $\mathbb{R}^n$ *is open.*

Proof (1) Let $A_1, A_2, \ldots$ be a collection of open sets (this collection may be finite, infinite and countable, or infinite and uncountable). We'll show that $A = A_1 \cup A_2 \cup \cdots$ is open as follows. Let x be an arbitrary element of A. Then $x \in A_i$ for some i. Since A_i is open, then there is $r > 0$ so that $B_r(x) \subset A_i$ by definition. Since $A_i \subset A$, then $B_r(x) \subset A$. Since this result holds for all $x \in A$, this means that A is open.

(2) Let $A_1, A_2, \ldots, A_N$ be a finite collection of sets. We'll show $A = A_1 \cap \cdots \cap A_N$ is open as follows. Let x be an arbitrary element of A. Then $x \in A_i$ for each i. Since A_i is open, there is some $r_i > 0$ so that $B_{r_i}(x) \subset A_i$. Can we construct a ball about x which is contained in *all* the A_i at once, i.e. such that the ball is contained in A? The answer is yes—let $r = \min\{r_1, \ldots, r_N\}$. Then $r > 0$ and $B_r(x) \subset B_{r_i}(x) \subset A_i$ for all i. Hence $B_r(x) \subset A$. Since this result holds for all $x \in A$, it follows that A is open.

(3) We argue that the empty set is open as follows. The definition requires that for a set A to be open, for every point $x \in A$, we can find ... Well, can we? In the empty set, there are no points to consider, so the conclusion holds for the entirety of the points in the empty set—i.e. none at all. The bottom line: the conclusion holds!

(4) Let $x \in \mathbb{R}^n$, and let $r > 0$ be any positive real number. Then $B_r(x) \subset \mathbb{R}^n$. Since x is arbitrary, this shows that $\mathbb{R}^n$ is open.

■

Remark 1.6 An infinite collection of open sets in $\mathbb{R}$ whose intersection is not open is $A_n = (-1/n, 1/n)$ for each $n = 1, 2, \ldots$ What is the intersection of all these sets? Prove that it is not open. Pinpoint where the proof of (2) fails for these sets.

Remark 1.7 Here is another way to think about statements about the empty set. Think of a procedure like that of determining whether a set is open as being a two-player game. The first player picks a point in a set, and the second player must produce a suitable r. Player 1 wins by producing a point for which player 2 cannot find a suitable r, and player 2 wins by finding such an r for every choice that player 1 makes. If the set is open, then player 2 has a winning strategy, whereas if the set is not open, then player 1 has a winning strategy. Who wins such a game in the case of the empty set? Player 2 of course, because player 1 can't even make a first move by presenting player 2 with a point.

Since vacuous statements are very important in mathematics but take some time to get used to, we should go through another (more frivolous) example of a vacuous statement. So consider *"All blue unicorns are pink."* This statement sounds like nonsense, since any blue unicorn would be blue and not pink, but it is actually true. In order for it *not* to be true, it would be necessary to exhibit a blue unicorn that fails to be pink. But there aren't any blue unicorns to begin with, so there is no chance of finding a counterexample.

In the future, the notation we will use for the empty set is $\varnothing$.

Open Sets in Topological Spaces. We finally arrive at the notion of an open set *inside* a given topological space S. This notion is sometimes referred to as *relative openness*.

Definition 1.8 Let S be a topological space in $\mathbb{R}^n$. An *open set* in S (also called a *relatively open* set in S) is the intersection of an open set U of $\mathbb{R}^n$ with S.

Example Let $\mathbb{S}^1$ be the circle $\{(x, y) \in \mathbb{R}^2 : x^2 + y^2 = 1\}$. The set $\{(x, y) \in \mathbb{S}^1 : y > 1/2\}$ is an open set of $\mathbb{S}^1$, because it is the intersection of $U = \{(x, y) \in \mathbb{R}^2 : y > 1/2\}$ with $\mathbb{S}^1$.

Figure 1.2 There aren't any of these.

There is another description of open sets in a topological space S that is easily seen to be equivalent to the previous one: a subset A of S is open in S if and only if, for every $p \in A$, we can find some r so that $B_r(p) \cap S \subset A$.

Related Notions. There are several other notions related to open sets which deserve to be mentioned, even though they do not appear explicitly in the definition of a surface that we have been studying. Basically, we would like to be able to describe a wider array of subsets of topological spaces. We begin with the idea of a *closed* set. First, recall that the complement of the set A, denoted A^c or $S \backslash A$, is defined as the set of points in S that are *not* in A. Mathematically speaking, $A^c = \{x \in S : x \notin A\}$.

Definition 1.9 Let S be a topological space, and let $A \subset S$ be a set. We say that A is *closed* if the complement of A is open.

Example The interval $[a, b] = \{x \in \mathbb{R} : a \leq x \leq b\}$ in $\mathbb{R}$ is closed. This set is known as the *closed interval* with endpoints a, b.

Example The set $\overline{B_r(p)} = \{x \in \mathbb{R}^n : d(x, p) \leq r\}$ is closed in $\mathbb{R}^n$. This set is known as the *closed ball* of radius r centered at p. We'll explain the meaning of the notation (i.e. the line hovering above the notation for the open ball) below.

Example The sets $[a, b) = \{x \in \mathbb{R} : a \leq x < b\}$ and the "half-open ball" in $\mathbb{R}^2$ given by

$$\{(x_1, x_2) : x_1^2 + x_2^2 \leq 1,\ x_1 < 0\} \cup \{(x_1, x_2) : x_1^2 + x_2^2 < 1,\ x_1 \geq 0\}$$

are neither open nor closed.

We see from the examples above that sets need not be either open or closed—and in some cases can even be both at once! But closed sets seem to contain all of their boundary points. To put this statement on a more rigorous footing, we make the following definitions.

Definition 1.10 Let S be a topological space, and let $A \subset S$ be a set. A point $x \in S$ is a *boundary point* of A if every ball centered at x contains points of A and of A^c. That is, for every $r > 0$ we have $B_r(x) \cap A \neq \varnothing$ and $B_r(x) \cap A^c \neq \varnothing$.

Definition 1.11 Let S be a topological space, and let $A \subset S$ be a set. The *boundary* of A, denoted ∂A, is the set of all boundary points of A. The *closure* of A is the set $\overline{A} = A \cup \partial A$. The *interior* of A is the set $A^\circ = A \cap \left(\partial A\right)^c$.

Example If $A = [a, b) \subset \mathbb{R}$ then $\partial A = \{a, b\}$, $\overline{A} = [a, b]$, and $A^\circ = (a, b)$.

The notion of boundary points allows us to make alternative characterizations of open and closed sets.

Proposition 1.12 *Let S be a topological space, and let $A \subset S$ be a set.*

(1) A is closed if and only if A contains all of its boundary points if and only if $A = \overline{A}$.
(2) A is open if and only if A^c *contains all of its boundary points if and only if* $A = A^\circ$.

Proof Exercise. ■

Finally, we conclude this section with another alternative description of closed sets that does not involve open sets. In order to do that, however, we will need the notion of a limit point.

Definition 1.13 Let S be a topological space, and let $a_1, a_2, \ldots$ be a sequence of points in S. Then a point $a \in S$ is a *limit* of the sequence $a_1, a_2, \ldots$ (also called a *limit point*) if, for every $\varepsilon > 0$, there is some $N \in \mathbb{N}$ so that, whenever $n > N$, $a_n \in B_\varepsilon(a)$.

It is easy to see that if a sequence has a limit point, then it is unique. Hence we may speak of *the* limit rather than merely *a* limit.

Example Let $S = \mathbb{R}$, and let $a_n = 1/n$. Then the limit of this sequence is 0.

Example Let $S = \mathbb{R}$, and let $b_n = n$. Then the sequence has no limit.

Remark 1.14 The choice of S in Definition 1.13 can be important. In the example above with $a_n = 1/n$, if we take $S = \mathbb{R}_{>0}$, then the sequence has no limit. The only possible candidate for the limit point would be 0, but $0 \notin S$.

Theorem 1.15 *Let S be a topological space. A subset* $A \subset S$ *is closed if and only if, whenever* $a_1, a_2, \ldots$ *is a sequence of points in A approaching some point* $a \in S$, *we have* $a \in A$.

Proof First, suppose A is a closed set, and let $a_1, a_2, \ldots$ be a sequence of points in A which approach some $a \in S$. We must show that $a \in A$. Suppose $a \in A^c$. Then, since A^c is open, there is some $r > 0$ so that $B_r(a) \subset A^c$. But since the sequence of a_i's approaches a, there is some n so that $a_n \in B_r(a)$. Hence, $a_n \in A^c$. But we assumed that each a_i was in A, so we have a contradiction. Hence $a \in A$, as desired.

Now suppose that whenever $a_1, a_2, \ldots$ is a sequence of points in A approaching some point $a \in S$, we have $a \in A$. Pick some point $x \in A^c$. Then there is no sequence of points $a_1, a_2, \ldots$ of points in A approaching x. Let us now consider the sequence of open sets $U_n = B_{1/n}(x) \cap S$ of S. We claim that we can find some n so that $U_n \subset A^c$. If not, then we can find some $a_n \in U_n \cap A$ for each n. But then the sequence $a_1, a_2, \ldots$ is in A and approaches x, which is a contradiction. Thus the open neighborhood U_n of x in S is contained in A^c. Since this works for an arbitrary $x \in A^c$, it follows that A^c is open and thus that A is closed. ■

1.4 Functions and Their Properties

The definition of a surface presented earlier uses several unfamiliar terms to describe functions. In fact, it also uses the term *function* in perhaps a slightly different way from what you might be used to.

To define the concept of a function $f : A \to B$ where A and B are two sets, we won't actually need much beyond the colloquial formulation "a function $f : A \to B$ is a rule which assigns a unique element $f(a) \in B$ to each element $a \in A$." For completeness, a more mathematically rigorous definition is the following.

Definition 1.16 A *function* $f : A \to B$ is a subset $\mathcal{F}$ of the Cartesian (or direct) product $A \times B$ which satisfies the properties:

- $A = \{a : (a, b) \in \mathcal{F}\}$;
- if (a, b_1) and (a, b_2) both belong to $\mathcal{F}$ then $b_1 = b_2$.

Remark 1.17 Very few mathematicians actually think about a function in terms of this definition. Instead, mathematicians tend to think of a function as a box, perhaps with an intricate set of gears and cranks, that eats an element of A as its input and spits out an element of B as its output. However, it is sometimes useful when proving things to have the more formal definition available to us.

In a calculus course, the sets A and B are usually subsets of $\mathbb{R}$ (or perhaps $\mathbb{R}^n$ in a multivariable calculus course), and functions are given by mathematical formulas that describe the operations to be carried out on the input numbers to produce the output numbers. By contrast, in a topology course, we would like to take a more *geometric* perspective in which A and B are topological spaces, and functions take points in A and convert them into points in B. Of course A and B are still subsets of $\mathbb{R}^n$ and the operations carrying points from A to B may still be described using mathematical formulas; we just want to think *geometrically* about what is going on.

Example Here are some different types of functions that can be understood from a geometric perspective.

(1) Familiar functions $f : \mathbb{R} \to \mathbb{R}$, such as $f(x) = x^3 - 5x$ and $f(x) = \cos(x)$ are functions.
(2) Functions of several variables, i.e., $f : \mathbb{R}^n \to \mathbb{R}$, such as $f(x, y) = x - y^2$ or $f(x, y, z) = xye^{e^z}$ are functions.
(3) We can also consider functions $f : \mathbb{R} \to \mathbb{R}^n$. Consider, for instance, $f(x) = (x^2, x^3)$, as a function $\mathbb{R} \to \mathbb{R}^2$. Considering such functions is useful for studying curves: if we look at the subset of $\mathbb{R}^2$ which is the image of f, we get an interesting curve in the plane. Another example of a curve in the plane is the image of the function $f : \mathbb{R} \to \mathbb{R}^2$ given by $f(t) = (\cos t, \sin t)$, i.e., a circle of radius 1.
(4) Functions $f : \mathbb{R}^2 \to \mathbb{R}^3$ can similarly be used to study surfaces in $\mathbb{R}^3$. For example, the image of the function $f : \mathbb{R}^2 \to \mathbb{R}^3$ given by $f(\theta, \phi) = (\cos\theta \sin\phi, \sin\theta \sin\phi, \cos\phi)$ is a sphere of radius 1.

(5) We can write down functions $f : \mathbb{R}^2 \to \mathbb{R}^2$ that do specific geometric things to points in the plane. For example, a function that rotates points counterclockwise about the origin by an angle of θ is given by $f(x, y) = (x \cos\theta - y \sin\theta, x \sin\theta + y \cos\theta)$.
(6) We can also write down equations of projections. For example, we have the projection $p : \mathbb{R}^3 \to \mathbb{R}^2$ onto the xy-plane, given by $p(x, y, z) = (x, y)$. Similarly, we can project onto the z-axis by taking $q : \mathbb{R}^3 \to \mathbb{R}$ to be $q(x, y, z) = z$.
(7) Slightly more deviously, we can project onto the xy-plane *inside of* $\mathbb{R}^3$ by taking $p : \mathbb{R}^3 \to \mathbb{R}^3$ to be $p(x, y, z) = (x, y, 0)$.

Before we continue, let us recall some standard terminology. For a function $f : A \to B$, the set A is called the *domain* of f, B is called the *codomain*, and the *range* or *image* of f is the set $f(A) = \{b \in B : b = f(a) \text{ for some } a \in A\}$. Therefore $f(A) \subset B$. We now return to the unfamiliar terminology relating to functions that first appeared in Definition 1.1.

Definition 1.18 Let A and B be two sets, and let $f : A \to B$ be a function from A to B.

- We say that f is *surjective*, or *onto*, if for every $b \in B$, there is some $a \in A$ so that $f(a) = b$. In other words, $f(A) = B$.
- We say that f is *injective*, or *one-to-one*, if whenever $f(a) = f(a')$ we have $a = a'$.
- We say that f is *bijective*, or a *bijection*, if it is both injective and surjective.

Exercise 1.19 Of the functions listed in the previous example, which are surjective? Injective?

Bijective functions are special in that they have *inverses*, as we now explain. The inverse of a function $f : A \to B$ is a function $g : B \to A$ that "undoes" the action of f; that is $g(f(x)) = x$ for all $x \in A$. We would also like this relation to hold with the roles of f and g reversed; in other words, f is the inverse of g and $f(g(y)) = y$ for all $y \in B$. A more succinct way of saying this is that $f \circ g = \mathrm{id}_B$ and $g \circ f = \mathrm{id}_A$, where id is the identity function on the appropriate space, defined by $\mathrm{id}(x) = x$ for all x in this space; and $\circ$ denotes the *composition* of functions $f \circ g(x) = f(g(x))$ defined whenever the range of g is contained in the domain of f.

Here is the reason why bijective functions have inverses. If $f : A \to B$ is bijective, then by surjectivity, for every $b \in B$ there is some $a \in A$ so that $f(a) = b$. Moreover, by injectivity, this a is unique: for if any other $a' \in A$ satisfies $f(a') = b$, then $a' = a$. Hence we can define a *function* (in the sense of the mathematically precise definition of this concept given above) $g : B \to A$ by the rule

$$g(b) = a \text{ where } a \text{ is such that } f(a) = b.$$

Note that both injectivity and surjectivity are needed for the inverse to be well-defined. A non-injective function will have more than one point in A mapping to the

same point in B, while for a non-surjective function, it will be the case that there is at least one $b \in B$ that has no points in A mapping to it.

Notation The inverse of a bijective function $f : A \to B$ is denoted $f^{-1} : B \to A$.

Example Let $f : \mathbb{R} \to \mathbb{R}_{>0}$ be defined by $f(x) = e^x$. Then f is a bijective function, and its inverse is $f^{-1}(x) = \log(x)$. As respectable mathematicians, we write "log" to denote the base-e logarithm.

In contrast to the previous definition of the inverse of a bijective function, the following definition holds for any function at all. Be careful not to confuse the notation!

Definition 1.20 Let $f : A \to B$ be a function and let $S \subset B$ be a subset. Then the *inverse image* or *preimage* of S under f is the set

$$f^{-1}(S) = \{a \in A : f(a) \in S\}.$$

In other words, $f^{-1}(S)$ is the set of points in A mapped into S by f.

Example Let $f : A \to B$ be a function, and let $b \in B$ be any point. Then the *level set* of f at b is the set $f^{-1}(\{b\}) = \{a \in A : f(a) = b\} \subset A$. Typically, we write $f^{-1}(b)$ instead of $f^{-1}(\{b\})$. If $B \subset \mathbb{R}$, then we also have a notion of a sublevel set: the *sublevel set* is the set $f^{-1}((-\infty, b]) = \{a \in A : f(a) \leq b\}$.

Example Let us consider Example (7) on page 10 above. Then:

- $f^{-1}((1, 3, 0)) = \{(1, 3, z) : z \in \mathbb{R}\}$.
- $f^{-1}(\{(x, y, 0) \mid x^2 + y^2 \leq 1\}) = \{(x, y, z) \mid x^2 + y^2 \leq 1\}$.
- $f^{-1}((2, 3, 4)) = \varnothing$.

Example Let $f : \mathbb{R}^3 \to \mathbb{R}$ be defined by $f(x, y, z) = x^2 + y^2 + z^2$. Then $f^{-1}(1)$ is the sphere of radius 1 centered at the origin.

1.5 Continuity

The final concept that we will need to describe in order to make sense of the definition of a surface given at the beginning is that of continuity of functions between topological spaces. But before we define this concept properly, let us say informally what it means.

Let $S \subset \mathbb{R}^m$. Roughly, a function $f : S \to \mathbb{R}^n$ is continuous if it sends nearby points in S to nearby points in $\mathbb{R}^n$.

Our immediate task is to convert this intuitive idea into formal mathematics. Our first hope might be that when we take a small open set U in S, then $f(U)$ is a small open set in $\mathbb{R}^n$. But a quick check shows that this is not quite right: If f is a constant

function, which sends everything in S to one point in $\mathbb{R}^n$, then the image of any open set in S is just a single point, which is not open.

This first idea didn't quite work, but we are on the right track. Let us instead look at nearby points in $\mathbb{R}^n$ and see where they come from in S. More precisely, let V be an open set in $\mathbb{R}^n$, and look at $f^{-1}(V)$. If $x \in f^{-1}(V)$, then $f(x) \in V$, so if y is close to x, then $f(y)$ should also be in V. We can rephrase that to say that, if V is open in $\mathbb{R}^n$, then $f^{-1}(V)$ is open in S. This will, in fact, be the definition of a continuous function.

Definition 1.21 Let $S \subset \mathbb{R}^m$. A function $f : S \to \mathbb{R}^n$ is said to be *continuous* if, for every open set $V \subset \mathbb{R}^n$, $f^{-1}(V) \subset S$ is an open set of S.

A different definition of continuity is commonly given in calculus classes—the famous ε–δ definition that you may be familiar with. In the case where S is an *open* subset of $\mathbb{R}^m$, it is equivalent:

Theorem 1.22 *Let S be an open subset of $\mathbb{R}^m$, and let $f : S \to \mathbb{R}^n$ be a function. Then f is continuous if and only if, for every point $x \in S$ and every $\varepsilon > 0$, there is some $\delta > 0$ so that, whenever $x' \in S$ and $d(x, x') < \delta$, then $d(f(x), f(x')) < \varepsilon$.*

Proof Assume that $f : S \to \mathbb{R}^n$ is continuous according to the topological definition of continuity. We will show that the calculus definition holds. Pick some $x \in S$, and let $y = f(x) \in \mathbb{R}^n$. Pick also some $\varepsilon > 0$, and let $V = B_\varepsilon(y) \subset \mathbb{R}^n$. Since f is continuous, $f^{-1}(V) \subset S$ is open, so it is of the form $U \cap S$, for some open set $U \subset \mathbb{R}^m$. Now, since U is open in $\mathbb{R}^m$ and $x \in U$, there is some $\delta > 0$ so that $B_\delta(x) \subset U$. Hence $U \cap S$ contains all points $x' \in S$ with $d(x, x') < \delta$, and $f(U) \subset V$. This shows that whenever $x' \in S$ and $d(x, x') < \delta$, then $d(f(x), f(x')) < \varepsilon$.

Now let's do the other direction: suppose that for every $x \in S$ and every $\varepsilon > 0$, we can find some $\delta > 0$ so that if $d(x, x') < \delta$, then $d(f(x), f(x')) < \varepsilon$. Let us pick some open set $U \subset \mathbb{R}^n$. For every $y \in U$, we can find some r_y so that $B_{r_y}(y) \subset U$, so that

$$U = \bigcup_{y \in U} B_{r_y}(y).$$

Note that

$$f^{-1}(U) = \bigcup_{y \in U} f^{-1}(B_{r_y}(y)),$$

which will be open in S if each $f^{-1}(B_{r_y}(y))$ is open in S. So, now we are reduced to showing that $f^{-1}(B_{r_y}(y))$ is open in S. Fix some particular y, and let $r = r_y$, to simplify notation.

We want to show that $f^{-1}(B_r(y))$ is open in S. Pick some point $a \in f^{-1}(B_r(y))$, and let $b = f(a) \in B_r(y)$. Since $B_r(y)$ is open, we can find some $\varepsilon > 0$ so that $B_\varepsilon(b) \subset B_r(y)$. By the hypothesis, we can find some $\delta > 0$ so that if $d(a, a') < \delta$, then $d(b, f(a')) < \varepsilon$, i.e., $f(a') \in B_\varepsilon(b) \subset B_r(y)$, so $a' \in f^{-1}(B_r(y))$. So, $B_\delta(a) \subset f^{-1}(B_r(y))$. Hence $f^{-1}(B_r(y))$ is open in S. We can therefore conclude that f is continuous. ∎

There are several advantages to using the topological definition of continuity over the calculus one. For one thing, it is a lot simpler to state and understand. It is also more general, and might be easier to work with at times. Let us give an example to see how easy it is to use.

Theorem 1.23 *Let $f : A \to B$ and $g : B \to C$ be two continuous functions. Then their composition $g \circ f : A \to C$ is continuous.*

Proof Let $W \subset C$ be an open set. We must show that $(g \circ f)^{-1}(W)$ is an open set in A. Since g is continuous, $V = f^{-1}(W)$ is an open set in B. Since f is continuous, $U = g^{-1}(V)$ is an open set in A. But $U = (g \circ f)^{-1}(W)$. So $g \circ f$ is continuous. ■

Imagine how much more annoying this would be using ε's and δ's!

Let us now look at some examples of continuous and discontinuous functions.

Example Most of the functions $f : \mathbb{R} \to \mathbb{R}$ that you are friends with are continuous. For example, the function $f(x) = x + 1$ is continuous. This is probably completely obvious, but let us prove it, just to make sure we can really trust our intuition. Let $U \subset \mathbb{R}$ be an open set. We have to show that $f^{-1}(U)$ is also open. We already know what all the open sets in $\mathbb{R}$ look like: they are unions of open intervals, $U = \bigcup_{i \in I}(a_i, b_i)$. Then

$$f^{-1}(U) = \bigcup_{i \in I} f^{-1}((a_i, b_i)),$$

so in order to show that $f^{-1}(U)$ is open, we only have to show that each $f^{-1}((a_i, b_i))$ is open. But $f^{-1}((a_i, b_i)) = (a_i - 1, b_i - 1)$, which is an open interval and hence open. Thus f is continuous.

Example $f(x) = x^2$ is also continuous. To see this, we must take an open set $U \subset \mathbb{R}$ and show that $f^{-1}(U)$ is also open. Suppose $U = \bigcup_{i \in I}(a_i, b_i)$. Since

$$f^{-1}(U) = \bigcup_{i \in I} f^{-1}((a_i, b_i)),$$

we can assume that U is just an interval $U = (a, b)$. The preimage of U is then $\{x \in \mathbb{R} : a < x^2 < b\}$. Let us now break up the problem into a few cases, based on whether a and b are positive or negative.

Case 1: $a \geq 0$. Then $f^{-1}(U) = (-\sqrt{b}, -\sqrt{a}) \cup (\sqrt{a}, \sqrt{b})$, which is the union of two open intervals and hence open.

Case 2: $a < 0 < b$. Then $f^{-1}(U) = (-\sqrt{b}, \sqrt{b})$, which as we just mentioned is an open interval and hence open.

Case 3: $b \leq 0$. Then $f^{-1}(U) = \varnothing$, which is open.

So, in all cases, the preimage of an open interval is an open set, so the preimage of any open set is open. Hence $f(x) = x^2$ is a continuous function.

Similarly, any polynomial $f : \mathbb{R} \to \mathbb{R}$ is continuous, as are rational and algebraic functions, trigonometric functions, exponential functions, logarithmic functions, and so forth, on the intervals on which they are defined. (So, $f(x) = \tan(x)$ is continuous on $(-\pi/2, \pi/2)$, but not on any interval containing $\pi/2$, since $f(x)$ is not defined at that point.)

Piecewise-defined functions, ones that have different sorts of definitions at different points, are likely to be discontinuous.

Example Let

$$f(x) = \begin{cases} 0 & x \leq 0, \\ 1 & x > 0. \end{cases}$$

Then f is discontinuous. In order to show this, we must find some open set $U \subset \mathbb{R}$ so that $f^{-1}(U)$ fails to be open. Let $U = (-1/2, 1/2)$. Then $f^{-1}(U) = (-\infty, 0]$, which is not open. This shows that f is discontinuous.

All is not lost with the function in this example, however. Something bad is going on at 0, but everywhere else, it "looks" continuous, in that nearby points get sent to nearby points. To make this precise, we now define the notion of continuity at a point. This notion of continuity, too, can be described both in terms of open sets as well as in terms of δ's and ε's. We give both definitions (which are equivalent to each other), starting with the open set definition.

Definition 1.24 Let $f : S \to \mathbb{R}^n$ be a function, and let $x \in S$ be a point. We say that f is continuous at x if, for every open set $V \subset \mathbb{R}^n$ containing $f(x)$, there is an open set $U \subset S$ containing x such that $f(U) \subset V$.

Definition 1.25 Let $f : S \to \mathbb{R}^n$ be a function, and let $x \in S$ be a point. We say that f is continuous at x if, for every $\varepsilon > 0$, there is some $\delta > 0$ so that if $d(x, x') < \delta$, then $d(f(x), f(x')) < \varepsilon$.

The only difference between this definition and the alternative characterization of continuous functions above is that here we are not allowed to vary x.

It is easy to see that the function f in the last example is continuous for all $x \neq 0$. But let us check, just so that we get a bit more practice using the definition. Pick some point $x \neq 0$. Now, if $|x - x'| < |x|$, then $f(x) = f(x')$. So, for any $\varepsilon > 0$, if we take $\delta = |x|$, then whenever $d(x, x') < \delta$, we have $d(f(x), f(x')) = 0 < \varepsilon$. Thus f is continuous at x.

A common way of interpreting continuity is that the graph can be drawn without lifting the pen from the paper. This notion is called *path-connectedness*, and we will discuss it later in the book. However, continuity is a bit more subtle than this. Let us see an example which shows some of the subtleties of continuity.

Example Let $f : \mathbb{R} \to \mathbb{R}$ be the function, known as the *Riemann function* or the *Thomae function*, defined by

$$f(x) = \begin{cases} \frac{1}{q} & \text{if } x = \frac{p}{q} \text{ in lowest terms,} \\ 0 & \text{if } x \text{ is irrational.} \end{cases}$$

Since the graph of $f(x)$ is a mess of points whose y-coordinates are between 0 and 1, we might expect that f is not continuous at any point. However, this is false: f is continuous at all the irrational numbers and discontinuous at all the rationals. Let us carefully examine why this is the case. First, let x be rational, say $x = \frac{p}{q}$. Pick $\varepsilon < \frac{1}{q} = f(x)$. For any $\delta > 0$, we can find some irrational number x' with $d(x, x') < \delta$. For such an x', we have $f(x') = 0$, so $d(f(x), f(x')) = \frac{1}{q} > \varepsilon$. In other words, we can't find any appropriate $\delta > 0$ for this choice of ε.

Now let's look at the irrational points, which are more interesting. Let x be irrational, so that $f(x) = 0$, and pick some $\varepsilon > 0$. Then there is some $Q > 0$ with $\frac{1}{Q} < \varepsilon$. Pick $\delta > 0$ to be less than the minimum of the distances from x to each $\frac{p}{q}$ with $q < Q$. Then if $d(x, x') < \delta$ and x' is rational, then the denominator of x' must be at least Q, so $f(x') < \varepsilon$. Thus if $d(x, x') < \delta$, then $d(f(x), f(x')) < \varepsilon$. In other words, f is continuous at x. So, even though the graph of f can't be drawn without lifting the pen off the paper at any point, it is still continuous at some—even most!—points.

Just for fun, let us present an exercise which is way too hard.

Exercise 1.26 Can you find an example of a function $f : \mathbb{R} \to \mathbb{R}$ that is continuous at all the rational numbers and at none of the irrational numbers?

The more "interesting" functions that we will be looking at in this book are functions between higher-dimensional topological spaces. That is, functions $f : S \to \mathbb{R}^m$ where $S \subset \mathbb{R}^n$ for either $m > 1$ or $n > 1$. Continuity here is also a more subtle concept, since an open set $V \subset \mathbb{R}^m$ is more complicated than a union of intervals, and so $f^{-1}(V)$ can fail to be open in much more interesting ways, and of course f itself can act in much more complicated ways.

There are some very important properties of continuous functions that we will use repeatedly throughout this book.

Theorem 1.27 *Let $S \subset \mathbb{R}^m$, and let $f : S \to \mathbb{R}^n$ be a continuous function. Let $p \in \mathbb{R}^n$ be any point. Then $f^{-1}(p)$ is a closed subset of S.*

Proof We begin with the observation that $f^{-1}(p) = S \setminus f^{-1}(\mathbb{R}^n \setminus \{p\})$, so it suffices to show that $f^{-1}(\mathbb{R}^n \setminus \{p\})$ is open. Since $\mathbb{R}^n \setminus \{p\}$ is an unwieldy name, let us rename it U. Since f is assumed to be continuous, it suffices to show that U is open in order to show that $f^{-1}(U)$ is open. Let $x \in U$ be an arbitrary point. We must show that there is some r, depending on x, so that $B_r(x) \subset U$. We can take $r = d(x, p)$ so that $p \notin B_r(x)$; note that p is the only point in $\mathbb{R}^n$ which is *not* in U, so any set that avoids p is contained in U. Hence U is open, so $f^{-1}(U)$ is open, so $f^{-1}(p)$ is closed, as desired. ■

A very similar argument shows, more generally, that the preimage under a continuous function of any closed set is closed. In fact, this could be used as a definition of

continuity instead of the same condition for open sets, and it is occasionally useful for checking whether functions are continuous.

1.6 Problems

(1) Prove the following. Be super rigorous for at least one of your proofs; you can be less so for the others.

   (a) The rectangle $A = (0, 1) \times (0, 1) := \{(x_1, x_2) : 0 < x_i < 1 \text{ for } i = 1, 2\}$ is open. Your proof should consist of picking a point $(x_1, x_2) \in A$ and then determining a value for r for which $B_r(x_1, x_2) \subseteq A$. Prove this last assertion as rigorously as you can.
   (b) The open half-plane $A = \{(x, y) : ax + by < 0\}$, where a, b are fixed real numbers, is open. (For concreteness you can choose $a = 1$ and $b = 2$ if you like. But try to construct a proof for general a, b.)
   (c) The interval $[a, b]$ is closed.
   (d) The closed ball $\{(x, y) \in \mathbb{R}^2 : x^2 + y^2 \leq 1\}$ is closed.
   (e) The line $L = \{(x, y) : ax + by = 0\}$, where $a, b \in \mathbb{R}$ are constants defining the slope of the line, is closed. What is its boundary?

(2) Write down a set in the plane that is neither open nor closed, and explain why. Is there a set which is *both* open and closed?
(3) Draw graphs of the functions $f_1(x) = \sin(x)$ and $f_2(x) = \arctan(x)$ and $f_3(x) = x^4 - x^2$ and $f_4(x) = \log(x)$ and $f_5(x) = (x - 1)/(x + 1)$. For each $i = 1, \ldots, 5$ do the following if possible. (Do as many as you need until you are convinced that you can do the rest without any significant effort.)

   (a) Find subsets $A, B \subseteq \mathbb{R}$ so that $f_i : A \to B$ is injective.
   (b) Find subsets $A, B \subseteq \mathbb{R}$ so that $f_i : A \to B$ is injective but not surjective.
   (c) Find subsets $A, B \subseteq \mathbb{R}$ so that $f_i : A \to B$ is surjective.
   (d) Find subsets $A, B \subseteq \mathbb{R}$ so that $f_i : A \to B$ is surjective but not injective.
   (e) Find subsets $A, B \subseteq \mathbb{R}$ so that $f_i : A \to B$ is bijective.
   (f) Find a formula for f_i^{-1} when f_i maps as in part (e).
   (g) Find $f_i^{-1}([-2, -1])$ and $f_i^{-1}([\frac{1}{2}, 1])$. (Here the notion $f^{-1}(A)$ can be defined even if f doesn't have an inverse. It means: the set of all points that map into A under f. I.e. $f^{-1}(A) = \{x : f(x) \in A\}$.)

(4) Let $f : \mathbb{R}^2 \to \mathbb{R}^3$ be defined by $f(x_1, x_2) = (x_1, x_2, 1)$. Is this function injective, surjective or both? What is $f^{-1}(B_2(0, 0, 0))$? Draw pictures.
(5) Describe with pictures and/or write down formulae for reasonably simple functions $f : \mathbb{R}^2 \to \mathbb{R}^3$ that are not injective; not surjective; neither injective nor surjective.
(6) Let $f : X \to Y$ be any function, and let $V \subset Y$ be any subset. Show that $f^{-1}(V^c) = (f^{-1}(V))^c$.

(7) Suppose that $f : X \to Y$ and $g : Y \to X$ are functions so that $g \circ f : X \to X$ is the identity: $(g \circ f)(x) = x$ for all $x \in X$. Show that f is injective and g is surjective.

(8) Use the definition of continuity in terms of inverse images of open sets in this problem.

(a) Show that $f(x) = 3x + 5$ is continuous as a function from $\mathbb{R}$ to itself.
(b) Show that $f(x, y) = x + y$ is a continuous function from the plane to the line.
(c) Show that $f(x, y) = 1$ is a continuous function from the plane to the line.
(d) Show that the function $f(x, y)$ that equals 1 when $(x, y) \neq (0, 0)$ and $f(0, 0) = 0$ is not continuous as a function from the plane to the line.

(9) Let $f : A \to B$ be a function (not necessarily bijective).

(a) Show that $U \subseteq f^{-1}(f(U))$ for all $U \subseteq A$. What is the most general condition under which you can prove $U = f^{-1}(f(U))$?
(b) Show that $f(f^{-1}(V)) \subseteq V$ for all $V \subseteq B$. What is the most general condition under which you can prove $f(f^{-1}(V)) = V$?

Mathematical Symbols

- $\forall$ "for all"
- $\exists$ "there exists"
- $\in$ "belongs to" or "in" (refers to a point p belonging to a set S as in $p \in S$)
- $\subseteq$ or $\subset$ "contained in" or just "in" (refers to a subset of points A contained in a set S as in $A \subseteq S$)
- $\subsetneq$ "strictly contained in" or "contained in but not equal to"
- $\{x$: blah blah$\}$ "the set of all x such that blah blah holds"
- $\equiv$ or $:=$ "defined as" (as in $\mathbb{R}_+ \equiv \{x \in \mathbb{R} : x > 0\}$)
- $f : A \to B$ "the function f maps the set A to the set B"
- $\cap$ "intersection" i.e. $A \cap B =$ the points simultaneously in A and in B
- $\cup$ "union" i.e. $A \cup B =$ the points either in A or in B

Greek Letters

Lower case:

α alpha	β beta	γ gamma	δ delta
ε epsilon	ζ zeta	η eta	θ theta
ι iota	κ kappa	λ lambda	μ mu
ν nu	ξ xi	o omicron	π pi
ρ rho	σ sigma	τ tau	υ upsilon
ϕ phi	χ chi	ψ psi	ω omega

Upper case:

A alpha	B beta	Γ gamma	Δ delta
E epsilon	Z zeta	H eta	Θ theta
I iota	K kappa	Λ lambda	M mu
N nu	Ξ xi	O omicron	Π pi
R rho	Σ sigma	T tau	Υ upsilon
Φ phi	X chi	Ψ psi	Ω omega

Chapter 2
Surfaces

2.1 The Definition of a Surface

Let us take a moment to remind ourselves of the definition of a surface given in the previous chapter (Definition 1.1). We introduce the terminology *homeomorphism* to mean *a continuous bijective function with a continuous inverse*.

> A *surface* S is a topological space such that for every point $p \in S$, there is an open set $U \subset S$ containing p, and a homeomorphism $f : U \to V$, where V is some open subset of $\mathbb{R}^2$.

Now that we know all the words in the definition, let us take another look at this definition and re-interpret it. We would like to be able to say that S is a surface if and only if "it locally looks like a piece of the plane." Mathematically speaking, the word "locally" is (usually) equivalent to "for every $p \in S$ there is an open set U containing p" such that the surface property holds; and "looks like a piece of the plane" is equivalent to the existence of the homeomorphism $f : U \to V \subset \mathbb{R}^2$. Informally, this homeomorphism "flattens" U into a subset V of $\mathbb{R}^2$ in an invertible way. Thus the definition above precisely captures what we would like!

Remark 2.1 According to our definition, we do allow surfaces to be empty, i.e. not contain any points, but we shall often tacitly assume that our surfaces are nonempty so as to avoid trivialities.

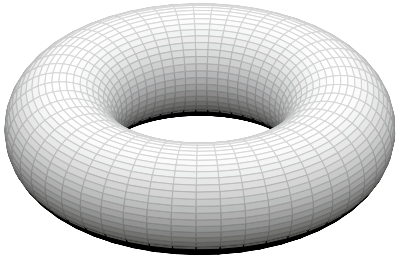
2.2 Examples of Surfaces

We now look at some examples and nonexamples of surfaces. We won't prove anything yet—we'll save an explicit proof of the surface property in a particular case for the next section. For now, we'll concentrate on building intuition.

Example *Planes.* For any point $p \in \mathbb{R}^2$, we can take U to be the full plane $\mathbb{R}^2$ and f to be the identity function $f(x) = x$. So a plane is a surface. More generally, any nonempty open subset of a plane is a surface.

C. Bray et al., *Algebraic Topology*,
https://doi.org/10.1007/978-3-030-70608-1_2

Figure 2.1 A torus.



Example *Spheres.* A sphere is a surface. We are all very familiar with this fact, because we live on a sphere. If we look at just a small piece of a sphere, then it looks as though it could be part of a plane; this is why people in past centuries believed that Earth was flat. The latitude and longitude coordinates constitute a homeomorphism from a large part of the sphere (but not all of the sphere—which part is missing?) to a part of the plane (i.e. latitude $\phi \in (-\pi/2, \pi/2)$ and longitude $\theta \in (0, 2\pi)$ give us the point $(\phi, \theta) \in (-\pi/2, \pi/2) \times (0, 2\pi) \subset \mathbb{R}^2$). We refer to the sphere as $\mathbb{S}^2$.

Example *Tori.* A torus is the surface of a bagel: something that has a hole in it. (See Figure 2.1.) A small piece of a (sufficiently large) torus once again looks like a piece of a plane. If we were to live on a torus-shaped planet, we would not be able to notice this easily without looking at a large portion of the planet. We refer to the torus as $\mathbb{T}$.

Example *Cubes.* A cube also looks locally like a plane. In fact, for a point on a face (rather than an edge or vertex), a neighborhood of that point *does* lie on a plane. For a point on the edge, a neighborhood lies in two adjacent faces. We must bend one of the faces to lie in the plane of the other, and then the neighborhood will lie in a plane. Can you see how to deal with the vertices?

We now come to a nonexample of a surface.

Nonexample Let X be the union of the xy-plane and the xz-plane in $\mathbb{R}^3$. (See Figure 2.2.) Then X is *not* a surface. A neighborhood of the point $(0, 0, 0) \in X$ does not look like a piece of a plane; it looks like a piece of two planes. (In fact, it is a piece of two planes.) Actually *proving* that X is not a surface is challenging, and we won't be able to do it completely, but here is something close to a proof: An open neighborhood U of $(0, 0, 0)$ is the union of two pieces Y and Z, where Y is an open subset of the xy-plane, and Z is an open subset of the xz-plane, and $Y \cap Z$ is a line segment (or union of line segments) on the x-axis. By passing to a smaller open neighborhood if necessary, we may assume that $Y \cap Z$ is a single line segment. Now, suppose $f : U \to V$ is a homeomorphism from U to an open subset $V \subseteq \mathbb{R}^2$. Then $U \setminus (Y \cap Z)$ consists of four pieces, but $V \setminus f(Y \cap Z)$ consists of only two. But the number of pieces (connected components) of a space does not change when we apply a homeomorphism.

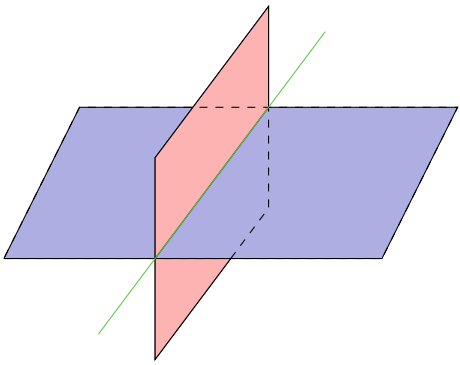


Figure 2.2 The union of two planes.

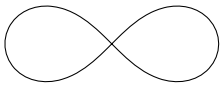


Figure 2.3 The figure-eight curve.

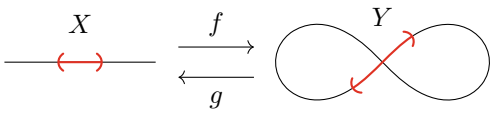


Figure 2.4 An interval X maps continuously (via f) to the figure-eight curve Y, but g is not continuous, since the preimage of the interval in red is not open.

The reason this argument is not completely satisfactory is that we do not yet know what the image of $Y \cap Z$ under a homeomorphism can look like. How do we know that it divides V into two pieces, and not four? It is true that it divides it into two pieces, but this is not obvious. A related result is the notorious *Jordan Curve Theorem*, which says that a simple closed curve (i.e. the image of a circle under an injective continuous map) in $\mathbb{R}^2$ divides the plane into two regions: an inside and an outside. While this is "obvious," the proofs are highly nontrivial. We will prove it in Chapter 14.

There is one piece of the definition of a surface that may seem a little bit mysterious: If f is a continuous bijection, is it possible that its inverse could fail to be continuous? Let us demonstrate that this is possible.

Nonexample Let X be the open interval $(0, 1)$, and let Y be the figure-eight curve. (See Figure 2.3.) We define a function $f : X \to Y$, which is obtained by wrapping the interval around the figure-eight. Then f is a continuous bijection. (Note that the only preimage of the nodal point in the middle is $1/2 \in X$, because X is an *open* interval.) Let $g : Y \to X$ be the inverse function of f. Then g is discontinuous: the preimage of the interval $(.4, .6)$ contains points on only two edges leaving the node, rather than on all four, whereas any open neighborhood of the node contains points on all four. (See Figure 2.4.)

2.3 Spheres as Surfaces

The most familiar example of a surface (other than an open set in $\mathbb{R}^2$) is a sphere $\mathbb{S}^2$, since we live on one. If we look around a bit at the surface of our planet, we might be inclined to suspect that Earth is flat, because it appears flat when we can only see a bit of it at a time.

Let us now start a rigorous proof that a sphere is a surface according to our definition. To do so, we need to show that for every point $p \in \mathbb{S}^2$, there is an open set U of $\mathbb{S}^2$ containing p, and a homeomorphism $f : U \to V \subset \mathbb{R}^2$. Thus we must first choose an appropriate open set U for each point p, and then construct the required homeomorphism. Note that the latitude and longitude coordinates we introduced above do not yet suffice. There are two reasons: the first is that they are only well-defined on *part* of $\mathbb{S}^2$, so we would only be able to prove that this part of $\mathbb{S}^2$ is a surface rather than all of $\mathbb{S}^2$; the second is that we have not defined f, nor shown the existence of f^{-1}, for these coordinates yet. We'll leave both of these issues for you to ponder on your own, and we will presently prove that $\mathbb{S}^2$ is a surface in a different way.

Proposition 2.2 *The unit sphere $\mathbb{S}^2$ in $\mathbb{R}^3$, which is defined as the set of points $\{(x, y, z) \in \mathbb{R}^3 : x^2 + y^2 + z^2 = 1\}$, is a surface.*

Proof Let us define six open sets in $\mathbb{S}^2$, namely U_{top}, U_{bottom}, U_{left}, U_{right}, U_{front}, and U_{back}. These are the six open hemispheres you would expect based on their names. For example $U_{\text{top}} = \{(x, y, z) \in \mathbb{S}^2 : z > 0\}$. The important points about these hemispheres are that each one is an open set in $\mathbb{S}^2$, and every point on the sphere is contained in at least one of them. Therefore if we can map each of these hemispheres bijectively to an open subset of $\mathbb{R}^2$ and show the existence of a continuous inverse, then we'll be done.

Let us first consider U_{top}. A very simple way of mapping this hemisphere to a piece of the plane is by *projection*. Namely, we use the map $f_{\text{top}} : U_{\text{top}} \to \mathbb{R}^2$ defined by $f_{\text{top}}(x, y, z) = (x, y)$. The range of f_{top} is the open ball $B_1(0) \subset \mathbb{R}^2$. Note that f_{top} is actually defined for all $(x, y, z) \in \mathbb{R}^3$, but we consider only the *restriction* of this map to $(x, y, z) \in U_{\text{top}}$. What about an inverse? The inverse should take a point $(x, y) \in B_1(0)$ and yield a point $f_{\text{top}}^{-1}(x, y) \in \mathbb{S}^2$. The point we want is clear: it is the point on $\mathbb{S}^2$ lying "above" $(x, y, 0)$, namely the point $(x, y, \sqrt{1 - x^2 - y^2})$. Thus we propose

$$f_{\text{top}}^{-1}(x, y) = (x, y, \sqrt{1 - x^2 - y^2}).$$

Note that both f_{top} and f_{top}^{-1} are continuous, as they are built from the reasonable functions of high-school calculus. Also we have $f_{\text{top}} \circ f_{\text{top}}^{-1}(x, y) = (x, y)$ and $f_{\text{top}}^{-1} \circ f_{\text{top}}(x, y, z) = (x, y, \sqrt{1 - x^2 - y^2}) = (x, y, z)$ for any $(x, y, z) \in U_{\text{top}}$, because the defining formula $x^2 + y^2 + z^2 = 1$ holds there. Thus f_{top} is a homeomorphism.

To complete the proof, we have to provide a similar analysis for the remaining five hemispheres of $\mathbb{S}^2$. For U_{bottom} we propose $f_{\text{bottom}}(x, y, z) = (x, y)$ once again, but

now we can check that $f_{\text{bottom}}^{-1}(x, y) = (x, y, -\sqrt{1 - x^2 - y^2})$ is the desired inverse function. For the remaining four hemispheres, we use similar ideas except using the "lateral projections" given by $(x, y, z) \mapsto (x, z)$ and $(x, y, z) \mapsto (y, z)$. The details are left to you! ■

2.4 Surfaces with Boundary

A natural question prompted by our consideration of hemispheres just now is: What is the nature of the *closed* hemisphere $\overline{U}_{\text{top}} := \{(x, y, z) \in \mathbb{S}^2 : z \geq 0\}$? Although this object is almost as surface-like as the familiar sphere $\mathbb{S}^2$, we are unfortunately not justified in calling it a surface—at least according to our definition. This is because any point on the *boundary* of the closed hemisphere, namely any point of the form $(x, y, 0) \in \overline{U}_{\text{top}}$, does not satisfy the surface property. For instance, we can form a relatively open set in $\overline{U}_{\text{top}}$ containing $(x, y, 0)$ by intersecting $\overline{U}_{\text{top}}$ with $B_r((x, y, 0))$. This open set is homeomorphic to a half-disk in $\mathbb{R}^2$ under the projection f_{top}, which is neither open nor closed. This is only one example, but it reflects a general phenomenon: Try as we might, we will never be able to map a relatively open set containing $(x, y, 0)$ to an open set in the plane, because the image of $\overline{U}_{\text{top}}$ will always be on only one side of the image of the boundary of $\overline{U}_{\text{top}}$.

We would, however, like to include the closed hemisphere $\overline{U}_{\text{top}}$ in our list of allowed "surface-like" objects. Therefore we make a special definition that covers the case of the closed hemisphere and similar surfaces with boundary curves. We'll need the *standard two-dimensional closed half-space* defined by $\mathbb{H}^2 := \{(x, y) \in \mathbb{R}^2 : y \geq 0\}$. We denote its boundary by $\partial\mathbb{H}^2 = \{(x, 0) : x \in \mathbb{R}\}$.

Definition 2.3 A *surface with boundary* S is a non-empty topological space such that for every point $p \in S$, there is an open set $U \subset S$ containing p, and a homeomorphism $f : U \to V$ onto a relatively open subset $V \subset \mathbb{H}^2$.

This definition admits two kinds of points in S. There are those points for which the original definition of "surface" holds, namely the homeomorphism $f : U \to V$ is such that V is contained in the interior of $\mathbb{H}^2$ and is thus an ordinary open set in $\mathbb{R}^2$. And there are those points whose image under f lie on $\partial\mathbb{H}^2$.

Remark 2.4 A technical note: there will typically be several homeomorphisms mapping neighborhoods of S to $\mathbb{H}^2$. It is possible to show that if f_1 and f_2 are homeomorphisms from a neighborhood of $p \in S$ to $\mathbb{H}^2$ and $f_1(p) \in (\mathbb{H}^2)^\circ$, then also $f_2(p) \in (\mathbb{H}^2)^\circ$; and if $f_1(p) \in \partial\mathbb{H}^2$, then this is also true for $f_2(p)$. Therefore the characterization of p into one of two kinds is unambiguous. The full proof requires more machinery than we currently have available, but here is the idea: let U be an open neighborhood around some point $p \in S^\circ$, and let $f : U \to \mathbb{R}^2$ be a homeomorphism onto an open subset. By passing to a smaller neighborhood V if necessary, we can assume that $f(V)$ is a disk in $\mathbb{R}^2$. Now, look at $f(V \setminus \{p\})$. This is a disk with a

hole in it. On the other hand, if $q \in S \setminus S^\circ$, then we can assume that a small neighborhood W around q in S is homeomorphic to a half-disk. Then let $g : W \to \mathbb{H}^2$ be a homeomorphism onto an open subset, with $g(q) \in \partial\mathbb{H}^2$. Now, $g(W \setminus \{q\})$ does not have a hole in it. Based on theory we will develop later on, these two regions $f(V \setminus \{p\})$ and $g(W \setminus \{q\})$ are not homeomorphic. Thus $f(V)$ and $g(W)$ are not either.

Definition 2.5 Let S be a surface with boundary. The *boundary* of S is the set of points $p \in S$ such that there exists a homeomorphism $f : U \to V \subset \mathbb{H}^2$ with $f(p) \in \partial\mathbb{H}^2$. The boundary of S is denoted ∂S.

2.5 Closed, Bounded, and Compact Surfaces

Recall that a surface S is a topological space, which for now means that S is a subset of some Euclidean space $\mathbb{R}^n$. Therefore we can ask about the nature of S as a point set inside $\mathbb{R}^n$. It is possible for a surface or surface with boundary to be a closed set (e.g. the sphere or the closed hemisphere) or a set which is neither open nor closed (e.g. the open hemisphere). In fact, an elementary property of a surface with boundary is that $S \cap (\partial S)^c$, i.e. S with its boundary removed, is a surface according to the original definition. But this surface is not closed unless $\partial S = \varnothing$. (Can you prove this?) Finally, a surface can never be an open set unless $n = 2$. (Can you see why this should be true?)

There is one additional feature of surfaces viewed as point sets in $\mathbb{R}^n$ that we should highlight. This concerns their behavior at infinity. We will say a surface S is *bounded* if there exists a perhaps large ball that fully encloses S, i.e. there exists $R > 0$ such that $S \subset B_R(0)$. We will say that S is unbounded if this fails to hold.

A very important technical property possessed by many surfaces that we'll encounter is the combination of *closedness* and *boundedness*. This has a special name.

Definition 2.6 A surface, or surface with boundary, is called *compact* if it is both closed and bounded as a subset of Euclidean space.

Remark 2.7 This definition is not the best or most general definition of compactness. The best definition is given in Problem 9.

2.6 Equivalence Relations and Topological Equivalence

A subject that we will explore in the next few chapters is the topological classification of surfaces. Broadly speaking, this is the partitioning of all possible surfaces into distinct categories, where two surfaces in the same category are considered topologically the same. This kind of partitioning based on sameness from a certain point

of view is actually an instance of a very general idea in mathematics known as a *partition into equivalence classes*. This is a very abstract concept that we will briefly explain here, given it will arise in various guises throughout the book.

Definition 2.8 Let X be a set. An equivalence relation on X is a subset of the product space $\mathcal{R} \subset X \times X$ that satisfies three key properties. We will use the notation $x \sim y$ instead of the usual (x, y) for pairs in $\mathcal{R}$, and we read $x \sim y$ as "x is equivalent to y." The properties are:

(1) *Reflexivity.* $x \sim x$ for every $x \in X$.
(2) *Symmetry.* $x \sim y$ if and only if $y \sim x$ for all $x, y \in X$.
(3) *Transitivity.* If $x, y, z \in X$ with $x \sim y$ and $y \sim z$, then $x \sim z$.

Exercise 2.9 Why doesn't property (1) in Definition 2.8 follow from properties (2) and (3)?

Example Let X be the set of all triangles in the plane and let $T_1, T_2 \in X$. We define $T_1 \sim T_2$ whenever T_1 is similar to T_2 (i.e. both triangles have the same internal angles). Now for any $T \in X$ it is clearly the case that $T \sim T$. Also, if $T_1, T_2 \in X$ and $T_1 \sim T_2$, then by the symmetry of "equal angles" it is true that $T_2 \sim T_1$. Finally, if $T_1, T_2, T_3 \in X$ with $T_1 \sim T_2$ and $T_2 \sim T_3$, then by the transitivity of "equal angles" it is true that $T_1 \sim T_3$. Thus similarity is an equivalence relation on X.

Example Let X be the set of all topological spaces (yes, this set is quite huge!), and let $S_1, S_2 \in X$. We define $S_1 \sim S_2$ whenever there exists a homeomorphism $f : S_1 \to S_2$. Now for any $S \in X$ we always have the identity homeomorphism $\mathrm{id} : S \to S$ defined by $\mathrm{id}(x) = x$. Hence $S \sim S$. Also, if $S_1 \sim S_2$, then the homeomorphism $f : S_1 \to S_2$ can be inverted to yield another homeomorphism $f^{-1} : S_2 \to S_1$. Hence $S_2 \sim S_1$. Finally, if $S_1, S_2, S_3 \in X$ with $S_1 \sim S_2$ and $S_2 \sim S_3$, then the homeomorphisms $f_{12} : S_1 \to S_2$ and $f_{23} : S_2 \to S_3$ can be composed to yield another homeomorphism $f_{23} \circ f_{12} : S_1 \to S_3$. Hence $S_1 \sim S_3$. Therefore the existence of a homeomorphism is an equivalence relation on X.

The key feature of an equivalence relation is that we can *partition* the set upon which it is defined into disjoint subsets called *equivalence classes*.

Definition 2.10 Let $\sim$ be an equivalence relation on X and let $x \in X$. The equivalence class of x is the subset $[x] = \{y \in X : y \sim x\}$ of X. In other words, $[x]$ consists of all $y \in X$ related to x. We write $X/\sim$ for the set of equivalence classes modulo $\sim$.

Proposition 2.11 *Let $\sim$ be an equivalence relation on X. The equivalence classes created by $\sim$ share the following key properties:*

(1) Equivalence classes are non-empty subsets of X.
(2) If $[x]$ and $[y]$ are two equivalence classes, then either $[x] \cap [y] = \varnothing$ or $[x] = [y]$.
(3) The union of all equivalence classes is X.

Proof The proof of (1) is simply to observe $x \sim x$ implies it is always the case that $x \in [x]$. Item (3) follows from this, because any $x \in X$ lies in an equivalence class (namely the class $[x]$) and so lies in the union of all equivalence classes. Hence X is a subset of the union of all equivalence classes. It is trivially true that the union of all equivalence classes is a subset of X. Hence X equals the union of all equivalence classes.

The proof of (2) is slightly more involved. Let $[x]$ and $[y]$ be two equivalence classes and suppose $z \in [x] \cap [y]$ is a common element. (If we could not find such a z that would mean that $[x] \cap [y] = \varnothing$ so we'd be done.) Hence $z \sim x$ and $z \sim y$. By symmetry we have $x \sim z$. Thus by transitivity we have $x \sim y$. By symmetry again we have $y \sim x$. Using this, we can prove both the $[x] \subset [y]$ and $[y] \subset [x]$ (which is what it takes to show $[x] = [y]$). Here's how: first pick an arbitrary element $\bar{x} \in [x]$. Then $\bar{x} \sim x$, and because $x \sim y$, we have $\bar{x} \sim y$ by transitivity. Thus $\bar{x} \in [y]$. Since $\bar{x}$ was chosen arbitrarily, this implies that $[x] \subset [y]$. The reverse inclusion is similar. ■

By the key property (3) above, X is equal to the union of its equivalence classes. By key property (2) above, there are only a certain number of distinct equivalence classes, and these do not overlap. Let $[x_1], \ldots, [x_n]$ be the distinct equivalence classes. (In fact, there may not be finitely many or even countably many of them, so the labeling of these with 1 through n is a bit of a misnomer.) We say that $x_1, \ldots, x_n$ are *representatives* of the distinct equivalence classes. Note that representatives of equivalence classes are not unique, because if $y \sim x_1$ then $[x_1] = [y]$, and y serves just as well as x_1 as a representative of the class $[x_1]$.

Exercise 2.12 What are the equivalence classes of the equivalence relations given in the two examples above? What are the possible representatives?

2.7 Homeomorphic Spaces

The question posed so glibly in the previous section, about characterizing the equivalence classes of homeomorphic equivalence on the set of all topological spaces, is actually a very deep mathematical question. Such questions have motivated and continue to motivate the development of the field of topology and of mathematics in general! In fact, it is fair to say that much of algebraic topology is about ways to tackle this question.

Let us build some intuition for the notion of homeomorphic equivalence of topological spaces. The question we'd like to consider here is: For what kinds of spaces S_1 and S_2 can we find a homeomorphism $f : S_1 \to S_2$?

Example A sphere S is homeomorphic to a cube C, as shown in Figure 2.5. To see this, let us describe the homeomorphism. Suppose the sphere is $\{(x, y, z) \in \mathbb{R}^3 : x^2 + y^2 + z^2 = 1\}$, and the cube has vertices $\{(\pm 1, \pm 1, \pm 1)\}$. Let p be a point on the cube. Then $\frac{p}{d(p,0)}$ is a point on the sphere. Let $f : C \to S$ be defined by $f(p) = \frac{p}{d(p,0)}$.

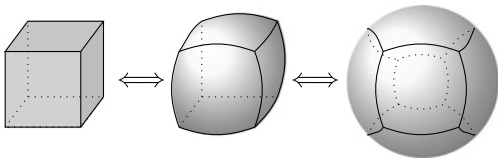


Figure 2.5 A sphere is homeomorphic to a cube.

Then f is a homeomorphism. (How would we go about writing down the inverse map?)

Nonexample The sphere, the plane, and the torus, all appear to live in different homeomorphism classes, although we do not yet have the tools to prove this.

Major Question 2.13 Given two topological spaces, how can we tell whether they are homeomorphic or not?

It is worth thinking a bit about the form that an answer to this question might have. Ideally, we would get an algorithm that runs for a while and then gives us an answer of "yes" or "no." Failing that, we could hope for an algorithm that runs for a while and then gives us an answer of "yes" or "no" or "I don't know." What we'll actually get is usually something resembling the latter type of answer, although in some cases we'll be able to get an answer of the former type.

2.8 Invariants

One tool that topologists have developed for determining whether or not two topological spaces are homeomorphic is called an *invariant*. An invariant is a function I on the set of all topological spaces of a certain type such that $I(S)$ is equal to a well-defined mathematical object for every appropriate topological space S. For instance I might be defined on the set of all compact surfaces, and $I(S)$ might be a natural number for each S. Moreover, this function must possess the following *invariance property*:

> Let S_1 and S_2 be two topological spaces in the domain of definition of the invariant I. If there exists a homeomorphism $f : S_1 \to S_2$ then $I(S_1) = I(S_2)$.

The existence of an invariant for a given type of topological space is very powerful because it allows us to tell spaces apart: if we find that $I(S_1) \neq I(S_2)$ then we can be sure that S_1 is not homeomorphic to S_2.

Invariants are very special objects, and some of the most important achievements in mathematics involve finding them—Fields Medals have been awarded for the discovery of new invariants! Of course, we're talking about non-trivial invariants here, because it is quite easy to come up with some useless invariants:

- $I(S) = 1$ for all topological spaces S is certainly invariant under homeomorphisms —but it isn't useful for telling anything apart from anything else!
- $I(S) = 1$ if S is homeomorphic to a sphere and $I(S) = 0$ otherwise is a perfectly good invariant for determining if S is homeomorphic to a sphere—except that you must already know whether S is homeomorphic to a sphere in order to compute $I(S)$!

A good invariant will be one for which it is much easier to compute $I(S)$ for a given S—or deduce properties of $I(S)$ depending on the nature of S—than it would be to prove the non-existence of homeomorphisms for such S using some ad hoc argument. Note that if we find $I(S_1) = I(S_2)$ we can *not* necessarily conclude that S_1 is homeomorphic to S_2. We call a *complete invariant* one for which it is true that $I(S_1) = I(S_2)$ if and only if S_1 is homeomorphic to S_2. Complete invariants are much rarer than ordinary invariants.

Example Here are some simple non-trivial invariants:

- $I(S) = 1$ if S is compact, and $I(S) = 0$ if it is not.
- $I(S) =$ the number of *connected components* of S. (A nonempty topological space T is said to be *connected* if it is not the union of two nonempty open subsets. A subset T of S is a said to be a *connected component* if T is connected, and every subset U of S properly containing T is not connected.)
- $I(S) =$ the number of components of ∂S.

The proof that the above "functions" are invariant under homeomorphisms is more or less straightforward. In the next chapter, we will study a much more interesting non-trivial invariant of surfaces called the *Euler characteristic*. It will take a bit of work to prove the invariance property for it! We will actually discover much more: It turns out that the Euler characteristic is very nearly a complete invariant of compact surfaces and can be used to solve the problem of characterizing the equivalence classes of homeomorphic equivalence in the set of all surfaces.

2.9 Problems

(1) (a) Show that the annulus $\{(x, y) \in \mathbb{R}^2 : 1 < x^2 + y^2 < 2\}$ is homeomorphic to the cylinder $\{(x, y, z) \in \mathbb{R}^3 : x^2 + y^2 = 1\}$.
(b) Show that the punctured sphere $\{(x, y, z) \in \mathbb{R}^3 : x^2 + y^2 + z^2 = 1, z \neq 1\}$ is homeomorphic to the plane.
(c) Do you think there exists a surjective continuous function from the torus to the sphere? If so, describe an example of such a function, in words.
(d) Do you think that there exists a surjective continuous function from the sphere to the torus? If so, describe an example of such a function, in words.
(e) Do you think that there exists an injective continuous function from the sphere to the torus? If so, describe an example of such a function, in words.

(f) Let γ_1 be a closed curve and γ_2 an open curve. (Here "closed" means the endpoints are glued together, like a loop, whereas "open" means that there are distinct endpoints (that are contained in the curve); note that these are different usages of the words "open" and "closed" from the ones we have seen before.) Do you think that there exists a continuous mapping from γ_1 to γ_2? What about from γ_2 to γ_1? Can they be surjective? Injective?

(2) Suppose you have four circles on a sphere for which the following hold:

- C_1, C_2 do not intersect and the centre of C_1 is "inside" the circle C_2, and
- C_1', C_2' also do not intersect, but the centre of C_1' is *not* "inside" the circle C_2'. (See Figure 2.6.) Is there a homeomorphism from the sphere to itself that maps C_1 to C_1' and C_2 to C_2'? If so, explain what the homeomorphism looks like. (You don't have to write down explicit equations.) If not, explain why there can be no such homeomorphism.

(3) Show that the cylinder $S = \{(x, y, z) \in \mathbb{R}^3 : x^2 + y^2 = 1\}$ is a surface.

(4) Either finish the proof started on Section 2.3 that the sphere is a surface, or give an alternative proof using a construction involving latitude and longitude.

(5) (a) Explain why the union of the xy- and xz-planes in $\mathbb{R}^3$ is not a surface with boundary.
(b) Show that the finite cylinder $S = \{(x, y, z) \in \mathbb{R}^3 : x^2 + y^2 = 1 \text{ and } z \in [-1, 1]\}$ is a surface with boundary.

(6) Explain why each of the following sets is *not* a compact surface. (A surface in $\mathbb{R}^3$ is *closed* if its complement is open. A surface in $\mathbb{R}^3$ is *bounded* if it can be enclosed in some large ball centered at the origin. A surface is *compact* if it is both closed and bounded.)

(a) A plane.
(b) The set of points in $\mathbb{R}^3$ with $z = x^2 + y^2$ and $z < 1$.
(c) The set of points in $\mathbb{R}^3$ with $z = x^2 + y^2$ and $z \leq 1$.
(d) The union of $\{(x, y, 0) : x^2 + y^2 < 1\}$ with $\{(0, y, z) : y^2 + z^2 < 1\}$.
(e) The union of the unit sphere centered at $(0, 0, 0)$ and the unit sphere centered at $(\frac{1}{2}, \frac{1}{2}, \frac{1}{2})$.

(7) In the definition of an equivalence relation, we could try to *deduce* reflexivity from symmetry and transitivity, as follows: pick $x, y \in X$ with $x \sim y$. Then $y \sim x$ as well. By transitivity, taking $z = x$, we have $x \sim y \sim x$, so $x \sim x$. Why doesn't this work?

(8) Stereographic projection of the sphere of radius 1 centered at $(0, 0, 0)$ in $\mathbb{R}^3$ into the plane $\mathbb{R}^2$ works as follows. Choose a point p on the sphere. Draw a line between the south pole and p. This line intersects the plane having $x_3 = 0$ in some point $(q_1, q_2, 0)$. Define $F(p) \equiv (q_1, q_2)$.

(a) What are the largest possible sets A, B we can put in the statement $F : A \to B$ such that F is surjective?
(b) Show that F is injective as a function $F : A \to B$.
(c) Write down a formula for $(q_1, q_2) \in B$ in terms of the coordinates of $p \in A$.

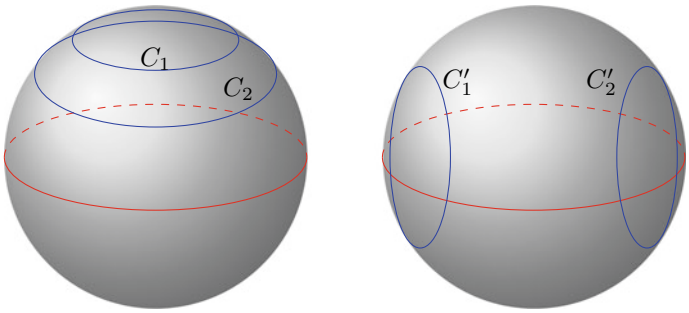


Figure 2.6 Left: C_1 and C_2. Right: C_1' and C_2'.

(d) Write down a formula for $F^{-1} : B \to A$.

(9) Let X be a compact set, and suppose $\{U_\alpha\}_{\alpha \in I}$ is a collection (not necessarily finite, or even countable) of open sets so that

$$X \subset \bigcup_{\alpha \in I} U_\alpha.$$

Show that we can find a *finite* subset $J \subset I$ so that

$$X \subset \bigcup_{\beta \in J} U_\beta.$$

(This is really the *definition* of compactness; the "definition" we gave in the text should actually be a theorem.)

Chapter 3
The Euler Characteristic and Identification Spaces

3.1 Triangulations and the Euler Characteristic

The goal of this chapter is to describe a useful homeomorphism invariant of surfaces known as the *Euler characteristic*. In order to do that, we need to discuss the notion of a *triangulation* of a surface.

Before we define a triangulation rigorously, let us explain intuitively what it is. The idea is to subdivide a surface S into patches, or *faces*, that are "triangularly shaped" and overlap in a controlled manner. Here is an example. Let S be the sphere $\mathbb{S}^2$. Of course there is no way to divide the smooth, round sphere into planar triangular patches. Instead, we use a homeomorphism as follows. Let $\mathcal{T}$ be the surface of a regular tetrahedron. Then $\mathcal{T}$ is homeomorphic to a sphere, and suppose $f : \mathcal{T} \to \mathbb{S}^2$ is such a homeomorphism. Now, $\mathcal{T}$ has four vertices, six edges, and four faces, and we can look at the images of these on the sphere. These divide the sphere up into four regions (the image of the faces), which are triangular in the sense that each face is delimited by three curves (the images of the edges) which intersect at the images of the three vertices. The homeomorphism f together with the regions and their boundaries created by the images of the faces, edges, and vertices is what we will call a *triangulation* of the sphere.

Remark 3.1 When we refer to a face in a polyhedron, we will always mean the *closed* face, so that it includes the edges and vertices on its boundary.

There are many ways to triangulate a surface. For example, another triangulation of the sphere is provided by the surface of a regular octahedron $\mathcal{O}$ (see Figure 3.1), because we can again take a homeomorphism $f : \mathcal{O} \to \mathbb{S}^2$ and map to the sphere its faces, edges, and vertices. Some common properties shared by this triangulation and the previous triangulation are:

- The image of every face is bounded by the images of exactly three edges and is homeomorphic to a disk.
- The intersection of any two faces is empty, a single vertex, or a single edge.

C. Bray et al., *Algebraic Topology*,
https://doi.org/10.1007/978-3-030-70608-1_3

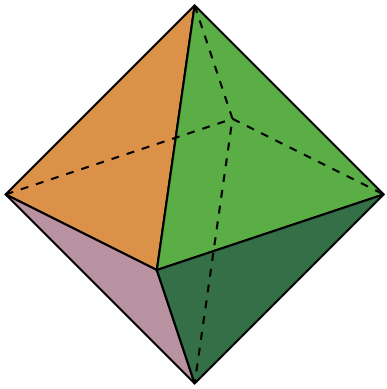


Figure 3.1 An octahedron.

- The intersection of any two edges is empty or a single vertex.

We will define a *general* triangulation of a surface to be the image, under a homeomorphism, of a domain that admits a subdivision into triangular faces such that these three properties hold. To be precise:

Definition 3.2 Let S be a surface, let P be a polyhedron, and let $f : P \to S$ be a homeomorphism. We say that f is a *triangulation* if it satisfies the following properties:

- Every face of P has exactly three edges and is homeomorphic to a closed disk.
- The intersection of any two faces of P is empty, a single vertex, or a single edge.
- The intersection of any two edges of P is empty or a single vertex.

Remark 3.3 Note that the choice of homeomorphism f is not used in the definition of a triangulation. That is, if $f : P \to S$ is some triangulation, and $g : P \to S$ is any other homeomorphism, then g is also a triangulation. Thus it is the subdivision of S into "triangularly shaped" patches with the correct intersection properties that counts, and not the shape of the patches.

Remark 3.4 In practice, we will tend to think of a triangulation by drawing edges on a surface S with the correct intersection properties, rather than mapping a polyhedron to S.

We have seen a few examples of triangulations of spheres so far, so now let us look at some nonexamples of triangulations to see what can go wrong.

Nonexample A homeomorphism f from a cube to a sphere is not a triangulation, because the faces of a cube are bounded by four edges rather than three. We can fix this problem by dividing each face of the cube (each square) into two triangles by drawing a diagonal edge. The resulting figure is then a triangulation.

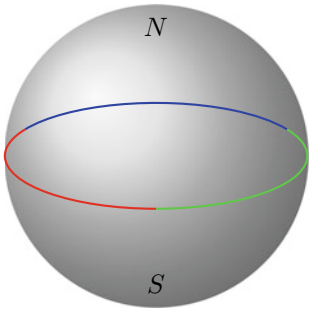


Figure 3.2 Dividing the sphere into two triangles (N and S) like this is not a triangulation, because the triangles intersect at all three edges.

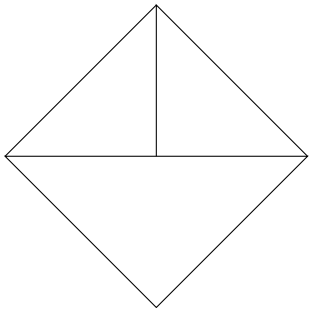


Figure 3.3 This can never be a piece of a triangulation.

Nonexample If we break the sphere into two faces, the northern and southern hemispheres, and we break the equator up into three edges, the resulting figure is not a triangulation: the two faces intersect at three edges, which is not allowed. See Figure 3.2.

Nonexample Consider the subdivision of the torus into two regions: one is a small triangle, and other is the rest of the torus. This is not a triangulation, for two reasons. First, the two regions intersect at three edges, which is not allowed. Second, the large region is not homeomorphic to a disk (or at least, it does not appear to be).

Nonexample Two faces cannot meet in only part of an edge. So, the configuration shown in Figure 3.3 cannot be part of a triangulation.

We would like to use triangulations in order to determine a homeomorphism invariant for surfaces. But the problem is that surfaces can have many different triangulations. Thus, we want to find a number that we can calculate from a triangulation that still somehow doesn't depend on the choice of triangulation.

Let us look carefully at a few different triangulations of the sphere. The easiest information we can get out of a triangulation is the number of vertices, edges, and faces it has, so let us tabulate those.

- The tetrahedron is a triangulation of the sphere, and it has 4 vertices, 6 edges, and 4 faces.

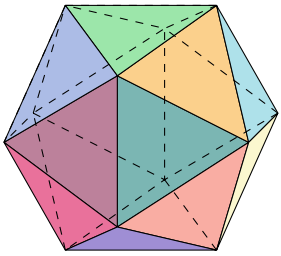


Figure 3.4 An icosahedron.

Table 3.1 Vertices, edges, and faces of various triangulated polyhedra.

Polyhedron	Vertices	Edges	Faces
Tetrahedron	4	6	4
Octahedron	6	12	8
Cube	8	18	12
Icosahedron	12	30	20

- The octahedron has 6 vertices, 12 edges, and 8 faces.
- The cube is not a triangulation, but we can modify it to become one. The cube has 8 vertices, 12 edges, and 6 faces. However, in order to make it into a triangulation, we need to draw a new edge on each face, thereby dividing each face into two new faces. As a result of this operation, we increase the edge count and face count each by 6, yielding 8 vertices, 18 edges, and 12 faces.
- The icosahedron (a regular 20-sided polyhedron; see Figure 3.4) has 12 vertices, 30 edges, and 20 faces. It is a triangulation of the sphere.

We collect our data in Table 3.1. A pattern pops out: the number of vertices (V) minus the number of edges (E) plus the number of faces (F) is always equal to 2. Or, more succinctly,

$$V - E + F = 2.$$

This formula is one of Euler's most famous results.

We can compute a similar quantity for any compact surface: we first triangulate the surface, count the number of vertices, edges, and faces, and then compute $V - E + F$.

Definition 3.5 Let S be a compact surface, and let T be a triangulation of S. Let V, E, and F denote the number of vertices, edges, and faces, respectively, of T. We call the quantity $V - E + F$ the *Euler characteristic* of S (with respect to T), and we write

$$\chi_T = V - E + F.$$

Note that, at the moment, χ_T might depend on the choice of triangulation of T. Our goal in the next section will be to show that this is not the case: χ_T is the same for all triangulations T of a fixed compact surface S.

3.2 Invariance of the Euler Characteristic

In this section, we will prove the following theorem:

Theorem 3.6 *Let T and T' be two triangulations of a compact surface S. Then $\chi_T = \chi_{T'}$.*

Proof Our strategy for proving Theorem 3.6 will be to take two triangulations T and T' of S and produce a new triangulation T'', which is a *refinement* of both T and T'. We will then show that χ_T and $\chi_{T'}$ both equal $\chi_{T''}$ which of course implies the truth of the theorem.

When we say that T'' is a refinement of T, this is what we mean:

Definition 3.7 Let T_1 and T_2 be triangulations of S. We say that T_2 is a *refinement* of T_1 if every face of T_1 is a union of faces of T_2.

In order to start the proof of Theorem 3.6, we need to show that any two triangulations T and T' of S share a common refinement T''. To do this, superimpose T and T' on S. The result will not in general be a triangulation, perhaps for several reasons. First, two edges can intersect at points other than vertices. When this happens, we add new vertices at the intersection points. Furthermore, the faces might not be triangles: they might be polygons with more than three sides. When this happens, we draw more edges to break these faces into several smaller ones, in just the way we did to turn the cube into a triangulation of the sphere. The result of these operations is the common refinement.

We will now analyze the common refinement to show that χ_T and $\chi_{T'}$ both equal $\chi_{T''}$. The hope is that any refinement of T or T' can be built up using a succession of steps in which we add one vertex at a time. In this way we obtain a sequence of triangulations $T = T_0, T_1, T_2, \ldots, T_N = T''$ where each T_k is a refinement of T_{k-1} and contains exactly one additional vertex, as well as a certain number of additional edges connecting this vertex to the original vertices and a certain number of additional faces resulting from subdividing the original faces.[1] There are in fact only two possibilities for where a vertex v can go, and what it can be connected to:

(1) v can be in the interior of a face f of T. In this case, we must draw edges connecting v to each vertex of f in order to obtain a new triangulation whose vertices consist of the vertices of T together with v. See Figure 3.5 for a picture.

[1] This isn't exactly true, but it *is* true that we can find a *further* refinement of T''' of T'' such that T''' can be obtained this way, even though T'' might not be one of the intermediate triangulations. We will not dwell further on this point, but see if you can convince yourself of this fact.

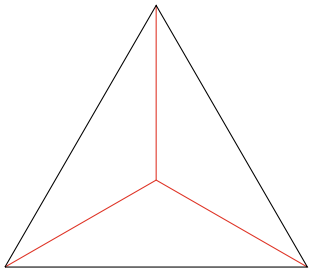


Figure 3.5 Here we have added a new vertex in the middle of a face f and connected it (the red edges) to all the vertices of f.

(2) v can be on an edge e of T. If T is a triangulation of a compact surface S, then e is an edge of exactly two faces f_1 and f_2. Now, f_1 and f_2 each have three vertices, two of which are the vertices of e, so they each have one additional vertex, say v_1 and v_2. Furthermore, v_1 and v_2 must be distinct, or else f_1 and f_2 would have an illegal intersection type. In order to create a triangulation whose vertices are the vertices of T together with v, we must draw edges connecting v to v_1 and to v_2, where these edges must lie inside f_1 and f_2, respectively. See Figure 3.6 for a picture.

Our final task will be to show that when we add a vertex to a triangulation, the Euler characteristic of the triangulation is unchanged. Therefore we will have $\chi_{T_k} = \chi_{T_{k-1}}$ for every k, and therefore $\chi_T = \chi_{T''}$ as desired. To see why this is so, we must analyze both of the scenarios above and count the number of vertices and faces before and after adding the vertex. So let us suppose that T_{k-1} has V vertices, E edges, and F faces. Let T_k be the triangulation obtained by adding a new vertex v.

(1) If v is in the interior of a face f of T_{k-1}, then T_k has $V + 1$ vertices. The edges are the same as the edges of T_{k-1}, together with the three new edges connecting v to the vertices of f, so T_k has $E + 3$ edges. The faces of T_k are the same as the faces of T_{k-1}, except that we have replaced f with three smaller faces. The effect of this is that we have lost one face of T_{k-1} and introduced three new ones, so the number of faces of T_k is $F + 2$. Hence

$$\chi_{T_k} = (V + 1) - (E + 3) + (F + 2) = V - E + F = \chi_{T_{k-1}}.$$

(2) If v is on an edge e of T_{k-1}, then T_k has $V + 1$ vertices again. The edges are now the same as the edges of T_{k-1}, except that we have removed e and replaced it with two smaller edges, and we also introduced the two new edges connecting v to v_1 and v_2 as above. Hence T_k has $E + 3$ edges again. Finally, both f_1 and f_2 have been split into two smaller faces, so T_k has $F + 2$ faces. Once again, we see that $\chi_{T_k} = \chi_{T_{k-1}}$.

■

Remark 3.8 There are several technical steps we are skipping in this last argument about common refinements. For example, it could be the case that some pair of edges

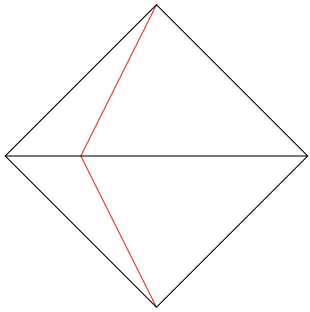


Figure 3.6 Here we have added a new vertex in the middle of an edge e and connected it (the red edges) to the other vertices of the neighboring faces.

intersect infinitely many times but are still not the same edge. In order to fix this, we will want to deform one of the triangulations a little bit so that this no longer happens. Try to convince yourself that this can be done!

Since we now know that the Euler characteristic of a surface S does not depend on the choice of triangulation, we can write $\chi(S)$ for the Euler characteristic of S.

3.3 Identification Spaces

Now that we have defined the Euler characteristic and seen that it is an invariant of a surface, we would like to be able to calculate it for different types of surfaces. Of course, we know already that $\chi(\mathbb{S}^2) = 2$. At the moment, it isn't very easy to go beyond this, because it is quite hard to visualize a triangulation of, say, a torus and count the vertices, edges, and faces. (But try and see if you can do it!)

In order to address this limitation, we will develop an easier way to work with surfaces—essentially by drawing them in the plane! To illustrate what we have in mind, let us use the torus as an example. The torus is a surface, so we know that every point of the torus has a small neighborhood that is homeomorphic to part of the plane (the definition of a surface!). But how do we get the *whole* torus to be part of the plane? We can do this by cheating a bit—but the "cheat" we'll use will actually end up being a rigorous mathematical operation. First, cut the torus along a circle, as in Figure 3.7. Once we have made this cut, we stretch out the cut torus into a cylinder. Then we make a cut along a line connecting the top and bottom of the cylinder and unroll it. The result is a rectangle, as we can see from looking at Figure 3.8.

We can go the other way too. If we start with a rectangle, we can glue one pair of opposite sides together to create a cylinder, and then we can glue the top and bottom of the cylinder to create a torus. In other words, we get a torus by gluing pairs of opposite sides of a rectangle.

Remark 3.9 We were a little bit sloppy when we talked about gluing opposite sides together, because we could glue them in either orientation. When we create a torus, we glue the top edge and the bottom edge in such a way that the left side of the top

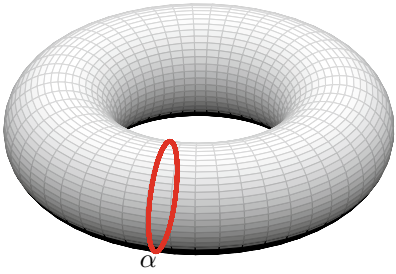


Figure 3.7 If we cut along the curve α, the red loop, we can stretch the resulting figure into a cylinder.

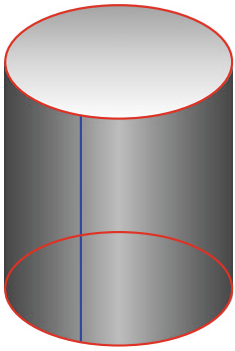


Figure 3.8 If we cut along the blue line, we can unroll the resulting figure into a rectangle.

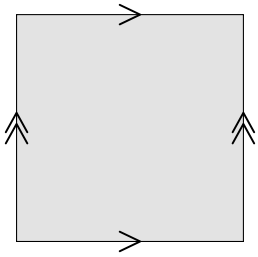


Figure 3.9 A rectangle as an identification space for a torus.

edge is glued to the left side of the bottom edge, but we could have switched it and glued the left side of the top edge to the right side of the bottom edge instead. This will be very important in the near future. In order to be unambiguous about which way to glue, we will draw arrows on the sides to specify the gluing direction, as shown in Figure 3.9.

The object we have created—a rectangle with edges that are meant to be glued together—is the planar representation of the torus that we had in mind. It is an example of an *identification space* which can be defined in general as follows.

Definition 3.10 An *identification space* (or *ID space*) is a polygon in $\mathbb{R}^2$ along with instructions for gluing edges together.

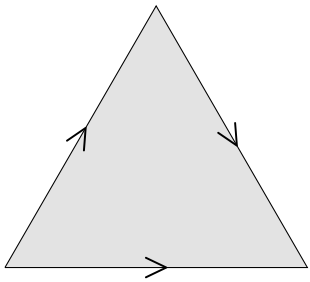


Figure 3.10 An identification space aaa^{-1}, sometimes called the *dunce cap*.

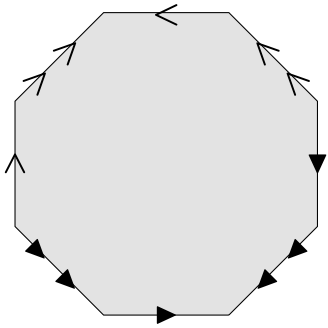


Figure 3.11 Another interesting identification space.

Exercise 3.11 Consider the ID space shown in Figure 3.10. For what sort of figure is this an ID space? Is it a surface?

Exercise 3.12 Consider the ID space shown in Figure 3.11. For what sort of figure is this an ID space? Is it a surface? Cutting up pieces of paper is encouraged!

Let us introduce some notation that will make it easier to describe ID spaces. Suppose we have an ID space whose underlying space is a polygon P. Each identified edge will be labeled by a different lowercase letter. Pick some vertex $v \in P$, and travel around the boundary of the polygon (in either direction), starting from v. If we travel along an edge labeled "e" (say) in the direction indicated by the arrow, then we write down an e; if we travel in the wrong direction, then we instead write e^{-1}. Continue in this manner until we get back to v.

Example The ID space for a torus can be written as $aba^{-1}b^{-1}$, and the ID space in Exercise 3.12 can be written as $aba^{-1}b^{-1}cdc^{-1}d^{-1}$.

3.4 ID Spaces as Surfaces

Two important questions that we must address are: (1) given a compact surface S, can we always obtain an ID space representation for it; (2) supposing we have an

ID space as defined in Definition 3.10, how can we tell whether it is an ID space for a surface or a surface with boundary? The answer of the first question is yes—this involves systematically cutting S into triangular faces and assembling these into an ID space. We'll see how this is done in the next chapter. The second question is interesting to ponder, because we have just seen some examples of ID spaces that are not surfaces amongst the examples above.

Here's another reason why this is an interesting question. Consider the ID space for the torus that we constructed earlier, namely the square with opposite sides glued together as shown in Figure 3.9. Call it S. We can show that only *part* of the definition of a surface is satisfied. That is, we can show that for every $p \in S$ there is an open set U containing p that can be mapped homeomorphically to an open set in the plane. To see this, consider the following three cases.

(1) If p belongs to the interior of S, then the condition is trivially satisfied.
(2) If p belongs to an edge of S but is not a corner of S, then we *define* an open neighborhood of p to be the union of the open half-disk containing p and the open half-disk containing the point p' on the opposite edge that is meant to be glued to p. Note that, as far as the topology of the torus is concerned, this union of two open half-disks is identical to the open disk from Case (1) because of the gluing instructions that come with S. Thus we can also easily map the glued union of open half-disks to the plane.
(3) If p is a corner of S, then we *define* an open neighborhood of p to be the union of four open quarter-disks at the four corners of S. Note that, as far as the topology of the torus is concerned, this union of open quarter-disks is identical to the open disk from Case (1) because of the gluing instructions that come with S. Thus we can also easily map the glued union of open quarter-disks to the plane.

Therefore we can show that every point in S has an open neighborhood that can be mapped to the plane, provided we are allowed to decide what "open" means. But we have not exhibited S as a set of points in $\mathbb{R}^3$ (or any other $\mathbb{R}^n$), nor have we shown that our open sets are of the form $\mathcal{O} \cap S$ where $\mathcal{O}$ is an open set in $\mathbb{R}^3$.

Well, the fact is that we cannot do this. Thus S is not a surface as we have defined it in Chapter 1. However, S is an example of an *abstract surface*. This is a "two-dimensional" topological space that exists in its own right, without reference to an ambient space such as $\mathbb{R}^3$. To define it, we need a more general notion of a topological space than merely a subset of $\mathbb{R}^n$, which was our earlier preliminary definition of a topological space.

3.5 Abstract Topological Spaces

Roughly speaking, a topological space is a set together with some notion of what it means for a subset to be considered open. The precise definition is as follows:

Definition 3.13 A *topological space* is a set S together with a collection $\mathscr{T}$ of subsets of S (called the open sets in S) so that

- $\varnothing, S \in \mathscr{T}$.
- If A is any set and $\{S_\alpha\}_{\alpha \in A}$ is a collection of subsets of S indexed by A, so that each $S_\alpha \in \mathscr{T}$, then $\bigcup_{\alpha \in A} S_\alpha \in \mathscr{T}$.
- If $S_1, S_2, \ldots, S_n \in \mathscr{T}$, then $\bigcap_{i=1}^{n} S_i \in \mathscr{T}$.

We call $\mathscr{T}$ a *topology* on S.

Remark 3.14 It is possible to put many different topologies on a set S. In particular, if $S = \mathbb{R}^n$, then the open sets of some exotic topology need not satisfy the ball property that we discussed in Chapter 1.

When we translate these into statements about open sets, these properties are saying the following:

- The empty set and all of S are open sets.
- The union of any collection of open sets is open.
- The intersection of a *finite* collection of open sets is open.

These are the properties of open sets that we proved in Proposition 1.5. Hence, $\mathbb{R}^n$ with the usual definition of open sets is a topological space, as is any subset of $\mathbb{R}^n$ with the notion of (relatively) open sets we discussed in Chapter 1. However, there are many other topological spaces out there, which do not live inside of $\mathbb{R}^n$, some of which are of paramount importance in other areas of mathematics.

Example The most important topology in algebraic geometry is the *Zariski topology*. We describe only a simple case here, that of the Zariski topology on the set $\mathbb{C}$ of complex numbers. We define the open sets to be the empty set $\varnothing$, and the complements of finite sets. Let us verify that this is a topology. We must check first that $\varnothing$ and $\mathbb{C}$ are open. $\varnothing$ is open because we said it is, and $\mathbb{C}$ is open because its complement is $\varnothing$, which is finite. Now we must check that unions of open sets are open sets. Let us suppose $\{U_\alpha\}_{\alpha \in A}$ is a collection of open sets. We now have two cases. If all the U_α's are $\varnothing$, then so is their union, which is open. If at least one U_α is nonempty, then it is the complement of a finite set. Taking the union with all the rest of the U_α's can only make the complement smaller, so the union is still the complement of a finite set. Finally, we must check that if $U_1, \ldots, U_n$ are open, then so is their intersection. If some $U_i = \varnothing$, then so is their intersection. Otherwise, suppose that their complements contain $a_1, \ldots, a_n$ elements. Then the complement of $U_1 \cap \cdots \cap U_n$ contains at most $a_1 + a_2 + \cdots + a_n$ elements, which is still finite. Hence the Zariski topology on $\mathbb{C}$ is a topology. The reason that the Zariski topology is important in algebraic geometry is that its closed sets are exactly the roots of polynomials.

Example Let p be a prime. Then we can put an interesting new topology, based on divisibility by p, on the set $\mathbb{Z}$ of integers. We define open balls first, and then declare the open sets to be exactly the unions of open balls. For an integer n and an integer $r \geq 0$, we define $A_{n,r}$ to be $\{m \in \mathbb{Z} : m - n \text{ is divisible by } p^r\}$. Note that the balls

get *smaller* as r increases; as a result, we could consider writing $1/r$ in place of r. Check that unions of these balls define a topology on $\mathbb{Z}$, called the *p-adic topology*. It is the starting point for the delightful p-adic numbers $\mathbb{Z}_p$—and its friends—and is the right setting for much of number theory; see [Gou97] for a good starting point on the subject.

For an abstract *surface*, there are also several other technical conditions that we must impose to make them into things that *could* be embedded into some $\mathbb{R}^n$. We relegate discussion of those to Appendix A.

3.6 The Quotient Topology

One particularly important type of topology is called the *quotient topology*, which is just the thing we need for ID spaces.

Definition 3.15 Let S be a topological space with topology $\mathscr{T}$, and let $\sim$ be an equivalence relation on S. We define a topology, called the *quotient topology*, on the set $S/\sim$ of equivalence classes modulo $\sim$ as follows: let $p : S \to S/\sim$ be the map that takes an element of S to its equivalence class modulo $\sim$. Then we define a set $U \subset S/\sim$ to be open if $p^{-1}(U) \in \mathscr{T}$, i.e. if $p^{-1}(U)$ is an open set of S.

This is relevant for ID spaces, because they are defined to be sets of equivalence classes. For example, we can view the torus, in its ID space form, as being the square $[0, 1] \times [0, 1]$ of its ID space, modulo the equivalence relation that points on the left side are equivalent to points on the right side, and similarly with top and bottom sides. To set this up as an equivalence relation, we declare that $(a, 0) \sim (a, 1)$ and $(0, b) \sim (1, b)$. The only other equivalences we allow are the trivial ones $(a, b) \sim (a, b)$. For (a, b) in the interior $(0, 1) \times (0, 1)$ of the square, a small disk in $(0, 1) \times (0, 1)$ centered at (a, b) is an open neighborhood of (a, b). But for (a, b) on an edge or vertex of a square, a neighborhood looks a little different, as shown (in the case of an edge) in Figure 3.12.

Similarly, we can view a circle as a quotient of the closed interval $[0, 1]$, by saying that $0 \sim 1$, and the only other equivalences are $a \sim a$ for $a \in [0, 1]$.

We now look at the quotient topology for ID spaces and surfaces. An open set in an ID space is just what you expect it to be: a set U is open if, for any point x in the ID space, U contains a small open disk around x. However, if the open set hits an

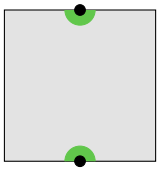


Figure 3.12 A neighborhood (shown in green) of a point on the edge of the square, under the quotient topology that turns it into a torus.

edge of the ID space, then it must continue on, out the other side: the topology does not "know" that is being drawn as an ID space. (It only "knows" about the open sets, not about how it's being drawn on the page.)

Exercise 3.16 Given an ID space, how can you tell whether it is an ID space for a surface or a surface with boundary?

3.7 Further Examples of ID Spaces

Let us look at the ID spaces for several other surfaces, starting with the sphere. The ID space represented by $abb^{-1}a^{-1}$ is a sphere, as shown in Figure 3.13.

What else can we do with a square? One possibility is to create a surface with boundary by gluing two opposite edges, so that we obtain a cylinder. A more exciting possibility is to glue two opposite edges with a twist: first we glue the left and right edges, and then we glue the top part of the left edge to the bottom part of the right edge. In doing so, we obtain a surface with boundary called a *Möbius strip* (see Figure 3.14). We can represent this as an ID space as $abac$.

Exercise 3.17 If you haven't played with Möbius strips before, do so! For example, how many boundary curves are there? What happens when you start drawing a path through the middle of the strip? What happens when you cut along this path? Can you figure out more cool properties of the Möbius strip?

We can obtain two more surfaces without boundary from a square as follows. The first is called the Klein bottle and is denoted by $\mathbb{K}$. We can get our hands on this in one of two ways: either we take a cylinder and identify the boundary circles in

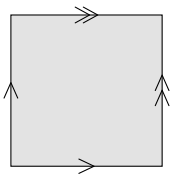


Figure 3.13 An ID space for a sphere.

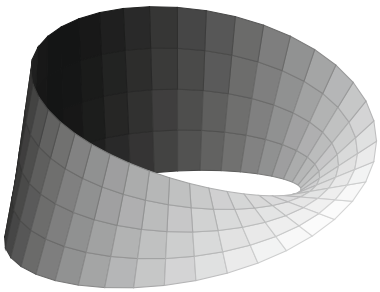


Figure 3.14 A Möbius strip.

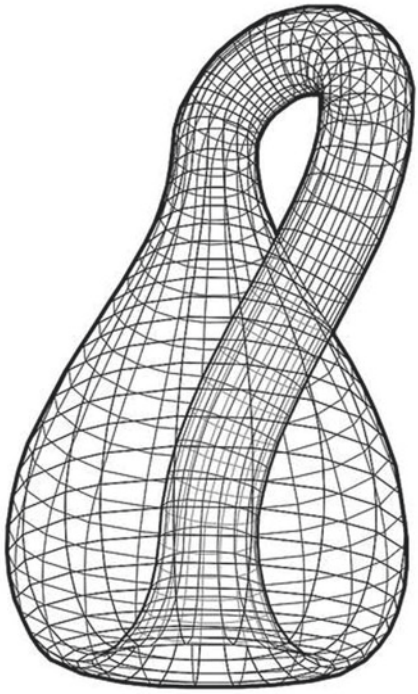


Figure 3.15 Here is a representation of a Klein bottle inside $\mathbb{R}^3$, with some self-intersections. It can be embedded in $\mathbb{R}^4$ without any self-intersections.

opposite directions (rather than in the same direction—which would give us a torus); or we can take a Möbius strip and identify pairs of points on the boundary curve. (How exactly? Exercise! Hint: try to see what is going on with a piece of paper.) Either way, the resulting ID space can be written as $abab^{-1}$. This space cannot be embedded into $\mathbb{R}^3$ without self-intersections, although it can in $\mathbb{R}^4$. See Figure 3.15 for a picture in $\mathbb{R}^3$ with some self-intersections. One can buy glass Klein bottles like it online from http://www.kleinbottle.com/.

The second additional surface without boundary obtainable from the square is called the *projective plane* and is denoted by $\mathbb{RP}^2$, for reasons to be explained in the next chapter. $\mathbb{RP}^2$ can be written in ID space notation as $abab$, although it also has an even simpler ID space description as aa. This space cannot be embedded into $\mathbb{R}^3$ without self-intersections, although it can in $\mathbb{R}^4$. See Figure 3.16 for a picture with some self-intersections. This figure is known as Boy's surface.

Exercise 3.18 Convince yourself that the two ID space descriptions $abab$ and aa of $\mathbb{RP}^2$ are homeomorphic.

Why is Boy's surface a representation of $\mathbb{RP}^2$? Start with the ID space $abab$ for $\mathbb{RP}^2$, and split b into two edges, which we'll call b and c. Thus $\mathbb{RP}^2$ has a hexagonal ID space labeled $abcabc$, as shown in Figure 3.17. We now bring the a edges together, rotating one of them in the process. That creates one of the bulbs in Boy's surface. We

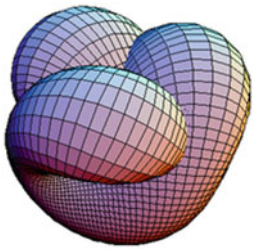


Figure 3.16 Here is a representation of $\mathbb{RP}^2$ inside $\mathbb{R}^3$, with some self-intersections. It can be embedded in $\mathbb{R}^4$ without any self-intersections.

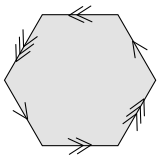


Figure 3.17 A hexagonal ID space for $\mathbb{RP}^2$.

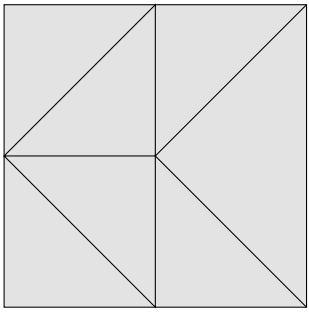


Figure 3.18 Suppose opposite sides are identified, so that this figure is an ID space for a torus. This is not a triangulation, because the vertex in the middle of the left side should be identified with a vertex in the middle of the right side.

get the other two bulbs by bringing together the two b edges and the two c edges. The fact that there are three pairs of edges to glue gives the surface a threefold symmetry. See https://www.mathcurve.com/surfaces.gb/boy/boy.shtml for an animation of the process of folding the hexagon into Boy's surface.

Exercise 3.19 Try playing chess on the square ID spaces of a torus, Möbius strip, projective plane, sphere, and Klein bottle.

3.8 Triangulations of ID Spaces

One nice thing we can do with ID spaces is to draw triangulations on them. In order to do that, we just subdivide the polygon in question into triangles. When we then glue the polygon's edges together, we obtain a triangulation on the resulting surface. However, it is possible for things to go wrong with this approach if we apply it naïvely. For example it might happen that the intersection of two faces is not of the desired type. (See Figure 3.18 for an example of something that can go wrong.)

Exercise 3.20 Draw an allowable triangulation on an ID space for a torus. What is its Euler characteristic?

Exercise 3.21 Draw an illegal "triangulation" on an ID space for a torus (i.e. one where we have the wrong type of face intersections). Count the vertices, edges, and faces of this "triangulation." What is the "Euler characteristic" for this triangulation? Is it the same as the actual Euler characteristic of a torus? Can you explain what is going on?

3.9 The Connected Sum

One useful operation when studying surfaces and other topological spaces is the *connected sum*.

Definition 3.22 Let S_1 and S_2 be two surfaces. We define their *connected sum* $S_1 \# S_2$ as follows: Choose open disks U_1 and U_2 inside S_1 and S_2, respectively, and let T_1 and T_2 denote $S_1 \setminus U_1$ and $S_2 \setminus U_2$, respectively. Now, glue T_1 and T_2 together by identifying the boundaries of U_1 and U_2. The resulting surface is the connected sum.

Exercise 3.23 Show that, up to homeomorphism, $S_1 \# S_2$ is independent of the choices of open disks U_1 and U_2, and also of the orientations of the boundaries along which we glue.

Example A two-holed torus is the connected sum of two ordinary tori. Similarly, a three-holed torus is the connected sum of a torus and a two-holed torus, and so forth.

Exercise 3.24 If S is a surface, what is $S \# \mathbb{S}^2$?

Remark 3.25 We tend to think of the connected sum operation as being analogous to multiplication for integers. By the previous exercise, we know that doing a connected sum with a sphere (spoiler alert!) doesn't change the surface, so the sphere acts like 1. Furthermore, we have a notion of primes for surfaces: a surface S is prime if whenever we have $S = S_1 \# S_2$, then one of S_1 and S_2 is a sphere (and also S is not a sphere). This is the start of a very deep connection between topology and number theory; see [Mor12] for a book-length treatment on one aspect of this fascinating connection.

Remark 3.26 We call a g-holed torus a *surface of genus* g.

Exercise 3.27 Let S and T be two surfaces. Can you express $\chi(S \# T)$ in terms of $\chi(S)$ and $\chi(T)$? What is the Euler characteristic of a surface of genus g?

It is also useful to understand how connected sums work in terms of ID spaces. Let us suppose we have ID spaces for two compact surfaces S and T, and let us suppose that they are represented by polygons P and Q, respectively, with some

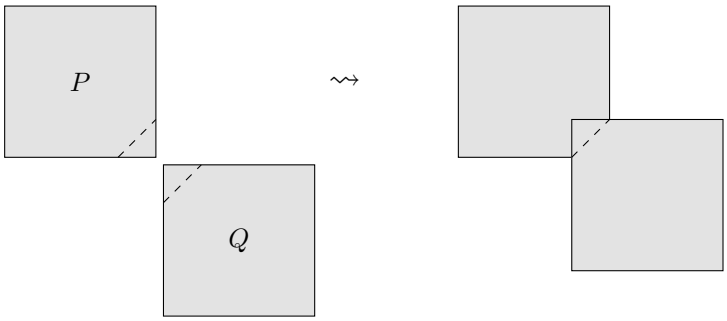


Figure 3.19 The connected sum of two ID spaces.

edge identifications. Here is one way of drawing an ID space for $S\#T$: Pick vertices v_P and v_Q of P and Q, respectively. Then v_P and v_Q are each adjacent to two edges of their respective polygons. Pick points on each of those adjacent edges, which are close to v_P or v_Q, and draw an edge between those two points. Then remove the new triangle formed. We now have two new edges, one on each of the mutilated original polygons, which are not identified with any other edges. So, identify them with each other. (See Figure 3.19.)

Remark 3.28 Let e be an edge adjacent to v_P, and let e' be the edge in P that is identified with e. After we cut out the triangle around P, we have to change the identification between the new e and e', since e is now a little bit shorter. The same thing holds for the other modified edges.

Exercise 3.29 Solve Exercise 3.27 using the ID space interpretation of connected sums.

Exercise 3.30 Is $\mathbb{RP}^2\#\mathbb{RP}^2$ a surface we have already studied? If so, which one is it?

3.10 The Euler Characteristic of a Compact Surface with Boundary

So far we have only looked at Euler characteristics for compact surfaces. However, we can also define the Euler characteristic for compact surfaces with boundary. They are defined in the same way as before: If S is a surface with boundary, we can show that S is homeomorphic to a compact surface with a finite number of open disks removed. Using this, we can show that S has a triangulation consisting of a finite number of triangles. Thus we can count the numbers V, E, and F of vertices, edges, and faces, and define the Euler characteristic as $\chi(S) = V - E + F$. Furthermore, $\chi(S)$ remains independent of the choice of triangulation.

Exercise 3.31 What is the Euler characteristic of a closed disk in $\mathbb{R}^2$?

Exercise 3.32 Suppose a surface with boundary S is equal to a compact surface S' from which a finite number n of open disks have been removed. Express $\chi(S)$ in terms of $\chi(S')$ and n.

3.11 Problems

(1) Draw a triangulation of a torus, and compute the Euler characteristic. Use this computation to explain why the torus and the sphere cannot be homeomorphic.
(2) Suppose we drop the restriction in a triangulation that the intersection of two faces or edges can only be an edge, a vertex, or nothing, but we allow it to be perhaps a union of several of these. Draw an invalid triangulation of a torus that only fails to be valid in this respect, and compute the Euler characteristic now. What happens?
(3) Prove that if we divide a surface up into polygons that might not be triangles and try to compute the Euler characteristic by counting the vertices, edges, and faces, then we get the same result as we do by triangulating.
(4) (a) Suppose we perform our connected sum in two different ways. Let S_1 and S_2 be connected surfaces. Let A be the connected sum you get when you perform the procedure described in Section 3.9 by removing disks containing the points $p_1 \in S_1$ and $p_2 \in S_2$. Let A' be the connected sum you get when you perform the procedure above by removing disks containing completely different points $p_1' \in S_1$ and $p_2' \in S_2$. Argue that A is homeomorphic to A'.
(b) Prove that $\chi(S_1 \# S_2) = \chi(S_1) + \chi(S_2) - 2$.
(c) Let S be any surface. Show that $\chi(S) = \chi(S \# \mathbb{S}^2)$. Is this reasonable?
(d) Show that the "orientable surface of genus g" (the connected sum of g tori) has Euler characteristic equal to $2 - 2g$.
(5) (a) Explain how the ID-space $aba^{-1}b^{-1}cdc^{-1}d^{-1}$ physically becomes a surface that looks like two tori joined together, by drawing the results of performing the gluings in the following order:
(i) Glue together the a's and the c's.
(ii) Observe that the two remaining b's and two remaining d's are now closed loops.
(iii) Glue together the b's and the d's.
(b) Using a similar procedure, determine what sort of surface you get from the ID-space $aba^{-1}dcb^{-1}c^{-1}d^{-1}$.
(6) What sort of object is the ID-space represented by aaa or by aaa^{-1}? Are these objects surfaces? Can they be visualized as something that lives in $\mathbb{R}^3$?
(7) Is an ID-space with an even number of oriented edges, identified in pairs, always a surface?
(8) Compute $\chi(\mathbb{RP}^2)$ and $\chi(\mathbb{K})$ using triangulations of ID-spaces.

(9) Show that, in any triangulation of the sphere where each vertex has at least 5 edges, there are at least 20 faces.

(10) Consider the following variation of the connected sum of two tori. Start out by removing disks in each torus. Before gluing the tori together along the boundary circles, reverse the orientation of one of the circles. What surface do you end up with?

Chapter 4
Classification Theorem of Compact Surfaces

4.1 The Geometry of the Projective Plane and the Klein Bottle

We now take a small diversion to discuss some interesting properties of the projective plane and the Klein bottle that we introduced in the previous chapter. Recall that these are *abstract surfaces* that exist in their own right, without reference to an embedding space like $\mathbb{R}^3$, but which nonetheless are locally homeomorphic to open sets in the plane.

The Projective Plane. We start by presenting another way of describing the projective plane. Consider the set of all lines through the origin in $\mathbb{R}^3$. It turns out that we can make a sort of space out of this set (which will turn out to be $\mathbb{RP}^2$) as follows. Since the origin is on all these lines, and every other point is on exactly one of them, it makes sense to throw out the origin and look at $\mathbb{R}^3 \setminus \{(0, 0, 0)\}$. Now we consider two nonzero points (a, b, c) and (a', b', c') to be the same if they lie on the same line. Alternatively, we can rephrase this condition by saying that points (a, b, c) and (a', b', c') are considered the same if there is some (nonzero) number $\lambda \in \mathbb{R}$ so that $a' = \lambda a$, $b' = \lambda b$, and $c' = \lambda c$. In this way, each point of $\mathbb{R}^3 \setminus \{(0, 0, 0)\}$ is glued to many other points in $\mathbb{R}^3 \setminus \{(0, 0, 0)\}$. Now the "space of all lines" that we're studying becomes $\mathbb{R}^3 \setminus \{(0, 0, 0)\}$ but with points glued like this.

Remark 4.1 We can use the language of *equivalence relations* introduced in Chapter 2 to define this space more rigorously. Define a relation on $\mathbb{R}^3 \setminus \{(0, 0, 0)\}$ by $(a, b, c) \sim (a', b', c')$ if and only if $a' = \lambda a$, $b' = \lambda b$, and $c' = \lambda c$ for some nonzero $\lambda \in \mathbb{R}$. It turns out that $\sim$ is an equivalence relation on $\mathbb{R}^3$ (exercise). An equivalence class of $\sim$ is a line through the origin in $\mathbb{R}^3$, but with the origin removed (exercise). So the set of lines through the origin in $\mathbb{R}^3$ is the set of equivalence classes of $\sim$, which we denote as Q for now. The set Q inherits a topology from the topology on $\mathbb{R}^3$, thanks to the quotient topology. Namely: A set of equivalence classes in Q is open if the union of all lines represented by these classes is an open set in the usual sense in $\mathbb{R}^3$.

C. Bray et al., *Algebraic Topology*,
https://doi.org/10.1007/978-3-030-70608-1_4

So now we have made a topological space out of the set of all lines through the origin in $\mathbb{R}^3$. We now show that this space is in fact an abstract surface. To see this, note that for every line ℓ through the origin, we can find some point (a, b, c) on ℓ with length 1: $a^2 + b^2 + c^2 = 1$. Well, actually two points—both (a, b, c) and its antipodal point $(-a, -b, -c)$. All these points form a sphere in $\mathbb{R}^3$, and the relation $(a, b, c) \sim (a', b', c')$ if and only if $(a', b', c') = (a, b, c)$ or $(-a, -b, -c)$ is an equivalence relation. If we now define $\mathbb{RP}^2$ as the set of equivalence classes (which we can think of as the sphere with antipodal points glued together and representing the same point), then this $\mathbb{RP}^2$ becomes a surface under the quotient topology. Furthermore, we have a natural map from the space of lines to $\mathbb{RP}^2$, by sending ℓ to (a, b, c) or $(-a, -b, -c)$, that takes open sets of lines to open sets in $\mathbb{RP}^2$. (Exercise: think this through.)

Now pick a point ℓ in the space of all lines through the origin (i.e. pick a line through the origin in $\mathbb{R}^3$). This line ℓ corresponds a point $(a, b, c) \in \mathbb{RP}^2$ according to the procedure above. A small open neighborhood of (a, b, c) is equivalent to two small neighborhoods on the sphere—one around (a, b, c) and one around $(-a, -b, -c)$. Furthermore, these neighborhoods are in correspondence with an open neighborhood of lines containing ℓ. Now, because the sphere is a surface, each of these neighborhoods can be mapped to an open neighborhood of the plane. Thus we obtain a mapping from a neighborhood of lines to a neighborhood of $\mathbb{RP}^2$ to a pair of neighborhoods in $\mathbb{S}^2$ to a pair of neighborhoods in $\mathbb{R}^2$. We have an abstract surface!

Exercise 4.2 Show that an ID space representation for this version of $\mathbb{RP}^2$ is aa. Thus, it is the same as the version of $\mathbb{RP}^2$ introduced in Chapter 3.

Remark 4.3 There are also more general projective spaces $\mathbb{RP}^n$, which are spaces of lines through the origin in $\mathbb{R}^{n+1}$. Another description of $\mathbb{RP}^n$ is that it's the quotient of an n-dimensional sphere by the equivalence relation that equates antipodal points.

Exercise 4.4 What is $\mathbb{RP}^1$?

$\mathbb{RP}^2$ is a wonderful setting for doing geometry: There is a notion of a line, and it has the advantage over Euclidean geometry that any two lines intersect at a point: there are no parallel lines! Many classical theorems of geometry, such as Pascal's hexagon theorem, are really theorems about projective geometry and are best understood inside $\mathbb{RP}^2$, because in this setting we can avoid annoying special cases involving parallel lines. We invite the interested reader to look into the beautiful subject of projective geometry; see for instance [Cox94].

The Klein Bottle. The Klein bottle is closely related to the Möbius strip. In fact, we can cut a Klein bottle into a Möbius strip, as is shown in the ID-space diagram in Figure 4.1: cutting the top/bottom horizontal edge leaves a Möbius strip. Furthermore, if we make two parallel cuts on the Klein bottle, we obtain a Möbius strip and a cylinder, as in Figure 4.2.

An Important Relationship. An important relationship between $\mathbb{RP}^2$ and $\mathbb{K}$ is given in the following proposition. Its proof shows the usefulness of surface triangulations!

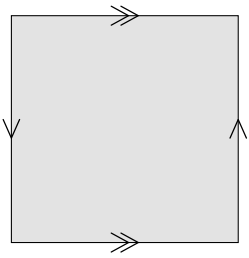


Figure 4.1 An ID space for the Klein bottle. If we cut along the top (or, equivalently, bottom) edge, we obtain a Möbius strip.

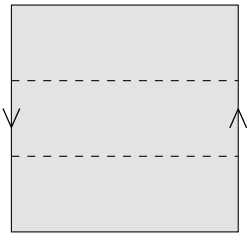


Figure 4.2 This is what happens when we make two parallel cuts (the solid line at the top/bottom, and the dashed lines in the middle, which together form one line) in the Klein bottle: we obtain a Möbius strip and something homeomorphic to a cylinder.

Proposition 4.5 *The surfaces* $\mathbb{RP}^2\#\mathbb{RP}^2$ *and* $\mathbb{K}$ *are homeomorphic.*

Proof We know that $\mathbb{RP}^2$ is homeomorphic to the ID space $abab$ and that $\mathbb{K}$ is homeomorphic to the ID space $cdcd^{-1}$. The sequence of "cut and paste" operations shown in Figure 4.3 shows that the ID spaces of $\mathbb{RP}^2\#\mathbb{RP}^2$ and $\mathbb{K}$ are in fact the same. ■

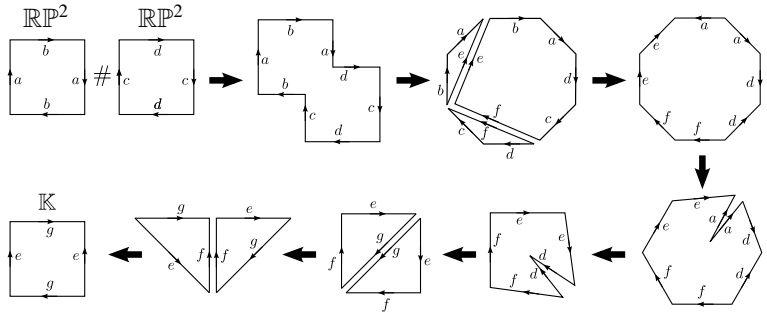


Figure 4.3 The cut-and-paste operations showing that $\mathbb{RP}^2\#\mathbb{RP}^2 = \mathbb{K}$.

4.2 Orientable and Nonorientable Surfaces

The Möbius strip, the Klein bottle, and the projective plane are examples of *nonorientable surfaces* (or nonorientable surfaces with boundary in the former case). In this section, we will define this notion more carefully.

The orientability of a compact surface or surface with boundary will be a *boolean* topological invariant—either a surface S is orientable or it is nonorientable; and this remains true for the image of S under any homeomorphism. One intuitive way to define this notion is by means of the "number of sides" that a surface embedded in $\mathbb{R}^3$ has. Take any such surface S and pick a point $p \in S$. Let's assume that S is in fact a very thin three-dimensional shell rather than an idealized, infinitesimally thin, two-dimensional membrane. Now pick a side of this shell at p and start painting it blue near p. Keep painting such that every new place you paint is physically reachable from a place that you've already painted. At some point, you will not be able to reach any unpainted parts of the shell. At this point you ask: have you painted the whole shell? If your shell is spherical and has two distinct sides, the answer is no. If your shell is the Möbius strip and has only one distinct side (boundary edges don't count) then the answer is yes. No means orientable, and yes means nonorientable!

The previous intuitive definition can actually be made rigorous in the case of surfaces embedded in $\mathbb{R}^3$. It requires a notion of "side," i.e. for every $p \in S$ there is a neighborhood in $\mathbb{R}^3$ containing p that can be subdivided into two subsets of $\mathbb{R}^3$ which we designate the "interior" and the "exterior" sides. We can encode the sidedness of S by defining a *unit normal vector* for each $p \in S$. This is a vector of length one that is orthogonal to the tangent plane of S—and note that S must be a smooth surface! The region into which the normal vector points is the "exterior." Now, S will be called orientable if it is possible to define this normal vector in a consistent and continuous way. More precisely:

Definition 4.6 A smooth, i.e. differentiable, compact surface S with or without boundary embedded in $\mathbb{R}^3$ is said to be *orientable* if there exists a continuous way to assign a unit normal vector at each $p \in S$. If no such assignment exists, then S is *nonorientable*.

Unfortunately, the preceding definition is not suitable for all the surfaces that we will encounter in this book. Certainly, we would have to modify this definition for non-smooth surfaces. But what about abstract surfaces for which an embedding into $\mathbb{R}^3$ is not given? It is still possible to define orientability in these cases, but we must proceed differently. To formulate a new definition, which is still equivalent to the old one, we proceed as follows. For each $p \in S$ we define an *orientation*. An orientation is an assignment of a small set of coordinate axes at p, where we decide what is the x-axis and positive x-direction, as well as the y-axis and the positive y-direction. Note that there are precisely two kinds of systems of coordinate axes: right-handed ones and left-handed ones. With a right-handed coordinate system, the "right-hand rule" describes the positive x-axis turning into the positive y-axis, whereas with a left-handed coordinate system, the "left hand rule" describes the positive x-axis

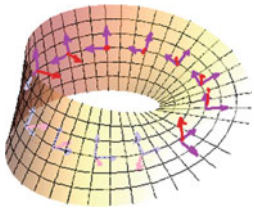


Figure 4.4 The Möbius strip is nonorientable, because the orientation changes as we travel around it.

turning into the positive y-axis. Now S will be called orientable if it is possible to define an orientation in a consistent way. Without loss of generality, this orientation can be right-handed. More precisely:

Definition 4.7 A smooth compact surface S—with or without boundary—is said to be *orientable* if there exists a continuous way to assign a right-handed orientation at each $p \in S$. If no such assignment exists, then S is said to be *nonorientable*.

Exercise 4.8 This alternate definition seems less "mathematical" than the first. Although even in the first definition (Definition 4.6), we haven't defined rigorously what it means for a normal vector to be assigned continuously. Both of these definitions are nevertheless perfectly rigorous. Can you explain how?

Remark 4.9 There is a simple way to detect the failure of orientability of a surface: the existence of an *orientation-reversing curve*. This is a non-trivial, continuous, closed curve $\gamma \subseteq S$ (one that starts and ends at a single point $p \in S$ and traverses a path that includes points other than p) upon which the following phenomenon occurs. Pick a direction in which to walk away from p along γ, and assign a right-handed orientation to each point on γ by assigning the positive x-axis to the direction you're moving and the positive y-axis to the side of the curve on your left. Now, eventually you will come back to p—and you might find that the positive y-axis has changed sides! See Figure 4.4 to see a path that changes the orientation on a Möbius strip.

Exercise 4.10 Find orientation-reversing curves on the Klein bottle and projective plane.

Exercise 4.11 Let S be an orientable surface and let S' be a nonorientable surface. What can we say about the orientability of $S \# S'$?

Exercise 4.12 Let S be a smooth, compact surface embedded in $\mathbb{R}^3$. Argue that given a continuous assignment of orientation as in Definition 4.7, we get a continuous assignment of normal vector as in Definition 4.6, and vice-versa. In other words, argue that these two definitions are equivalent in the case of smooth, embedded, compact surfaces.

Now that we have defined orientation, albeit in a somewhat non-rigorous fashion, we must establish that orientability and non-orientability are homeomorphism invariants.

Theorem 4.13 *Let S, S' be compact surfaces with or without boundary and let $\phi : S \to S'$ be a homeomorphism. Then S is orientable (or nonorientable) if and only if S' is orientable (or nonorientable).*

We can't prove this theorem rigorously at this point, but we can at least sketch the idea. If S is nonorientable, then there is a curve γ in S so that when we start with a right-handed orientation and travel along γ, we eventually end up with a left-handed orientation. Now, ϕ will drag γ to some curve γ' in S', and γ' will also switch a right-handed orientation in S' to a left-handed orientation.

The reason that this is not completely rigorous is that we don't really know anything yet about what ϕ does to γ at a global level. So, how can we be so sure that it isn't possible to find a homeomorphism such that γ' is orientation-preserving, even though γ is orientation-reversing? There are ways of dealing with this issue, mostly involving coming up with more technical definitions of orientability that are more conducive to proving theorems, but they would take us too far off course here. To point you in the right direction as you go through the rest of the book, orientability can be defined cleanly in terms of homology; once we have defined homology, you will immediately notice the difference between the top-dimensional homology of a Klein bottle and that of a torus.

4.3 The Classification Theorem for Compact Surfaces

In the material we have covered so far in this book, we have defined two invariants—the Euler characteristic and orientability—that are sufficiently powerful to *classify* all compact surfaces without boundary up to homeomorphism. In other words, we first define an equivalence relation on the set of all compact surfaces without boundary by saying that two surfaces are equivalent if and only if there exists a homeomorphism between them. Then we are able to prove that every equivalence class of surfaces can be uniquely described by two numbers: the Euler characteristic and the *orientation bit* (i.e. 1 if S is orientable and 0 otherwise). Moreover, for each equivalence class, we can specify a natural representative that has these two numbers. We will prove this in the remainder of this chapter. This is an important mathematical result with a long history. Perhaps Möbius first attempted a proof of this in [Möb61], but his proof was flawed. The first correct proof seems to have been given by Brahana in 1921 in [Bra21].

Theorem 4.14 (Classification of Compact Surfaces without Boundary) *Let S be a connected compact surface without boundary. Then S is homeomorphic to exactly one of the following surfaces:*

- *The sphere $\mathbb{S}^2$, which is orientable and has Euler characteristic $\chi = 2$.*
- *A connected sum of g tori, which is orientable and has Euler characteristic $\chi = 2 - 2g$.*

- *A connected sum of g projective planes, which is nonorientable and has Euler characteristic $\chi = 2 - g$.*

Therefore the homeomorphism type of S can be determined by knowing only the Euler characteristic and orientation bit of S.

Exercise 4.15 Why don't the surfaces of type

$$\mathbb{T}\# \cdots \#\mathbb{T}\#\mathbb{RP}^2\# \cdots \#\mathbb{RP}^2$$

appear on this list? Are these surfaces orientable?

Exercise 4.16 Where is the Klein bottle on this list?

For completeness, we state the classification of compact 1-manifolds, which are the 1-dimensional analogue of surfaces:

Theorem 4.17 *Let M be a compact connected 1-manifold, possibly with boundary. Then M is homeomorphic to the closed interval* $[0, 1]$*, or to a circle.*

We do not give a proof of this here. Instead, we refer the reader to David Gale's paper [Gal87] for an elementary proof.

4.4 Compact Surfaces Have Finite Triangulations

In the previous section, we stated that one of the invariants used in the classification theorem is the Euler characteristic. However, so far we do not necessarily know how to compute Euler characteristics for all compact surfaces. The problem is that we have defined the Euler characteristic in terms of triangulations. Thus, in order to guarantee that Euler characteristic makes sense for all compact surfaces, we need to show that *every compact surface admits a triangulation*.

Theorem 4.18 *Let S be a compact surface. Then S has a triangulation into finitely many triangles.*

This theorem is not easy to prove, and we will skip the proof here. You can find a relatively elementary—but rather long and intricate—proof in Thomassen's paper [Tho92].

It is worth pointing out that Theorem 4.18 is not nearly as obvious as it seems. While it is true for surfaces, and also for 3-dimensional manifolds broken up into tetrahedra, it is false in higher dimensions. Freedman in [Fre82] provided an example of a 4-dimensional manifold that is not triangulable, and Manolescu in [Man16] proved that there are also examples in all dimensions ≥ 5.

4.5 Proof of the Classification Theorem

The proof of the classification theorem uses the ID space representation of a surface. Therefore, we have to begin by showing that *every connected compact surface is homeomorphic to an ID space consisting of a polygon of some finite number* $2N$ *of sides that are identified in pairs*. We can argue, as follows, that this is true. First, apply Theorem 4.18 to the compact surface S to decompose it into a union of cells $\{T_1, \ldots, T_N\}$ that satisfy all the properties of a valid triangulation given in the last chapter. Label each edge on S with a unique identifier $a_1, \ldots, a_M$, and transfer these labels to the appropriate edges of all the T_i. Thus, each identifier is used exactly twice as a label among the edges of all the T_i's. Also, for each i, there is a planar triangle T_i' that is homeomorphic to the cell T_i. Let us label the edges of the T_i''s using the same labels as the T_i's.

At this point, we have already shown that S is homeomorphic to an ID space—namely the union of all triangles $T_1', \ldots, T_N'$ whose edges are identified according to the assignment of labels we have made. However, this ID space isn't a polygon! To get a polygon, first start by choosing any two triangles in $\{T_1', \ldots, T_N'\}$ that have an edge with the same label, and gluing them together along the labeled edge, as shown in Figure 4.5. Now we have an ID space for S consisting of $N - 2$ triangles and one lozenge. Now keep repeating this process until there are no triangles left. Every time you add a triangle, a pair of commonly labeled edges disappears. Thus the process terminates in a polygon with $2N$ boundary edges identified in pairs.

We will now prove the classification theorem by applying cut-and-paste operations to this ID space until we have an equivalent ID space that we recognize as one of the three different kinds of surfaces listed in the theorem. The following five steps allow us to reach this goal.

Step 1. We start with an ID space S, orient its boundary, and name its distinct edges as $a_1, a_2, a_3, \ldots, a_M$ for some M. Note that each a_i appears twice in S with either the same orientation or the opposite orientation. If it's the case that a_i appears with a_i^{-1}, then we say that the edge a_i is of Type I. If it's the case that a_i appears with a_i or a_i^{-1} appears with a_i^{-1}, then we say that the edge a_i is of Type II.

The first step of the proof of the classification theorem is to *remove adjacent Type I edges*. That is to say, if we find that two edges of S are labeled $a_i a_i^{-1}$ or $a_i^{-1} a_i$ for some i, then we can apply the "folding away trick" used in the previous cut-and-paste

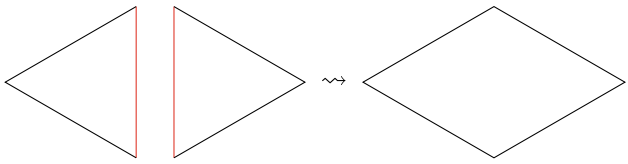


Figure 4.5 Converting two triangles into a lozenge. The red edges have the same label and can thus be glued.

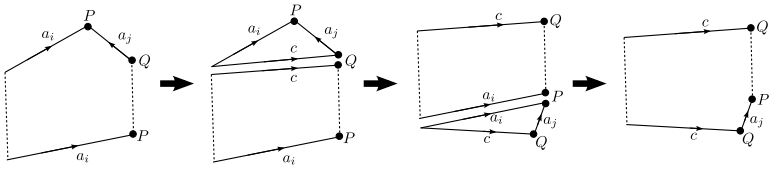


Figure 4.6 The cut-and-paste operations for Step 2.

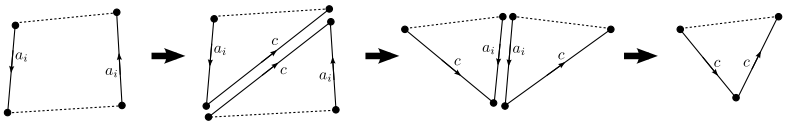


Figure 4.7 The cut-and-paste operations for Step 3.

proof giving a homeomorphism between $\mathbb{RP}^2\#\mathbb{RP}^2$ and $\mathbb{K}$, to show that S is identical to the ID space obtained from S by simply removing the pair of edges a_i and a_i^{-1}.

Note that it may be possible to collapse S all the way down to a single aa^{-1} pair. Of course we can't go any further, but no matter: aa^{-1} is the ID space for the sphere $\mathbb{S}^2$. Thus S is homeomorphic to $\mathbb{S}^2$ in this case.

Step 2. If we reach this stage, we now have an ID space S with no adjacent Type I edges. But it may be the case that not all of the vertices of the ID space are identified under the gluing instructions. (Exercise: give an example!) The second step of the proof of the classification theorem is to replace S with an equivalent ID space where *all vertices are identified*. (Exercise: give an example of a ID space with all vertices identified!)

Given the above, we can apply an inductive sequence of steps to S that allows us to replace S with an equivalent ID space where all vertices are identified with each other. To see how, suppose that there are at least two distinct vertices P and Q that appear i and j times, respectively, in S. Now apply the sequence of cut-and-paste operations shown in Figure 4.6, and obtain an equivalent ID space where P and Q appear $i-1$ and $j+1$ times. We continue to apply these operations until no P's are left. (Exercise: What happens at the end?)

Step 3. Next, we show that *all Type II edges of S can be made adjacent*. The sequence of cut-and-paste operations given in Figure 4.7 achieves this. If we do this for all Type II edges, we are in a situation where S looks like $S'\#\mathbb{RP}^2\#\cdots\#\mathbb{RP}^2$ for some other ID space S' and some number (possibly zero, if there were no Type II edges to begin with) of projective planes. Note that it may be the case that S' isn't there, so S is homeomorphic to $\mathbb{RP}^2\#\cdots\#\mathbb{RP}^2$ in that case.

Step 4. If we reach this stage, we now have

$$S = S'\#\mathbb{RP}^2\#\cdots\#\mathbb{RP}^2$$

(or perhaps no $\mathbb{RP}^2$'s), and S' contains only Type I edges. By Step 1, none of these edges are adjacent. Moreover, it is a consequence of Step 2 that, for every pair of

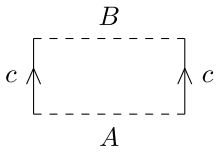


Figure 4.8 We show that in Step 4, there must be an edge of A identified with an edge of B.

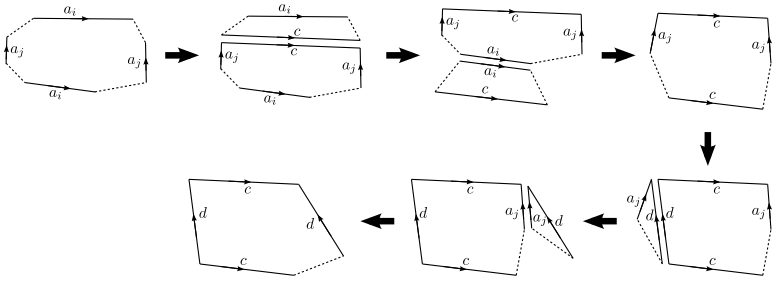


Figure 4.9 The cut-and-paste operations for Step 4.

Type I edges in S', there is a second pair of Type I edges in S' that separates them. To see this, suppose not: there is some pair of Type I edges c that do not separate any pair of Type I edges. Then, in Figure 4.8, every edge in A must be identified with another edge in A, and every edge in B must be identified with another edge in B. But then the two endpoints of c are not identified, contradicting Step 2, where we arranged for all the vertices to be identified.

Now the fourth step of the proof of the classification theorem is to show that *every such quartet of Type I edges in S' can be put into torus order*. This is achieved by applying the sequence of cut-and-paste operations shown in Figure 4.9.

Step 5. Once we reach this stage, we have

$$S = \mathbb{T}\# \cdots \#\mathbb{T}\#\mathbb{RP}^2\# \cdots \#\mathbb{RP}^2$$

for some number of each factor (possibly zero factors of one kind but not of both kinds). The remaining step is to show the following: *If there is one $\mathbb{RP}^2$ factor in S then S can be re-arranged to have only $\mathbb{RP}^2$ factors*. It thus suffices to show that $\mathbb{T}\#\mathbb{RP}^2$ is homeomorphic to $\mathbb{RP}^2\#\mathbb{RP}^2\#\mathbb{RP}^2$. There is a similar cut-and-paste argument for doing this. We leave this step as Problem 9.

This concludes the proof of the classification theorem. ■

4.6 Problems

(1) Prove the following rephrasing of the Classification Theorem:

> Any compact surface can be obtained from a sphere *in a unique way* by first doing a connect sum with some number of tori, and then doing a connect sum with either zero, one, or two projective planes.

(2) Determine which pairs of the following surfaces are homeomorphic to one another:

(a) $\mathbb{T}\#\mathbb{T}\#\mathbb{T}\#\mathbb{RP}^2$
(b) $\mathbb{T}\#\mathbb{T}\#\mathbb{K}\#\mathbb{RP}^2$
(c) $\mathbb{T}\#\mathbb{K}\#\mathbb{K}\#\mathbb{RP}^2$
(d) $\mathbb{K}\#\mathbb{K}\#\mathbb{K}\#\mathbb{RP}^2$
(e) $\mathbb{T}\#\mathbb{T}\#\mathbb{RP}^2\#\mathbb{RP}^2$
(f) $\mathbb{T}\#\mathbb{T}\#\mathbb{T}\#\mathbb{K}$
(g) $\mathbb{K}\#(\mathbb{RP}^2)^5$ (5 copies of $\mathbb{RP}^2$)
(h) $\mathbb{T}\#\mathbb{T}\#\mathbb{T}\#\mathbb{T}$
(i) $\mathbb{T}\#\mathbb{T}\#\mathbb{RP}^2\#\mathbb{RP}^2\#\mathbb{RP}^2$

(3) Show that it is *not* possible to subdivide the surface of a sphere into regions, each of which has six sides, such that any two regions have no more than one side in common.

(4) Suppose we have a sphere that is divided up into regions by n great circles, no three of which intersect in a common point. How many regions is the sphere divided into?

(5) Let S_1 be a surface (either orientable or not, perhaps with boundary) and let S_2 be a nonorientable surface (perhaps with boundary). Show that $S_1\#S_2$ is nonorientable.

(6) Let S be a compact surface with boundary.

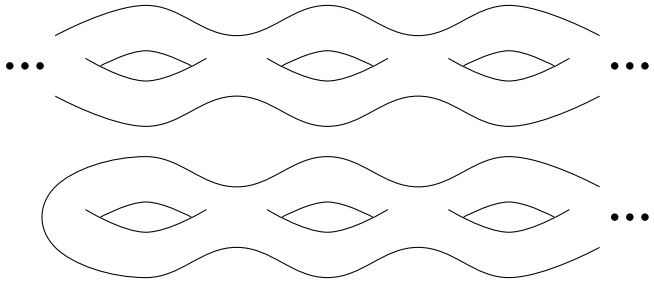
(a) Argue that the boundary of S consists of a finite collection of circles. Argue that it is possible to "cap off" each of these circles with a disk, and the object S' that results from this process is a compact surface.
(b) Using these ideas, determine a formula for the Euler characteristic of S in terms of that of S' and the number of boundary circles of S.
(c) Extend the classification theorem to compact surfaces with boundary. To see how, take a surface with boundary and use the "capping off" procedure to obtain a compact surface S'. Now use the original classification theorem to determine all the possibilities for S'. Now, what happens when you remove the disks? Try to formulate a clear statement and outline a proof, even if you cannot supply all details.

(7) Let S be a compact surface of Euler characteristic $\chi \leq 0$. Let $N = \frac{1}{2}(7 + \sqrt{49 - 24\chi})$; N is a root of the equation $6(1 - \chi/N) = N - 1$.

(a) If we have a polygonization (like a triangulation, but with polygons instead of triangles) $\mathcal{P}$ of S with V vertices, E edges, and F faces, then show that $E \leq 3(F - \chi)$.

(b) Show that $2E < \lfloor N \rfloor F$. Conclude that there is some region with fewer than $\lfloor N \rfloor$ edges.
(c) Show that it is possible to color the faces of $\mathcal{P}$ with at most $\lfloor N \rfloor$ colors so that any two adjacent regions are colored differently. (The result is also true when $\chi = 1$, but the proof is different. The result is also true when $\chi = 2$, but the proof is *much* harder: this is the infamous *Four-Color Theorem* proven in [AH77] and [AHK77]. This is also best possible except in the case of the Klein bottle, where only 6 colors are needed rather than 7.)

(8) Are the following two surfaces homeomorphic?



Imagine trying to classify noncompact surfaces. What might such a classification look like? Can you guess what the statement of this theorem might look like? What difficulties arise?[1]

(9) Consider the surfaces $\mathbb{T}\#\mathbb{RP}^2$ and $\mathbb{RP}^2\#\mathbb{RP}^2\#\mathbb{RP}^2$. In order to complete the classification of compact surfaces, we must show that they are homeomorphic. Show that this is true by using an ID-space rearrangement.

[1] In case you are curious, a complete description of the classification of noncompact surfaces can be found in [Ric63].

Chapter 5
Introduction to Group Theory

5.1 Why Use Groups?

So far, in order to understand topological spaces, we have been using numerical invariants such as the Euler characteristic in order to detect whether spaces are homeomorphic or not. However, there is a wide class of other invariants, which associate other sorts of objects to spaces. For the next few chapters, we will build up to the fundamental group, and then we will work on understanding its behavior.

At this point, it would be reasonable to wonder why we would bother looking for more invariants of topological spaces now that we have already completely classified compact surfaces. There are many reasons to do this, but let us focus on two at the moment:

(1) Compact surfaces turn out to be uncommonly easy to classify, because they can be completely described only by their Euler characteristic and orientability. But if we look at different classes of topological spaces, for example 3-dimensional manifolds, then the situation is far more complicated, and the set of invariants needed to classify them is quite a bit longer. The classification of 3-manifolds was only recently more-or-less resolved, with Perelman's proof of Thurston's Geometrization Conjecture in [Per02, Per03b, Per03a]. As a result, if we wish to figure out whether topological spaces that aren't just surfaces are homeomorphic or not, we need a wider range of tools. In general, it will not be possible to write down a complete invariant for interesting classes of topological spaces. This statement can be made into a precise theorem, but that's far beyond the scope of this book; see [Mar60].

(2) The Euler characteristic does not give us very much insight into the possible relationships between two spaces. For example, if we have two surfaces S and S' and a continuous function from S to S'—which need not be a homeomorphism—what can we say about the Euler characteristics of S and S'? Or, if we know the Euler characteristics of S and S', can we say something about the map? The answer here is, more or less, no. There is one theorem, due to Riemann and Hurwitz (see [Mir95, Theorem 4.16]), which gives us a small amount of

C. Bray et al., *Algebraic Topology*,
https://doi.org/10.1007/978-3-030-70608-1_5

information, but the relationship is rather weak. Other invariants allow us to say interesting things about maps between spaces as well as just the spaces themselves. A modern theme in mathematics, pioneered by Grothendieck, Eilenberg, and Mac Lane, and others (see for instance [EML45]), is that we can best understand mathematical objects not in isolation, but based on their maps to and from other similar types of objects.

5.2 A Motivating Example

Let us start with a square S in the plane. What are the rigid motions of the plane that send the square to itself? That is, what are the distance-preserving functions $f : \mathbb{R}^2 \to \mathbb{R}^2$ for which $f(x) \in S$ for every $x \in S$? One could count these, and we will do so later, but for now let us instead look at the properties that such functions have.

Let G be the set of such functions. Inside G, there is a special element: the element that takes every point back to itself. This element is called the *identity*. If we pick any two elements $\sigma, \tau \in G$, then we can investigate what happens when we first apply σ and then apply τ. If we start with a point $x \in S$, then $\sigma(x) \in S$. But since $\sigma(x) \in S$, we also have $\tau(\sigma(x)) \in S$. Hence the composition of σ and τ, written $\tau\sigma$, is also an element of G. We think of $\tau\sigma$ as the "product" of τ and σ. Note, however, that $\tau\sigma$ need not be the same as $\sigma\tau$. For example, if σ is a rotation by $\pi/2$ counterclockwise and τ is a reflection about the y-axis, then $\sigma\tau$ and $\tau\sigma$ are not the same. (Convince yourself of this!)

If we multiply the identity by any element σ (in either order), we just end up with σ again. This is the basic property of the identity. If we have any element $\sigma \in G$, we can "undo" the effect of σ: there is some element, called σ^{-1}, so that $\sigma^{-1}\sigma(x) = x$. In this case, multiplying in the reverse order also has the same effect: $\sigma\sigma^{-1}(x) = x$ as well. We call σ^{-1} the *inverse* of σ. For example, if σ is a rotation by $\pi/2$ counterclockwise, then σ^{-1} is a rotation by $\pi/2$ clockwise.

Finally, if we have three elements $\rho, \sigma, \tau \in G$, then consider the elements $(\rho\sigma)\tau$ and $\rho(\sigma\tau)$. These two elements are actually the same: $(\rho\sigma)(\tau(x)) = \rho(\sigma\tau(x))$ for all $x \in S$, because this is always true of compositions of functions. This is called the *associative property*, or *associativity*. It tells us that we can unambiguously compute threefold (or higher) products of elements by chaining together, in the correct order, a sequence of pairwise products.

5.3 Definition of a Group

We are now ready to introduce the notion of a group. A group, roughly, is an object that satisfies all the properties that G had in the above example. More precisely, we have the following definition:

Definition 5.1 A group is a pair $(G, \cdot)$, where G is a set, and $\cdot : G \times G \to G$ is a binary operation, satisfying the following three properties:

- (Identity property.) There is some element $e \in G$ so that, for any $g \in G$, $e \cdot g = g \cdot e = g$.
- (Inverses.) For any $g \in G$, there is an element $g^{-1} \in G$ so that $g \cdot g^{-1} = g^{-1} \cdot g = e$.
- (Associativity.) If g, h, k are any three elements of G, then $g \cdot (h \cdot k) = (g \cdot h) \cdot k$.

Remark 5.2 Frequently, we suppress the $\cdot$ for multiplication and just write gh rather than $g \cdot h$. However, if we use a different notation for the group multiplication, such as "+" (which we will frequently do), then we do not suppress the symbol for the operation. Also, we will generally refer to a group as G, rather than the pair $(G, \cdot)$, when it is clear from context what the appropriate operation is.

Theorem 5.3 *Let G be a group. Then g has a* unique *identity element.*

Proof Let us suppose that G has two elements e and e' that both satisfy the identity property. Then consider the element $e \cdot e'$. On the one hand, the identity property for e states that $e \cdot g = g$ for any choice of $g \in G$. In particular $e \cdot e' = e'$. On the other hand, the identity property for e' states that $g \cdot e' = g$ for any choice of $g \in G$. In particular $e \cdot e' = e$. Hence we have $e = e'$. ■

Theorem 5.4 *Let G be a group, and let g be an element of G. Then g has a* unique *inverse.*

Proof Since G is a group, we know that g has at least one inverse. Let us suppose that it has two, say h and k. Then consider the element $h(gk)$. Since k is an inverse of g with $gk = e$, this element is equal to $he = h$. However, by associativity, $h(gk) = (hg)k$. Now, since h is also an inverse of g with $hg = e$, this element is equal to $ek = k$. Hence this element is equal to both h and k, so $h = k$. Thus these two supposedly different inverses were actually the same. ■

5.4 Examples of Groups

We now take the opportunity to introduce several important groups and collections of groups, as well as give some nonexamples of groups.

Elementary Examples. We first rephrase the very familiar sets of "ordinary numbers" in the language of groups.

Example The integers $(\mathbb{Z}, +)$ form a group under the operation of addition. The identity is $e = 0$, and the inverse of $n \in \mathbb{Z}$ is $-n$. Similarly, the rationals $(\mathbb{Q}, +)$ and the real numbers $(\mathbb{R}, +)$ form groups under addition.

Nonexample The integers $(\mathbb{Z}, \times)$ under multiplication do not form a group. The only possible identity element is $e = 1$, but then there are no inverses: for example, there is no integer x so that $2x = 1$, so 2 does not have an inverse. Similarly, the rational numbers $(\mathbb{Q}, \times)$ under multiplication do not form a group, because there is no inverse for 0.

Example But the rational numbers under multiplication come close to forming a group: 0 is the only element without an inverse. Hence the nonzero rationals, denoted $\mathbb{Q}^\times$ and pronounced "Q star," do form a group under multiplication. Similarly, the nonzero real numbers form a group under multiplication. There are other closely related groups to these: for example, the *positive* rational or real numbers under multiplication also form groups. The negative ones do not!

The Integers Modulo n. We now define a very important group, denoted $\mathbb{Z}/n\mathbb{Z}$, where n is a positive integer. The set of elements here is the integers, except that we consider two integers to be the same if they leave the same remainder upon division by n. For example, 1 and 7 leave the same remainder upon division by 6, so they are the same element in $\mathbb{Z}/6\mathbb{Z}$. We call $\mathbb{Z}/n\mathbb{Z}$ the *integers modulo n*.

Remark 5.5 We can rephrase the construction of $\mathbb{Z}/n\mathbb{Z}$ in terms of an equivalence relation. We put an equivalence relation $\sim$ on $\mathbb{Z}$ by saying that $a \sim b$ if $a - b$ is a multiple of n. (Exercise: Verify that this is an equivalence relation!) Then the set of $\mathbb{Z}/n\mathbb{Z}$ is the set of equivalence classes of the equivalence relation $\sim$.

How many elements are there in $\mathbb{Z}/n\mathbb{Z}$? Well, if we start with any integer k, we can divide k by n and end up with some remainder between 0 and $n - 1$; that is, we have $k = qn + r$ for some integers q and r with $0 \leq r \leq n - 1$. Then k is equal to r in $\mathbb{Z}/n\mathbb{Z}$. Furthermore, all the numbers from 0 to $n - 1$ are different in $\mathbb{Z}/n\mathbb{Z}$. Hence $\mathbb{Z}/n\mathbb{Z}$ consists of exactly n elements, which we can think of as being the integers from 0 to $n - 1$.

Some notation will be convenient here: If two numbers a and b correspond to the same element of $\mathbb{Z}/n\mathbb{Z}$, then we write $a \equiv b \pmod{n}$. This is equivalent to saying that $a - b$ is a multiple of n.

Now let us see how to make $\mathbb{Z}/n\mathbb{Z}$ into a group, under addition. If we take two elements x and y of $\mathbb{Z}/n\mathbb{Z}$, we can pretend that they are normal integers and add them together, and we would end up with some integer $x + y$. Since $x + y$ may now be larger than n, we then reduce $x + y$ modulo n. Thus we define $+$ in $\mathbb{Z}/n\mathbb{Z}$ as $x + y \pmod{n}$. We have to be a little bit careful though: Does this definition really make sense?

What could go wrong? Let us suppose that, instead of picking x and y in $\mathbb{Z}$ that reduce to the desired elements of $\mathbb{Z}/n\mathbb{Z}$, we chose other integers x' and y' that are equivalent to x and y in $\mathbb{Z}/n\mathbb{Z}$, i.e. $x \equiv x' \pmod{n}$ and $y \equiv y' \pmod{n}$. Is $x' + y' \equiv x + y \pmod{n}$? It has to be, in order for our group operation to make sense!

Exercise 5.6 Show that if $x \equiv x' \pmod{n}$ and $y \equiv y' \pmod{n}$, then $x + y \equiv x' + y' \pmod{n}$.

Exercise 5.7 What is the identity and what are inverses in $\mathbb{Z}/n\mathbb{Z}$?

As a result of these exercises, addition in $\mathbb{Z}/n\mathbb{Z}$ makes sense, and $\mathbb{Z}/n\mathbb{Z}$ under the operation of addition forms a group.

Remark 5.8 The notation for the integers modulo n might look strange, but it is an example of a general construction of groups called *quotient groups*. We will see in Chapter 7 how to define quotient groups in general. There is also another notation for the group of integers modulo n, written C_n, which stands for "the cyclic group of order n." We will also soon see what a cyclic group is.

Multiplication Tables. Sometimes, it will be important to be able to define group operations, not by having general rules for how to multiply two elements, but rather by just listing all possible products of the elements. To do this, we can write down tables for the products of all pairs of elements in a set and check that this does in fact give us a group structure. Let us show an example, of a group with four elements, which we call a, b, c, d.

$\times$	a	b	c	d
a	a	b	c	d
b	b	a	d	c
c	c	d	a	b
d	d	c	b	a

(5.1)

We read this table just like we would read an ordinary multiplication table. For example, the entry in the c row and the b column (which in this case is d) is the product cb.

It won't be very fun, but it is possible to check that the operation $\times$ on the set $\{a, b, c, d\}$ as described by (5.1) satisfies all the properties needed to be a group. This group is called the *Klein 4-group*, or the *Vierergruppe*. This is the same Klein as the one responsible for the Klein bottle.

Exercise 5.9 What is the identity in this group? What are the inverses of all the elements?

Exercise 5.10 Show that, in any group multiplication table, every element appears exactly once in each row and each column.

Dihedral Groups. Let us now return to the motivating example above: the rigid motions preserving a square. This is a group called D_4, where the D stands for "dihedral." More generally, the group of rigid motions of a plane preserving a regular n-gon is called D_n. These are important examples of groups, but we will defer their discussion until Section 5.5.

Exercise 5.11 How many elements does D_n have?

Symmetric Groups. Another important family of groups consists of the permutations of all the elements of some set. Let us consider the set $\mathcal{X}_n = \{1, 2, 3, \ldots, n\}$. Let S_n

denote the set of all ways of rearranging the elements of $\mathcal{X}_n$. There are $n!$ elements of S_n.

We would like to put a group structure on S_n. To do so, we view a permutation of $\mathcal{X}_n$ as a bijective function $\sigma : \mathcal{X}_n \to \mathcal{X}_n$ from $\mathcal{X}_n$ to itself. In this interpretation, multiplication of two permutations is the same as composition of two bijective functions. The identity permutation is the same as the identity function. The inverse of a permutation is the same as the inverse of the function representing that permutation. This set-up is perhaps best illustrated with an example. Let us suppose $n = 4$, and let us take two elements of S_4: σ will be the element that puts $\mathcal{X}_4$ into the order 1432, and τ will be the element that puts $\mathcal{X}_4$ into the order 3421. This means that, as functions, σ and τ behave as follows:

$$\sigma(1) = 1 \qquad \sigma(2) = 4 \qquad \sigma(3) = 3 \qquad \text{and} \qquad \sigma(4) = 2$$

as well as

$$\tau(1) = 3 \qquad \tau(2) = 4 \qquad \tau(3) = 2 \qquad \text{and} \qquad \tau(4) = 1\,.$$

We can write these identities in more concise notation as

$$\sigma = \begin{pmatrix} 1\ 2\ 3\ 4 \\ 1\ 4\ 3\ 2 \end{pmatrix}, \qquad \tau = \begin{pmatrix} 1\ 2\ 3\ 4 \\ 3\ 4\ 2\ 1 \end{pmatrix}.$$

By inspection, we can read off the inverse functions as

$$\sigma^{-1}(1) = 1 \qquad \sigma^{-1}(4) = 2 \qquad \sigma^{-1}(3) = 3 \qquad \sigma^{-1}(2) = 4$$

as well as

$$\tau^{-1}(3) = 1 \qquad \tau^{-1}(4) = 2 \qquad \tau^{-1}(2) = 3 \qquad \tau^{-1}(1) = 4\,.$$

Rearranging these identities, and using our concise notation, yields

$$\sigma^{-1} = \begin{pmatrix} 1\ 2\ 3\ 4 \\ 1\ 4\ 3\ 2 \end{pmatrix}, \qquad \tau^{-1} = \begin{pmatrix} 1\ 2\ 3\ 4 \\ 4\ 3\ 1\ 2 \end{pmatrix}.$$

Note that $\sigma = \sigma^{-1}$ (this also implies $\sigma^2 = 1$). This is interesting!

With this interpretation in mind, we can now multiply σ and τ in the order $\sigma\tau$. We view this as the composite function $\sigma \circ \tau : \mathcal{X}_n \to \mathcal{X}_n$. Consequently, $\sigma \circ \tau(1) = \sigma(\tau(1)) = \sigma(3) = 3$, and similarly for all other elements in $\mathcal{X}_n$. We obtain

$$\sigma\tau = \begin{pmatrix} 1\ 2\ 3\ 4 \\ 3\ 2\ 4\ 1 \end{pmatrix}.$$

It is useful to write elements of S_n in a different form, called cycle notation. Let us take a slightly longer example: consider the permutation of $\mathcal{X}_8$ given by

$$\sigma = \begin{pmatrix} 1\,2\,3\,4\,5\,6\,7\,8 \\ 8\,2\,6\,5\,7\,3\,1\,4 \end{pmatrix}.$$

Let us now start with 1 and figure out where it goes: it goes to position 8. Now that we're focused on 8, let's figure out where that goes: it goes to position 4. Where does 4 go? It goes to position 5. Where does 5 go? It goes to position 7. And 7? It goes to position 1. And now we're back to where we started. So, let us write down the list of numbers we encountered in that process: (18457). It describes the permutation that permutes the numbers 18457 in cyclic order. We call this kind of a permutation a *cycle*.

Now, the cycle we just found above doesn't completely describe the permutation σ because we still have not investigated the action of σ on the remaining numbers 2, 3, and 6. So, let's now try the same process starting from 2. The number 2 goes to position 2, and now we have already looped back. We can write that as a (very short) cycle though: it's just (2). Once again, we can repeat the process starting with 3. The number 3 goes to 6, and then 6 goes back to 3, so we have another cycle (36). If we put all the cycles together, as (18457)(2)(36), we have completely described the permutation. This is called the cycle decomposition of the cycle. Notationally, we tend to delete any cycles of length one, e.g. the cycle (2) in the cycle decomposition we have just found, because all cycles of length one are really the same: the identity permutation. Finally, note that we can honestly write $\sigma = (18457)(36)$ as the multiplication of two permutations, i.e. the composition of two functions that act by permuting the numbers in $\mathcal{X}_8$. These are the permutations $\tau_1 = (18457)$ and $\tau_2 = (36)$. So we have $\sigma = \tau_1\tau_2$. Note that $\tau_1\tau_2 = \tau_2\tau_1$.

Exercise 5.12 Show that disjoint cycles, i.e. cycles without any common elements, always commute.

The following type of cycle is especially important:

Definition 5.13 An element of S_n that switches exactly two elements of $\mathcal{X}_n$ is called a *transposition*.

Abelian Groups. One key property that certain groups have is that any two elements *commute*: if a and b are two elements, then $ab = ba$. If this happens, we say that the group is *abelian*. This was named after Norwegian mathematician Niels Henrik Abel. True fame in mathematics is indicated by having your name made into an improper adjective.

Example The group $\mathbb{Z}$ of integers is an abelian group.

Nonexample The symmetric group S_n is not abelian, because (12) and (23) do not commute with one another.

Exercise 5.14 Classify the groups we have looked at so far on the basis of whether they are abelian or not.

Remark 5.15 Frequently, when we are especially interested in abelian groups, we write the operation as addition rather than multiplication, we denote the identity by 0 rather than 1, and we denote the inverse of an element g by $-g$ rather than g^{-1}. This is by analogy with the group of integers under addition.

5.5 Free Groups, Generators, and Relations

So far, all of our examples of groups have been pretty concrete: we have seen cyclic groups, which can be described in terms of explicit elements, and we have seen groups such as dihedral groups and symmetric groups, which we have interpreted as the symmetries of certain objects. These are important ways of thinking about groups, and many groups naturally arise in this way. However, sometimes we will want to think of groups in a more abstract way: by describing how certain key elements relate to each other.

Let us see how we could have described the dihedral group D_4 in such a manner. The first step will be to find several elements in D_4 such that every element can be obtained by multiplying these together; such a set of elements will be called a *generating set*. One possibility is to take all the elements of D_4 as a generating set. This is perfectly valid, but it isn't very efficient. In order to state how all the elements relate to each other, we would need to write down an entire multiplication table, which has 64 entries.

We do better by choosing two elements, which we call ρ and σ. We let ρ be rotation by $\pi/2$ in the counterclockwise direction, and we let σ be a reflection about the y-axis. Every element of D_4 can be written as a product of powers of ρ and powers of σ, possibly with many repetitions. (Exercise: Why?)

However, just specifying that we can build D_4 out of products of ρ and σ is not enough: that doesn't tell us, for example, that σ^2 is the identity. So, we also need to specify certain identities that these two elements satisfy; in this case, the three relations

$$\sigma^2 = e, \tag{5.2}$$

$$\rho^4 = e, \tag{5.3}$$

$$\sigma\rho = \rho^{-1}\sigma \tag{5.4}$$

are enough to specify all the group behavior. (Exercise: Why do these relations hold?)

From this information, how can we list all the elements of the group? First, let us see what a typical element of D_4 should "look like" in terms of ρ and σ. Since every element is expressible in terms of them, this means that we can write a typical element as

$$\rho^{a_1}\sigma^{b_1}\rho^{a_2}\sigma^{b_2}\cdots\rho^{a_n}\sigma^{b_n},$$

for some integers a_i and b_i (which might be zero). However, we can simplify such an expression using the rules (5.2)–(5.4). For example, if any a_i is at least 4, we can replace it with $a_i \pmod 4$, which we interpret as being an integer between 0 and 3. Similarly, if any a_i is negative, we can replace it with an integer between 0 and 3. The same thing holds for the powers of σ, except that now we can replace a power of σ by either $\sigma^0 = e$ or $\sigma^1 = \sigma$. If any exponents are 0, then we can remove those terms and shorten the expression.

The relation (5.4) allows us to do even more: If we ever have a σ before a ρ, then we can switch the order, at the cost of replacing the ρ with a ρ^{-1}. So, we can simplify the expression by forcing all the ρ's to come before all the σ's. Hence every element can be expressed as one of the following eight:

$$e, \rho, \rho^2, \rho^3, \sigma, \rho\sigma, \rho^2\sigma, \rho^3\sigma.$$

Let us note that there are other groups that have generating sets $\{\rho, \sigma\}$ satisfying (5.2)–(5.4). For example, in the trivial group $G = \{e\}$, we can take both ρ and σ to be e, and then they clearly satisfy (5.2)–(5.4). But D_4 is in some sense the "best" group described in this way: it satisfies those relations, and no others that do not automatically follow from them.

Let us now look at groups described in this way more generally. Let G be a group, and let S be a generating set for G—i.e. every element of G can be expressed as a product of elements in S, possibly with many repetitions. We call the elements of S *generators*. Let R be a set of identities in the elements of S that are satisfied in G—so that a group generated by the set S subject to the identities in R is forced to be G or smaller. The elements of R are called *relations*. The pair (S, R) is called a *presentation* for G.

Notation We usually write $G = \langle S \mid R\rangle$ when we want to say that G is a group generated by the elements S, subject to the relations R. Hence, we can write

$$D_4 = \langle \rho, \sigma \mid \rho^4 = e, \sigma^2 = e, \sigma\rho = \rho^{-1}\sigma\rangle.$$

Example Let us work out some presentations for the symmetric group S_n. There are many natural choices of presentations, and we will look at two of the most important. The first one is

$$S_n = \langle \tau_1, \ldots, \tau_{n-1} \mid \tau_i\tau_j = \tau_j\tau_i \text{ if } |i-j| \geq 2, \\ \tau_i\tau_{i+1}\tau_i = \tau_{i+1}\tau_i\tau_{i+1},\ \tau_i^2 = e\rangle.$$

In order to interpret this as the symmetric group we are used to, we need to explain which elements (written in terms of permutations) the τ_i's are. In this case, τ_i is the transposition $(i, i+1)$. (Exercise: Verify that these elements actually satisfy the

relations we claim they do. Can you show that any $\sigma \in S_n$ can be written as a product of generators?)

The other important presentation for the symmetric group, with just two generators, is

$$S_n = \langle x, y \mid x^2 = y^n = (xy)^{n-1} = e, (xy^{-1}xy)^3 = e, \\ (xy^{-j}xy^j)^2 = e \text{ for } 2 \leq j \leq \lfloor n/2 \rfloor \rangle.$$

In this case, x is the transposition (12) and y is the n-cycle $(123\cdots n)$.

Exercise 5.16 Convince yourself that the first presentation is actually a presentation for S_n. In the second case, at least convince yourself that x and y generate S_n. You are encouraged to Play with decks of cards. (Hint: it suffices to check that every transposition can be expressed in terms of x and y, because the transpositions generate S_n.)

Exercise 5.17 What is the shortest presentation you can give for the cyclic group $\mathbb{Z}/n\mathbb{Z}$?

Free Groups and Free Abelian Groups. Some groups have particularly simple presentations, in that we do not need any relations in their presentation. Such groups are called *free groups*.

Example The group $\mathbb{Z}$ of integers is a free group. Its presentation is

$$\mathbb{Z} = \langle 1 \mid \ \rangle.$$

Note that when we specify a free group, we do not put anything after the vertical bar, because there are no relations. In the case of the integers, we can write everything as $n \times 1$, for some n. That completely describes the group: we don't need any further information to cut it down to the right size. Indeed, any relation must take the form $1 \times n = 0$ for some n, and that is false in $\mathbb{Z}$ for $n \neq 0$.

Example The trivial group is also a free group. It has no generators and also no relations. Hence we can write $\{e\} = \langle \mid \rangle$. But generally, people do not use this notation.

Free groups with at least two generators have a different feel to them from the trivial group and the group of integers. Let us consider the free group on two generators:

$$F_2 = \langle a, b \mid \ \rangle.$$

As we saw when discussing the dihedral group, a typical element of F_2 has the form

$$a^{m_1}b^{n_1}a^{m_2}b^{n_2}\cdots a^{m_k}b^{n_k}$$

for some k and some integers m_i and n_i. However, unless some m_i or n_i is equal to 0, we cannot shorten this expression, as every similar expression corresponds to

a different element of F_2. The situation is completely similar for free groups on more—and possibly infinitely many—generators.

Free groups with at least two generators are not abelian. However, we have a notion similar to free groups in the setting of abelian groups: we can have groups that have no relations other than the ones that are needed to make the groups abelian. Such a group has the following presentation:

$$F_n^{\text{ab}} = \langle x_1, x_2, \ldots, x_n \mid x_i x_j = x_j x_i \rangle.$$

We call F_n^{ab} the *free abelian group on n generators*, or just a free abelian group.

5.6 Free Products

Let us take another look at the free group F_2 on two generators, a and b. We know that a typical element has the form

$$a^{m_1} b^{n_1} \cdots a^{m_k} b^{n_k}.$$

There are two important smaller groups ("subgroups," which we will introduce formally in the next chapter) inside F_2: there is the group of all powers of a, and there is the group of all powers of b. Each of these looks like a copy of the integers $\mathbb{Z}$. Let us call these two groups A and B, respectively.

To write a general element of F_2, then, we take some element of A, then multiply it by some element of B, and then we go back to A, and so forth. We can perform this operation more generally, in the following way. Let G and H be two groups. We wish to form a new group out of them, as we formed F_2 out of A and B. The group we form is called the *free product* of G and H, and it is denoted $G * H$. A typical element is of the form

$$g_1 h_1 g_2 h_2 \cdots g_k h_k,$$

for some $g_i \in G$ and $h_i \in H$. (It might also start with an element of H or end with an element of G.) If we also require that none of the g_i and h_i are the identity elements in G and H, respectively, then such a representation is unique.

We can also write down a presentation for $G * H$ in terms of presentations of G and H. Suppose that $G = \langle S_G \mid R_G \rangle$ and $H = \langle S_H \mid R_H \rangle$. Suppose furthermore that S_G and S_H are disjoint. Then we have

$$G * H = \langle S_G \cup S_H \mid R_G \cup R_H \rangle.$$

Remark 5.18 Free products come up naturally when studying fundamental groups; we'll see them all over the place shortly. In fact, the study of fundamental groups is the main reason that people are interested in free products. But one free product shows up, rather unexpectedly, in number theory: $(\mathbb{Z}/2\mathbb{Z}) * (\mathbb{Z}/3\mathbb{Z})$. Consider the

group of matrices

$$\mathrm{SL}_2(\mathbb{Z}) = \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} : a, b, c, d \in \mathbb{Z} \text{ and } ad - bc = 1 \right\}.$$

The "SL" in the name of this group stands for "special linear," and the subscript "2" tells us that we are looking at 2×2 matrices. (Exercise: Check that this is a group!) This group isn't quite a free product, but it is very close. If we consider the two matrices A and $-A$ to be the same, similar to what we did to construct $\mathbb{Z}/n\mathbb{Z}$ from $\mathbb{Z}$, we obtain a new group: $\mathrm{PSL}_2(\mathbb{Z})$. The "$P$" stands for "projective." Then $\mathrm{PSL}_2(\mathbb{Z})$ is the free product of $\mathbb{Z}/2\mathbb{Z}$ and $\mathbb{Z}/3\mathbb{Z}$. Hence, $\mathrm{PSL}_2(\mathbb{Z})$ has the presentation

$$\langle a, b \mid a^2 = b^3 = 1 \rangle.$$

Here, we can take

$$a = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \qquad b = \begin{pmatrix} 0 & -1 \\ 1 & 1 \end{pmatrix}.$$

5.7 Problems

(1) Prove that $\mathbb{Z}/5\mathbb{Z} \setminus \{0\}$ is a group under multiplication modulo 5. Find its multiplication table. More generally, for which n is $\mathbb{Z}/n\mathbb{Z} \setminus \{0\}$ a group under multiplication modulo n?

(2) Find all possible groups with four elements.

(3) Which of the following are groups? If not, can the set be modified in a simple way to make it into a group? Which are abelian?

   (a) The set of all translations, rotations about the origin *and* reflections across arbitrary lines of the plane.

   (b) $(\{\text{Continuous functions on } \mathbb{R}\}, +)$ where $(f + g)(x) = f(x) + g(x)$ defines $f + g$.

   (c) $(\{\text{Continuous functions on } \mathbb{R}\}, \cdot)$ where $(f \cdot g)(x) = f(x)g(x)$ defines $f \cdot g$.

   (d) The set of 2×2 matrices with real entries under matrix multiplication. We multiply 2×2 matrices with the following rule:

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} e & f \\ g & h \end{pmatrix} = \begin{pmatrix} ae + bg & af + bh \\ ce + dg & cf + dh \end{pmatrix}.$$

(4) Consider the transformation group G consisting of all isometries of $\mathbb{R}^2$—i.e. consisting of all translations, rotations about an arbitrary point and reflections across arbitrary lines, along with all their compositions and inverses. If you would like to have a simple presentation of this group in terms of simple building

blocks, you would need to work out the commutation relations between different motions. Let τ_a denote translation in the direction $a \in \mathbb{R}^2$; let $\rho_{b,\theta}$ denote a counterclockwise rotation by the angle $\theta \in [0, 2\pi)$ about the point $b \in \mathbb{R}^2$; and let $\pi_{c,m}$ denote the reflection across the line whose slope is $m \in [-\infty, \infty)$ and that passes through the point $c \in \mathbb{R}^2$. We'd like to show that G is generated in some way by $\rho_{0,\theta}$, τ_a and $\pi_{0,0}$. The standard form you would like to achieve is $g \in G$ can be written $g = \pi_{0,0}^{\varepsilon} * \tau_a * \rho_{0,\theta}$ where $\varepsilon = 0$ or 1, and $a \in \mathbb{R}^2$ and $\theta \in [0, 2\pi)$.

(a) Write $\rho_{b,\theta}$ in terms of translations and $\rho_{0,\theta}$.
(b) Write $\pi_{c,m}$ in terms of translations, rotations, and $\pi_{0,0}$.
(c) How can you commute $\pi_{0,0}$ past translations? (In other words: suppose T is a translation; now what group element g satisfies $\pi_{0,0} * T = g * \pi_{0,0}$?)
(d) How can you commute $\pi_{0,0}$ past rotations?
(e) What happens when you encounter $\pi_{0,0}^k$ where $k \geq 2$?
(f) Describe the procedure for putting an arbitrary composition of rotations about points in space, reflections about various lines, and translations into standard form.

(5) The *dihedral group of order n* is the group of symmetries of a regular n-gon and is denoted D_n. Consider the group D_4 of symmetries of the square. Let ρ be the rotation by one quarter turn counterclockwise (i.e. by $\pi/2$ radians), and let σ be the reflection across the horizontal line going through the center of the square.

(a) How many different elements does D_4 have?
(b) Write all of these elements in terms of ρ and σ.
(c) Write out a multiplication table for D_4.
(d) Is D_4 abelian?

(6) Find a group with eight elements (e.g., write down its multiplication table or construct it from simpler groups) such that every element is its own inverse.

(7) (a) Find the cycle decompositions of

(i)
$$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\ 6 & 4 & 1 & 3 & 2 & 7 & 9 & 8 & 5 \end{pmatrix}$$

(ii)
$$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\ 6 & 4 & 5 & 3 & 2 & 7 & 9 & 8 & 1 \end{pmatrix}$$

(b) Find the cycle decomposition of the product

$$(1327)(453)(287)$$

in S_8.

(c) Let $\tau = (a_1, a_2, \ldots, a_k)$ be a cycle in S_n and let $\sigma \in S_n$. Show that

$$\sigma\tau\sigma^{-1} = (\sigma(a_1), \ldots, \sigma(a_k)) .$$

(d) Now let $\rho = c_1 c_2 \cdots c_k$ be a product of cycles in S_n. Show that $\sigma\rho\sigma^{-1} = c_1' c_2' \cdots c_k'$, where each c_i' is obtained from c_i according to the rule you just determined.

(e) Let $\tau = (a_1, a_2, \ldots, a_k)$ be a cycle in S_n. Show that

$$\tau = (a_1 a_2)(a_2 a_3)(a_3 a_4) \cdots (a_{k-1} a_k) .$$

In this way, show that every permutation can be written as a product of cycles of length 2.

Chapter 6
Structure of Groups

6.1 Subgroups

This chapter is an introduction to the rich structure possessed by a set endowed with a group operation. The first notion we will explore is that of *subgroups*, or subsets of a group that themselves satisfy all the properties of a group.

Definition 6.1 Let G be a group. A nonempty subset $H \subset G$ is called a *subgroup* if it is closed under taking products and inverses. In other words:

- If $h_1, h_2 \in H$, then the product $h_1 h_2 \in H$.
- If $h \in H$, then the inverse $h^{-1} \in H$.

We will use the notation $H \leq G$ to denote that H is a subgroup of G. Here are some elementary properties of subgroups.

Proposition 6.2 *Let $H \leq G$ be a subgroup. Then the following hold.*

(1) An alternative defining property for a subgroup is: H is a subgroup if and only if H is nonempty and $h_1 h_2^{-1} \in H$ for all $h_1, h_2 \in H$.
(2) The identity e belongs to H.
(3) H is a group in its own right.

Proof (1) If H is a subgroup and $h_1, h_2 \in H$, then we know that $h_2^{-1} \in H$ as well, and thus that $h_1 h_2^{-1} \in H$ by the primary defining properties of a subgroup. Hence we have established the alternative defining property. Conversely, suppose the alternative defining property holds and choose any $h \in H$. Then, letting both h_1 and h_2 be h, we have $hh^{-1} = e \in H$. Next, by choosing $h_1 = e$ and $h_2 = h$, we know that $h_1 h_2^{-1} = h^{-1} \in H$. Next, choose $h_1, h_2 \in H$. Hence $h_2^{-1} \in H$ by what we have just shown. Hence $h_1 h_2 = h_1 (h_2^{-1})^{-1} \in H$ by the alternative defining property. In this way, we have established both primary defining conditions.
(2) To show that $e \in H$, we proceed as follows. Pick any $h \in H$ and let $h_1 = h_2 = h$. Now, when we apply the alternative defining property, we find that $h_1 h_2^{-1} = hh^{-1} = e \in H$, as desired.

C. Bray et al., *Algebraic Topology*,
https://doi.org/10.1007/978-3-030-70608-1_6

(3) Finally, H is a group because it contains the identity and inverses of all its elements, and associativity is inherited from G.

■

It is time to give several examples of subgroups. The *order* of a finite group or subgroup is equal to the number of elements it contains.

Example Let $G = D_3 = \langle \sigma, \rho \mid \sigma^2 = \rho^3 = e \text{ and } \sigma\rho\sigma = \rho^2 \rangle$. A subgroup of order two is $H = \{e, \sigma\}$. A subgroup of order three is $H = \{e, \rho, \rho^2\}$. (Exercise: Show that Proposition 6.2(1) holds for each H. What are the other subgroups of D_3?)

Example Let G be any group, and let $g \in G$ be an arbitrary element. We can build a subgroup out of g that we denote $\langle g \rangle$, namely $\langle g \rangle = \{g^n \mid n \in \mathbb{Z}\}$. In other words, $\langle g \rangle$ consists of e, all products of g with itself, and all products of g^{-1} with itself. This subgroup is called the *cyclic subgroup generated by* g. If the subgroup $\langle g \rangle$ has order N, then we say the element g itself has order N. (Exercise: Prove that N is the smallest positive integer such that $g^N = e$.)

Example We can generalize the previous example as follows. Let $g_1, \ldots, g_k \in G$ be elements. Then the set

$$\langle g_1, \ldots, g_k \rangle = \{\text{products of } g_1, \ldots, g_k \text{ and their inverses}\}$$

is also a subgroup. It is called the *subgroup generated by* $g_1, \ldots, g_k$.

Example Let $G = D_4 = \langle \sigma, \rho \mid \sigma^2 = \rho^4 = e \text{ and } \sigma\rho\sigma = \rho^3 \rangle$. A subgroup of order four that is not cyclic is $H = \langle \sigma, \rho^2 \rangle = \{e, \sigma, \rho^2, \sigma\rho^2\}$.

Example Let $G = S_n$, and let A_n be the set of all possible products of an even number of transpositions. (Exercise: Show that A_n is a subgroup.) We call A_n the *alternating group*. (Exercise: Can you express it in the form $A_n = \langle \cdots \rangle$?)

6.2 Direct Products of Groups

If the theme of the previous section was finding smaller groups within bigger groups, the theme of this section is constructing bigger groups from smaller ones. There are several such constructions in group theory, and we will present only one of these—which is the most straightforward and ubiquitous of such constructions.

Let G_1 and G_2 be two groups. We will show that the *direct product* of G_1 and G_2, namely the set of pairs of elements, one from G_1 and one from G_2, defined by

$$G_1 \times G_2 = \{(g_1, g_2) : g_1 \in G_1 \text{ and } g_2 \in G_2\},$$

can be made into a group. To do this, we first must define a binary operation on $G_1 \times G_2$. The obvious choice is

$$(g_1, g_2) \cdot (h_1, h_2) = (g_1 h_1, g_2 h_2)$$

for any (g_1, g_2) and (h_1, h_2) in $G_1 \times G_2$. Next, we must define an identity element, and the obvious choice is

$$e = (e_1, e_2),$$

where e_1 is the identity in G_1 and e_2 is the identity in G_2. Finally, we must define inverses, and the obvious choice is

$$(g_1, g_2)^{-1} = (g_1^{-1}, g_2^{-1})$$

for every $(g_1, g_2) \in G_1 \times G_2$.

Remark 6.3 If G_1 and G_2 have finite orders N_1 and N_2, respectively, then the order of $G_1 \times G_2$ is equal to $N_1 N_2$.

Proposition 6.4 *The set $G_1 \times G_2$, equipped with the binary operation and the identity element and inverses as defined above, is a group.*

Proof We must show that the group properties of Definition 5.1 hold for $G_1 \times G_2$. First, we must show that the putative identity element we have defined truly "behaves like" an identity element. To this end, let $(g_1, g_2) \in G_1 \times G_2$ be any element. Then the computation

$$e \cdot (g_1, g_2) = (e_1, e_2) \cdot (g_1, g_2) = (e_1 g_1, e_2 g_2) = (g_1, g_2)$$

confirms this behavior. Similarly, we must show that the putative inverse element (g_1^{-1}, g_2^{-1}) truly "behaves like" the inverse of (g_1, g_2). The computation

$$(g_1^{-1}, g_2^{-1}) \cdot (g_1, g_2) = (g_1^{-1} g_1, g_2^{-1} g_2) = (e_1, e_2) = e$$

confirms this. Finally, we must show that the operation $\cdot$ is associative. To this end, let (g_1, g_2), (h_1, h_2) and (k_1, k_2) be any three elements of $G_1 \times G_2$. Then the computation

$$\begin{aligned}
\big((g_1, g_2) \cdot (h_1 h_2)\big) \cdot (k_1, k_2) &= (g_1 h_1, g_2 h_2) \cdot (k_1, k_2) \\
&= \big((g_1 h_1) k_1, (g_2 h_2) k_2\big) \\
&= \big(g_1 (h_1 k_1), g_2 (h_2 k_2)\big) \\
&= (g_1, g_2) \cdot (h_1 k_1, h_2 k_2) \\
&= (g_1, g_2) \cdot \big((h_1, h_2) \cdot (k_1, k_2)\big)
\end{aligned}$$

confirms associativity. We have used the associativity of G_1 and G_2 to pass from the second line to the third line above. ■

Example The direct product of $\mathbb{Z}/2\mathbb{Z}$ with itself is $\mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/2\mathbb{Z} = \{(0, 0), (1, 0), (0, 1), (1, 1)\}$. The group operation is $(a_1, a_2) + (b_1, b_2) = (a_1 + b_1 \pmod 2, a_2 + b_2 \pmod 2)$. This group is the same as, i.e. isomorphic to (in the language we will see very soon), the Klein 4-group that we saw in Chapter 5.

Example Let G and H be two groups, and let $A \leq G$ and $B \leq H$ be any subgroups. Then $A \times B$ is a subgroup of $G \times H$. (Exercise: Show this. Can you come up with an example of groups G and H such that $G \times H$ has a subgroup that is not of this form?)

Remark 6.5 We mentioned at the beginning of this section that there are other sorts of product-like constructions to build new groups out of smaller ones. We won't define the others here, but in case you're interested in learning about them, here are a few others. There are *semidirect products* and their generalizations known as *group extensions*. Then there are *wreath products*, which have a rather different feel to them. All of these notions are discussed in [Rot95, Chapter 7].

6.3 Homomorphisms

A general principle in mathematics is that once you have defined an interesting structure, you should also study the maps that preserve that structure. Thus when you are studying topology, you should study continuous functions and especially homeomorphisms. We now consider the types of maps between groups that preserve the basic structure. These are known as *homomorphisms*; the homomorphisms that are bijective are called *isomorphisms*.

Remark 6.6 Do not get homomorphisms confused with homeomorphisms. Despite the similarities in the words, they are very different notions. Homomorphisms are for groups, or more generally for algebraic structures, whereas homeomorphisms are for topological spaces. In fact, homeomorphisms of topological spaces more closely resemble isomorphisms of groups. We will see that there are relationships of this type once we have studied our group-theoretic invariants of topological spaces.

Definition 6.7 Let $(G, \cdot)$ and $(G', \otimes)$ be two groups. Then a function $f : G \to G'$ is said to be a *homomorphism* if for any $g_1, g_2 \in G$ we have

$$f(g_1 \cdot g_2) = f(g_1) \otimes f(g_2).$$

In this way, the function f "preserves the structure" of G and G'. This is because the most important structure, the group multiplication (the multiplication $\cdot$ of G on the left hand side above, and the multiplication $\otimes$ of G' on the right hand side above), is "respected" by f. In the future, we'll continue to suppress the multiplication symbol when convenient, so we'll write $f(g_1 g_2) = f(g_1) f(g_2)$.

Example Let G and G' be any groups. Then there is always at least one homomorphism from G to G'—the *trivial* homomorphism defined by $f(x) = e'$, where $x \in G$ is arbitrary and e' is the identity in G'. (Exercise: Show that f is indeed a homomorphism.)

Example Let $G = G' = \mathbb{Z}$. Group homomorphisms $f : \mathbb{Z} \to \mathbb{Z}$ are of the form $f(x) = nx$, where $n \in \mathbb{Z}$ is a fixed integer. Note that for such a function, we do indeed have $f(x + y) = n(x + y) = nx + ny = f(x) + f(y)$.

Example Let $G = \mathbb{Z}$ and $G' = \mathbb{Z}/2\mathbb{Z}$, and define $f : \mathbb{Z} \to \mathbb{Z}/2\mathbb{Z}$ by

$$f(x) = \begin{cases} 1 & x \text{ is odd,} \\ 0 & x \text{ is even.} \end{cases}$$

This is a homomorphism, because we can show $f(x + y) = f(x) + f(y)$ by considering even and odd x, y separately; we can also use the fact that the sum of two even numbers and the sum of two odd number is even, while the sum of an even and an odd number is odd.

Example Generalizing the example above, define $f : \mathbb{Z} \to \mathbb{Z}/n\mathbb{Z}$ by $f(x) = x \pmod{n}$. This is a homomorphism; to see this, we must show that $x + y \pmod{n} = x \pmod{n} + y \pmod{n}$. This follows because we have defined the group $\mathbb{Z}/n\mathbb{Z}$ in Chapter 5 as the *set of equivalence classes* of the equivalence relation $x \sim y$ on $\mathbb{Z}$ if and only if $x - y$ is a multiple of n; and we have defined addition in this group as ordinary addition of representatives of the classes of numbers, followed by taking remainders upon division by n. In the next chapter, we'll see a generalization of this sort of homomorphism.

Group homomorphisms satisfy several elementary properties.

Proposition 6.8 *Let $f : G \to G'$ be a homomorphism between groups G and G'.*

- *$f(e) = e'$, where e is the identity in G and e' is the identity in G'.*
- *$f(x^{-1}) = \big(f(x)\big)^{-1}$ for any $x \in G$.*
- *If $x \in G$ has order n, then $f(x)$ has order at most n.*

Proof For the first statement, consider the equalities $f(e) = f(ee) = f(e)f(e)$ that follow by the homomorphism property. We still don't know what $f(e)$ is, but since it belongs to the group G', it has to have an inverse $(f(e))^{-1}$. Now multiply both sides of the equality by this inverse; we get

$$e' = f(e)(f(e))^{-1} = f(e)f(e)(f(e))^{-1} = f(e)e' = f(e).$$

For the second statement, we compute

$$f(x^{-1})f(x) = f(x^{-1}x) = f(e) = e'.$$

This follows from the homomorphism property and from what we have just shown. So we see that $f(x^{-1})$ "behaves like" the inverse of $f(x)$. We know from Chapter 5 that inverses are unique, so we can conclude that $f(x^{-1})$ actually is the inverse of $f(x)$ or, in other words, $f(x^{-1}) = (f(x))^{-1}$.

The third statement follows from $(f(x))^n = f(x^n) = f(e) = e'$. We will see a stronger version of this statement in Chapter 7. ■

A homomorphism $f : G \to G'$ between groups G, G' is always associated with two special subgroups, one inside G and the other inside G'.

Definition 6.9 The *kernel* of the homomorphism $f : G \to G'$ is the set

$$\ker(f) = \{g \in G : f(g) = e'\},$$

where e' is the identity in G'. The *image* of f is the set

$$\mathrm{im}(f) = \{g' \in G' : \text{there exists } g \in G \text{ so that } g' = f(g)\}.$$

The image of f is also denoted $f(G)$.

Proposition 6.10 *Let $f : G \to G'$ be a homomorphism. Then $\ker(f) \leq G$ and $im(f) \leq G'$.*

Proof For the first statement, let $x, y \in \ker(f)$. Then $xy^{-1} \in \ker(f)$, because

$$f(xy^{-1}) = f(x)f(y^{-1}) = f(x)(f(y))^{-1} = e(e^{-1}) = ee = e.$$

Thus $\ker(f)$ satisfies Proposition 6.2(1). Also, $\ker(f)$ is nonempty, because $e \in \ker(f)$.

For the second statement, let $g, h \in \mathrm{im}(f)$. Then we know that $g = f(x)$ and $h = f(y)$ for some $x, y \in G$. But now

$$gh^{-1} = f(x)(f(y))^{-1} = f(x)f(y^{-1}) = f(xy^{-1}).$$

Hence we have succeeded in writing gh^{-1} in the form required to belong to $\mathrm{im}(f)$; in other words, $gh^{-1} \in \mathrm{im}(f)$. Also, $\mathrm{im}(f)$ is nonempty, because for any $x \in G$, $f(x) \in \mathrm{im}(f)$ and G is nonempty. Thus $\mathrm{im}(f)$ also satisfies Proposition 6.2(1). ■

The kernel and the image of f characterize the degree to which f fails to be bijective in the following sense. Consider injectivity first. If $f(x) = f(y)$ for some pair of elements $x, y \in G$, then $f(xy^{-1}) = f(x)(f(y))^{-1} = e$ or xy^{-1} belongs to $\ker(f)$. Conversely, if k is a non-trivial element of $\ker(f)$ then x and kx are different elements in G so that

$$f(kx) = f(k)f(x) = ef(x) = f(x).$$

Therefore f is injective if and only if $\ker(f) = \{e\}$. Surjectivity is simpler: f is onto G' if *every* element $g' \in G'$ can be written as $g' = f(g)$ for some $g \in G$, or in other words $\mathrm{im}(f) = G'$. If $\mathrm{im}(f)$ is a strictly smaller subgroup, then there are elements in G' that can not be "reached" from G by mapping under f, and f is not surjective.

6.4 Isomorphisms

A homomorphism $f : G \to G'$ is called an *isomorphism* if it is bijective. In terms of kernels and images, if $f : G \to G'$ is an isomorphism, then the kernel of f is the trivial subgroup $K = \{e\} \subset G$, and the image of G is the entire group G'. However, notice in particular that if only the kernel of f is trivial, then $f : G \to f(G) = H' \subset G'$ is always an isomorphism between G and H'.

An easy but important fact is that because f is bijective, it has an inverse, and this inverse f^{-1} is also a homomorphism. This is because $f(g_1 g_2) = f(g_1) f(g_2)$ iff $f^{-1}(g_1') f^{-1}(g_2') = f^{-1}(g_1' g_2')$, where $g_j' = f(g_j)$.

Isomorphisms should be thought of as the natural notion of equivalence of groups, so if two groups are isomorphic, then they are really "the same" even though they may have been described in very different ways initially. (Exercise: Check that indeed $G \sim G'$ if and only if G is isomorphic to G' is an equivalence relationship amongst groups.[1])

When G and G' are isomorphic, we shall write $G \cong G'$ (rather than $G = G'$). One interesting and important fact is that if $G \cong G'$, then there may be many different isomorphisms between these groups. The notation $\cong$ omits the explicit choice of isomorphism and simply records that there is at least one. An isomorphism from a group to itself, i.e. $f : G \to G$, is called an *automorphism*. The identity map (so $f(g) = g$ for every g) is always an automorphism, but sometimes—in fact, nearly always—there are nontrivial automorphisms.

Example Here are two simple examples that show how two groups may initially look fairly different yet still be isomorphic.

- The groups $G = \{\pm 1\}$ (with multiplication) and $G' = \mathbb{Z}/2\mathbb{Z} = \{0, 1\}$ (with addition) are isomorphic. Indeed, we obviously just need to define f by

$$f(1) = 0, \qquad f(-1) = 1.$$

  In fact, it is easy to see that *any* group with precisely two elements is isomorphic to this G! Let's prove this (though the proof is simpler than one of the homework exercises you have already done). The point is that there is precisely one coherent

[1] Actually this isn't quite true, because there is no set of all groups: the collection of all groups forms a proper class. However, nothing goes wrong, at least initially, if we put equivalence relations on proper classes. Things can become a little more complicated, though, when we wish to take a set of representatives or look at the set of all equivalence classes.

way to fill in the multiplication table for a group with two elements, in which a is the identity:

$*$	a	b
a	$\cdot$	$\cdot$
b	$\cdot$	$\cdot$

Namely, we have to put in a on the two principal diagonal slots, and b in the two off-diagonal slots. Once we know this, then clearly we can define a homomorphism by mapping a to 1 and b to -1.

- Let $G = \mathbb{Z}/6\mathbb{Z}$ and $G' = \mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/3\mathbb{Z}$. Then $G \cong G'$. In fact, define $f : G \to G'$ by setting $f(1) = (1, 1)$. Since G is cyclic, this is enough to determine f completely, and you can check that this really is a bijective homomorphism. It is an interesting problem to decide when the two groups $\mathbb{Z}/k\mathbb{Z}$ and $\mathbb{Z}/n\mathbb{Z} \times \mathbb{Z}/m\mathbb{Z}$ are isomorphic.

6.5 Existence of Homomorphisms

Given any two groups G and G', it is an interesting question to decide whether there are any homomorphisms $f : G \to G'$, and if so, how many. Actually, there is always at least one—the trivial homomorphism. Beyond this one, however, there are sometimes many others and sometimes no others at all. Let's try to understand this through a few observations and examples.

First consider the case where G is cyclic, so $G = \mathbb{Z}$ or $\mathbb{Z}/n\mathbb{Z}$ for some n, and let G' be any other group. The element 1 is a generator for G. Therefore it makes sense to try to first define a homomorphism f by its action on this generator, and then see what consequences follow. For example, if we define $g' = f(1)$ for some $g' \in G'$, then using additive notation for G', we know that $f(2) = f(1 + 1) = f(1) + f(1) = g' + g'$. In general, for any integer ℓ, we have $f(\ell) = f(1 + \cdots + 1)$ (ℓ times), hence in G' this is equal to $f(1) + \cdots + f(1)$ (ℓ times), or simply $\ell g'$. Therefore if $G = \mathbb{Z}$, this creates no problems, and in fact we can define a homomorphism f in this way by choosing g' to be an arbitrary element of G'. Note that f really is a homomorphism, because

$$f(k + \ell) = kg' + \ell g' = (k + \ell)g',$$

which is all that we need to check. Thus we have proved that there are very many homomorphisms $f : \mathbb{Z} \to G'$ for any group G'. (In fact $|G'|$ many.)

However, now suppose that $G = \mathbb{Z}/n\mathbb{Z}$. Then $\ell \equiv 0$ if $\ell = kn$ for some integer k, and this means that unless $f(n) = n \cdot f(1) = e'$, the identity in G', then we are in trouble and f cannot be a homomorphism. This means that unless $g' \in G'$ has the very special property, that $ng' = e'$, then there is no homomorphism $f : \mathbb{Z}/n\mathbb{Z} \to G'$ such that $f(1) = g'$. If e' is the only such $g' \in G'$, this means that the only homomorphism

from $\mathbb{Z}/n\mathbb{Z}$ to G' is the trivial homomorphism. For example, suppose that $G' = \mathbb{Z}$. If f is a (putative) homomorphism between G and G' and $f(1_n) = k$ (where we use the notation 1_n to denote the element 1 in $\mathbb{Z}/n\mathbb{Z}$ so as to distinguish it from $1 \in \mathbb{Z}$), then

$$f(\ell \cdot 1_n) = f(1_n + \cdots + 1_n) = f(1_n) + \cdots + f(1_n) = \ell k,$$

and as we observed before, if $\ell = n$, then the left hand side equals 0 and the right hand side is nonzero unless $k = 0$. This proves the following proposition:

Proposition 6.11 *If $n \in \mathbb{Z}$, $n \geq 2$, then there is no nontrivial homomorphism $f : \mathbb{Z}/n\mathbb{Z} \to \mathbb{Z}$.*

We can recast this slightly more generally in the following proposition. Note that we use multiplicative notation so that $g + g = 2g$ becomes $g \cdot g = g^2$, and ℓg for some integer ℓ becomes g^ℓ.

Proposition 6.12 *Suppose every element $g \in G$ has finite order, i.e. for each g there exists some ℓ (depending on g) such that $g^\ell = e$. Suppose on the other hand that the only element of finite order in G' is e'. Then there are no nontrivial homomorphisms from G to G'.*

Proof The proof uses the same reasoning as we just used above. ■

Here is another useful (and somewhat related) fact. Suppose that G is abelian. Then the subgroup $\mathrm{im}(f)$ in G' must be an abelian subgroup. The reason is simply that for any $g_1, g_2 \in G$,

$$f(g_1)f(g_2) = f(g_1g_2) = f(g_2g_1) = f(g_2)f(g_1),$$

or in other words, any two elements $f(g_1)$ and $f(g_2)$ in $f(G)$ commute with one another as elements of G'. Now, *any* group G' has some nontrivial abelian subgroups. Indeed, just as above, choose any $g' \in G'$, $g' \neq e'$, and consider the cyclic subgroup generated by g': $H' = \{(g')^\ell : \ell \in \mathbb{Z}\}$. This is obviously abelian. Remember that this may contain infinitely many elements, which happens if and only if $(g')^\ell \neq e'$ except when $\ell = 0$. Or, it may be finite and so is isomorphic to $\mathbb{Z}/n\mathbb{Z}$ for some n, and this is the case when $(g')^n = e'$ but $(g')^\ell \neq e'$ for $1 \leq \ell < n$.

Here is a much broader generalization of all of this. To state it, first recall an idea from Chapter 5. We said that G is generated by some subset of its elements $g_1, \ldots, g_n$ if any element in G can be written—not necessarily uniquely—as a "word" in these elements and their inverses. That is, every $g \in G$ can be written in at least one way as

$$g = g_{i_1}^{\ell_1} g_{i_2}^{\ell_2} \cdots g_{i_N}^{\ell_N},$$

where each $i_j \in \{1, \ldots, n\}$ and each $\ell_j \in \mathbb{Z}$. We also discussed some interesting examples of this, e.g. when a cyclic group $\mathbb{Z}/n\mathbb{Z}$ is generated by one of its elements p, and the group of all rigid motions in $\mathbb{R}^2$ or $\mathbb{R}^3$ is generated by the subset of

reflections (across lines or planes, respectively). In this last example, both the group and the set of generators are infinite. For simplicity, we shall usually just work with groups with only finitely many generators, but—unless we say so explicitly—everything works just the same if there are infinitely many generators. Now clearly, if $g_1, \ldots, g_n$ generate G, and $f : G \to G'$ is a homomorphism, then $f(g_1), \ldots, f(g_n)$ generate the subgroup $\mathrm{im}(f)$ of G'.

Finally, we come to the point we wish to make. If G is generated by some collection of elements $\{g_j\}$, then *every relationship satisfied by the g_j must also be satisfied by the images $g'_1 = f(g_n), \ldots, g'_n = f(g_n)$ in G'*. Thus, the two instances of this we have already discussed are that if $g_j^\ell = e$ for some g_j and ℓ, then $f(g_j)^\ell = e'$; similarly, if $g_j g_k = g_k g_j$, then $f(g_j)f(g_k) = f(g_k)f(g_j)$ in G'. All that this more general statement is asserting is that no matter how complicated the relationships satisfied by these generators, e.g. if

$$g_1 g_2^{-25} g_1^{-17} g_2^2 g_3 g_4^{-1} = e$$

in G, then it is also true that

$$g'_1 (g'_2)^{-25} (g'_1)^{-17} (g'_2)^2 g'_3 (g'_4)^{-1} = e'$$

in G'. More precisely:

Theorem 6.13 *Suppose $G = \langle g_1, \ldots, g_n \mid r_1, \ldots, r_m \rangle$ is a presentation of G, and suppose $r_i = g_{i1}^{e_{i1}} g_{i2}^{e_{i2}} \cdots g_{ik}^{e_{ik}}$, where the g_{ij}'s are generators, and the e_{ij}'s are integers. Then, for any group H, there is a natural bijection between the following two sets:*

- *Homomorphisms $\phi : G \to H$.*
- *Tuples $(h_1, \ldots, h_n)$ of elements of H (not necessarily distinct), for which $h_{i1}^{e_{i1}} h_{i2}^{e_{i2}} \cdots h_{ik}^{e_{ik}}$ is the identity of H, where $h_{ij} = h_\ell$ if $g_{ij} = g_\ell$.*

The bijection works as follows: if $\phi : G \to H$ is a homomorphism, then $h_\ell = \phi(g_\ell)$. On the other hand, if we have a collection of elements $h_1, \ldots, h_n$ with the above property, then we can construct a homomorphism $\phi : G \to H$ by setting $\phi(g_\ell) = h_\ell$ and using the homomorphism property to extend ϕ to all of G starting from the generators.

Or, more informally, a homomorphism $\phi : G \to H$ is equivalent to the data of the image of the generators, subject to the constraint that all the relations map to the identity. In particular, to define a homomorphism from a free group to any other group, it suffices to specify what the images of the generators are, and any images will work.

One possible moral of all of this is that if G is generated by a set of elements that have rather complicated relationships, and if G' is another group for which any set of generators satisfies only much simpler relationships, then any homomorphism $f : G \to G'$ must be somehow "close" to the trivial homomorphism.

6.6 Finitely Generated Abelian Groups

A natural question to ask is whether one can classify all groups up to isomorphism, i.e. to write down a list of all possible groups $\{G_1, G_2, \ldots\}$ such that every group is isomorphic to one and only one element of this list. This turns out to be an unreasonably hard question, and is in a certain sense known to be impossible! However, if we just consider groups of certain special types, then one can sometimes answer this question. Here's an elementary example: Classify all cyclic groups. We know the answer. Any cyclic group is isomorphic either to the trivial group, or else to some $\mathbb{Z}/n\mathbb{Z}$ where $n = 2, 3, \ldots$, or else to $\mathbb{Z}$. A less trivial problem, which requires a real proof, is to classify all *finitely generated abelian groups*. We will discuss this classification theorem in the remainder of this section.

Let us review what we know. First of all, amongst all possible groups of this type, there are some which are infinite, such as $\mathbb{Z}$ or $\mathbb{Z}^2 = \mathbb{Z} \times \mathbb{Z}$, and others which are finite, such as $\mathbb{Z}/n\mathbb{Z}$ or $\mathbb{Z}/n\mathbb{Z} \times \mathbb{Z}/m\mathbb{Z}$. Furthermore, if both G and G' are finitely generated abelian, then so is $G \times G'$. However, we do know that "redundancies" can occur, e.g. $\mathbb{Z}/6\mathbb{Z} \cong \mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/3\mathbb{Z}$. So, even though the first reasonable guess is that an arbitrary finitely generated abelian group should be obtained by starting with the examples we know, namely $\mathbb{Z}$ and $\mathbb{Z}/n\mathbb{Z}$ and taking some finite number of direct products of these, we will still be faced with the problem of winnowing down this list to cull out all these redundancies. Furthermore, we still also need to prove that there are no weird extra examples that we didn't know about beforehand, or that there are no other ways of combining two abelian groups—other than by a direct product—to obtain an abelian group. Here is the general result.

Theorem 6.14 *Let G be any finitely generated abelian group. Then there exists a finite collection of (not necessarily distinct) prime numbers $p_1, \ldots, p_n$, positive integers $\ell_1, \ldots, \ell_n$, and a nonnegative integer k such that*

$$G \cong \mathbb{Z}^k \times \mathbb{Z}/p_1^{\ell_1}\mathbb{Z} \times \cdots \times \mathbb{Z}/p_n^{\ell_n}\mathbb{Z}.$$

Furthermore, this decomposition is unique up to rearrangement of the factors.

Sketch of the Proof Notice some special cases of this. First, if G is a *finite* abelian group, then $k = 0$. Second, if G is a finitely generated abelian group, and no element of G has finite order, then $G \cong \mathbb{Z}^k$ for some k. A slightly simpler statement of a part of this theorem is that if G is any finite abelian group, then

$$G \cong \mathbb{Z}/k_1\mathbb{Z} \times \mathbb{Z}/k_2\mathbb{Z} \times \cdots \times \mathbb{Z}/k_m\mathbb{Z}$$

for some nonnegative integers $k_1, \ldots, k_m$. However, this decomposition is not unique. This is the reason for the further decomposition into cyclic groups with prime power orders.

The least elementary part of this proof is when G is infinite, so that $k > 0$. The first step in the general case is to show that there exists some k so that $G = \mathbb{Z}^k \times G'$, where

G' is a finite abelian group. Thus, for simplicity, we shall restrict to this latter case. The proof is then accessible to you at this stage, though it takes enough space and time that we will not prove it in full detail here. (One of the standard proofs is based on the *Smith algorithm* and *Smith normal form*, which we discuss in Chapter 13.) However, let us describe some of the ideas in the proof.

Let G be a finite abelian group. The key idea is to first find a largest cyclic subgroup $\mathbb{Z}/k_1\mathbb{Z} \subset G$, i.e. for which k_1 is as large as possible. To see why this is possible, note that if $g \in G$ is arbitrary, then g generates a cyclic group of finite order in G (because G is finite), so we simply need to choose the element g for which the size of this cyclic group is maximal. The main step is to show that if H_1 denotes this particular cyclic subgroup, then there exists another subgroup $G_1 \subset G$ such that $G \cong H_1 \times G_1$. A decomposition like this is very far from true for an arbitrary subgroup, even when G is finite abelian, so this is nontrivial. It must now be checked that G_1 is again finite abelian, which is easy. (Why?) Now we repeat this process. Namely, let us find a largest cyclic subgroup $\mathbb{Z}/k_2\mathbb{Z} \subset G_1$ and then (by the same argument as before) write $G_1 \cong \mathbb{Z}/k_2\mathbb{Z} \times G_2$. If we continue in this way, we must reach our conclusion in finitely many steps. The reason is that the size (or order) of the successive subgroups G_j gets smaller and smaller, and because the initial G was finite, we must reach a subgroup G_{m+1} of order 1 eventually, so that $G_{m+1} = \{e\}$. This means that the subgroup reached at the previous stage, G_m, was already cyclic.

Once we have written G as a finite product of cyclic groups, we must continue to break down these cyclic groups into products of cyclic groups of prime power order. Finally, we must also prove that once we have ensured that the cyclic factors are of this special form, the decomposition is unique up to rearrangement of the factors. ■

The last step of this proof involves a sequence of ideas that includes the *Chinese Remainder Theorem* and an interesting use of a group isomorphism. We'll conclude this chapter by presenting these ideas.

Theorem 6.15 (Chinese Remainder Theorem) *Suppose $n_1, \ldots, n_k$ are positive integers that are* coprime in pairs. *This means that the greatest common divisor of any n_i, n_j with $i \neq j$ is equal to one. Then for any given integers $a_1, \ldots, a_k$, it is possible to find an integer x that solves all of the following congruences simultaneously:*

$$\begin{aligned} x &\equiv a_1 \pmod{n_1} \\ &\vdots \\ x &\equiv a_k \pmod{n_k}\,. \end{aligned}$$

Moreover, all solutions of these equations differ by a multiple of $N = n_1 n_2 \cdots n_k$.

Proof We know that n_i and N/n_i are coprime integers for every i. Therefore we can find r_i, s_i so that $r_i n_i + s_i N/n_i = 1$ by definition of the greatest common divisor. (The proof of this fact is the so-called "Euclidean algorithm.") Let $e_i = s_i N/n_i$. Then $e_i = 1 \pmod{n_i}$ and $e_i = 0 \pmod{n_j}$ for all $j \neq i$. (This is because N/n_i is divisible by all other n_j.) Now let $x = \sum_i a_i e_i$. The first part of the theorem follows

by the homomorphism property of congruence modulo an integer, meaning $x + y \pmod{n} = x \pmod{n} + y \pmod{n}$. Finally, if x and y are two solutions of the equations above, then $x - y$ is congruent to zero modulo n_1 through n_k. Therefore $x - y$ is divisible by the product of the n_i, namely N. ■

Corollary 6.16 *Suppose the integer N can be decomposed into powers of prime numbers as $N = p_1^{\ell_1} \cdots p_k^{\ell_k}$. Then $\mathbb{Z}/N\mathbb{Z} \cong \mathbb{Z}/p_1^{\ell_1}\mathbb{Z} \times \cdots \times \mathbb{Z}/p_k^{\ell_k}\mathbb{Z}$.*

Proof Let $n_i = p_i^{\ell_i}$, and define a homomorphism $\phi : \mathbb{Z}/n\mathbb{Z} \to \mathbb{Z}/n_1\mathbb{Z} \times \cdots \times \mathbb{Z}/n_k\mathbb{Z}$ by

$$\phi(x) = \big(x \pmod{n_1}\ ,\ \ldots\ ,\ x \pmod{n_k}\big)\,.$$

(Why is ϕ a homomorphism?) Since $n_1, \ldots, n_k$ are coprime in pairs and $N = n_1 \cdots n_k$, the Chinese Remainder Theorem applies, and for every set of integers $a_1, \ldots, a_k$ we can find $x \in \mathbb{Z}/N\mathbb{Z}$ (why can we say $x \in \mathbb{Z}/N\mathbb{Z}$ and not merely $x \in \mathbb{Z}$?) such that $\phi(x) = (a_1, \ldots, a_k)$. In other words, ϕ is surjective. Finally, suppose $\phi(x) = 0$. Then $x \pmod{n_i} = 0$ for each i, and so x is in fact a multiple of N. Therefore $x = 0 \pmod{N}$. Therefore $x = 0$ in $\mathbb{Z}/N\mathbb{Z}$, and we have now shown that $\ker(\phi)$ is trivial. Thus ϕ is injective. ■

6.7 Problems

(1) Let n, m be positive integers. Show that $S_n \times S_m$ is a subgroup of S_{n+m} in a natural way.

(2) Describe carefully how the dihedral group D_3 can be regarded as a subgroup of S_6, in the most interesting way possible.

(3) A *cyclic group* is a group G generated by one element; i.e. $G = \{e, a^{\pm 1}, a^{\pm 2}, \ldots\}$. The group G is *finite cyclic* having *order* $n \in \mathbb{N}$ if n is the smallest positive integer such that $a^n = e$, and it is *infinite cyclic* otherwise.

   (a) Prove that every cyclic group is abelian.
   (b) Show that in a cyclic group G of order n generated by an element a, every $g \in G$ can be expressed uniquely in the form $g = a^i$, where $0 \leq i \leq n - 1$.
   (c) Show that every cyclic group must be isomorphic to either $\mathbb{Z}/n\mathbb{Z}$ for some $n \in \mathbb{N}$, or to $\mathbb{Z}$.

(4) Define $\pi : S_n \to \mathbb{Z}/2\mathbb{Z}$ by

$$\pi(\sigma) := [\text{number of 2-cycles making up } \sigma] \mod 2.$$

Show that π is well-defined (meaning that if σ can be written in two different ways as a product of 2-cycles, then their number (mod 2) is the same) and is a group homomorphism. The kernel of π consists of all permutations that can be decomposed into an even number of cycles of length 2, and is called the *alternating group* and denoted A_n. Find A_2, A_3, and A_4.

(5) Let G be a group and $g_1, \ldots, g_k \in G$. Then the subgroup generated by $g_1, \ldots, g_k$ is denoted by $\langle g_1, \ldots, g_k \rangle$ and is the smallest subgroup of G containing $g_1, \ldots, g_k$.

(a) Determine the subgroup generated by (123), (134), (234), and (124) in S_4.
(b) Determine the subgroup generated by (12), (23), and (34) in S_4.

(6) (a) Write down all homomorphisms $\mathbb{Z}/2\mathbb{Z} \to \mathbb{Z}/6\mathbb{Z}$.
(b) Write down all homomorphisms $\mathbb{Z}/3\mathbb{Z} \to \mathbb{Z}/6\mathbb{Z}$.
(c) Write down all homomorphisms $\mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/3\mathbb{Z} \to \mathbb{Z}/6\mathbb{Z}$.
(d) Which of the homomorphisms from part (c) are isomorphisms?

(7) Let $f : \mathbb{Z}/mn\mathbb{Z} \to \mathbb{Z}/m\mathbb{Z} \times \mathbb{Z}/n\mathbb{Z}$ be a homomorphism defined by $f(1) := (1, 1)$. Show f is an isomorphism if and only if m and n are relatively prime. (This explains how $\mathbb{Z}/6\mathbb{Z} \cong \mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/3\mathbb{Z}$.)

(8) (a) How many distinct abelian groups containing exactly 36 elements are there?
(b) Explain why $\mathbb{Z}/8\mathbb{Z}$, $\mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/4\mathbb{Z}$ and $\mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/2\mathbb{Z}$ are non-isomorphic groups.

(9) (a) Let G_1 and G_2 be groups containing subgroups H_1 and H_2, respectively. Show that $H_1 \times H_2$ is a subgroup of $G_1 \times G_2$.
(b) If G_1 is generated by g_1 and G_2 is generated by g_2, show that $G_1 \times G_2$ is generated by (g_1, e_{G_2}) and (e_{G_1}, g_2). Are there other ways of generating $G_1 \times G_2$? Under what circumstances is it possible to generate $G_1 \times G_2$ by only one element?
(c) Can you find a pair of groups G_1 and G_2 and a subgroup of $G_1 \times G_2$ that is not of the form $H_1 \times H_2$ for some subgroups $H_i \leq G_i$?

Chapter 7
Cosets, Normal Subgroups, and Quotient Groups

7.1 Cosets

There are several very important constructions that one can obtain from a group G and a subgroup H. Before we define them formally, let us look at an example.

Let $G = S_n$ be the symmetric group, consisting of all permutations of the elements of $\mathcal{X}_n$. Suppose we have some element $\sigma \in S_n$, but for some reason we can only tell what $\sigma(1)$ and $\sigma(2)$ are; in other words, we are only able to see some of the information that σ possesses. Now, this information doesn't allow us to recover σ completely, but it does cut down on the possibilities for what σ might be. Let's start with a warmup: let H be the set of elements of S_n that have the property that $\sigma(1) = 1$ and $\sigma(2) = 2$.

Example Show that $H \leq G$, i.e. that H is a subgroup.

Okay, subgroups we already understand, so let's move on to a different sort of subset. Let S denote the subset of the elements of S_n that have the property that $\sigma(1) = 4$ and $\sigma(2) = 3$. What structure does S have?

Example Show that S is *not* a subgroup of G.

Nonetheless, S is still interesting and has some relevant structure. Here is an example of structure possessed by S: suppose that $\tau \in H$. Now choose any $\sigma \in S$ and observe that $\sigma\tau$ is also in S, since for example $\sigma\tau(1) = \sigma(1) = 4$ and $\sigma\tau(2) = \sigma(2) = 3$. Another example of structure possessed by S is that if σ_1 and σ_2 are both in S, then there is some $\tau \in H$ with $\sigma_1\tau = \sigma_2$. In fact, we have $\tau = \sigma_1^{-1}\sigma_2$.

We can formalize the observations above as follows: H and S have the property that $SH = S$, meaning that if $\sigma \in S$ and $\tau \in H$, then $\sigma\tau \in S$. Alternatively, if $\sigma \in S$, then $S = \sigma H$. Finally, every element in S can be written as $\sigma\tau$ for some $\sigma \in S$ and $\tau \in H$. Here, S is a typical example of something called a *coset*.

Definition 7.1 Let G be a group and $H \leq G$ a subgroup. Also, let $g \in G$ be some element. Then the set

C. Bray et al., *Algebraic Topology*,
https://doi.org/10.1007/978-3-030-70608-1_7

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

is called a *left coset* of H. Similarly,

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

is a *right coset* of H.

Remark 7.2 No one can ever remember which is a left coset and which is a right coset. It is, in general, only important to remember that these are different. In the rest of this chapter, we will do some constructions for left cosets only; all these things will work for right cosets as well, and vice versa.

Example The introductory discussion suggests an example in which left and right cosets differ. Let $G = S_4$, and let H be the subgroup of G consisting of those σ for which $\sigma(1) = 1$ and $\sigma(2) = 2$. Let $g = (23)$. Then gH consists of all the elements $\tau \in S_4$ for which $\tau(1) = 1$ and $\tau(2) = 3$. On the other hand, Hg consists of all the elements $\tau \in S_4$ for which $\tau(1) = 1$ and $\tau(3) = 2$. Note that these are different subsets of G!

Example Let $G = \mathbb{Z}$ be the group of integers under addition and let $H = 3\mathbb{Z} = \{\ldots, -6, -3, 0, 3, 6, \ldots\}$ be the subgroup of all multiples of three. Since addition is the operation in $\mathbb{Z}$, we adapt the notation for cosets accordingly: $k + 3\mathbb{Z} = \{\ldots, k - 6, k - 3, k, k + 3, k + 6, \ldots\}$ is the coset of $k \in \mathbb{Z}$.

Remark 7.3 We will soon see a condition that allows us to conclude that the right and left cosets gH and Hg do in fact coincide.

Let us look at some basic properties of cosets.

Theorem 7.4 *Let G be a group and $H \leq G$. Let g_1H and g_2H be two cosets of H. Then g_1H and g_2H are either equal or disjoint, i.e., $g_1H = g_2H$ or else $g_1H \cap g_2H = \varnothing$.*

Proof Suppose that g_1H and g_2H are not disjoint, so that there is an element $x \in g_1H \cap g_2H$. Then there are $h_1, h_2 \in H$ so that $x = g_1h_1 = g_2h_2$. We can then solve for g_1 in this equation, so that $g_1 = g_2h_2h_1^{-1}$. Now, we show that $g_1H \subset g_2H$. Let $g_1h \in g_1H$. Then we have

$$g_1h = (g_2h_2h_1^{-1})h = g_2(h_2h_1^{-1}h) \in g_2H.$$

Since g_1h was an arbitrary element of g_1H, we have $g_1H \subset g_2H$. By symmetry, we also have $g_1H \supset g_2H$. Hence $g_1H = g_2H$. ■

Example We can see this theorem at work in the example $k + 3\mathbb{Z}$ inside $\mathbb{Z}$ introduced above. There are in fact exactly three distinct cosets: $0 + 3\mathbb{Z} = 3\mathbb{Z}$ and $1 + 3\mathbb{Z} = \{\ldots, -5, -2, 1, 4, 7, \ldots\}$ and $2 + 3\mathbb{Z} = \{-4, -1, 2, 5, 8, \ldots\}$. Every other coset coincides with one of these three cosets. For example, $3 + 3\mathbb{Z} = 3\mathbb{Z}$ and $10 + 3\mathbb{Z} = 1 + 9 + 3\mathbb{Z} = 1 + 3\mathbb{Z}$.

It is useful to be able to distinguish between these two possibilities: Given two cosets g_1H and g_2H, how can we tell whether they are equal or disjoint? The following proposition answers this question:

Proposition 7.5 *Let g_1H and g_2H be two cosets of H. Then $g_1H = g_2H$ if and only if $g_1^{-1}g_2 \in H$.*

Proof Since $e \in H$, we know that $g_2 \in g_2H$. So, by Theorem 7.4, we know that $g_1H = g_2H$ if and only if

$$g_2 \in g_1H. \tag{7.1}$$

Now, we can multiply (7.1) on the left by g_1^{-1} to obtain $g_1^{-1}g_2 \in H$. In other words, we have shown that $g_1H = g_2H$ if and only if $g_1^{-1}g_2 \in H$, which is what we wanted. ■

Example Again, we can see this proposition at work in the $k + 3\mathbb{Z}$ example. Consider another coset $\ell + 3\mathbb{Z}$. A typical element of this coset has the form $\ell + 3n$ for some integer n. We can find this element inside $k + 3\mathbb{Z}$ if and only if $\ell + 3n$ can be written as $k + 3m$ for some integer m. Hence $\ell + 3n = k + 3m$ if and only if $\ell - k = 3(m - n)$, or in other words $\ell - k \in 3\mathbb{Z}$. Also, in the example above, $10 - 1 = 9 = 3 \times 3$, so that $1 + 3\mathbb{Z} = 10 + 3\mathbb{Z}$ as we have observed.

Proposition 7.6 *Let G be a* finite *group and H a subgroup. Then any two cosets of H have the same size.*

Proof Let g_1H and g_2H be two cosets (which could be the same). We will find a bijection between the elements of g_1H and g_2H. A typical element of g_1H has the form g_1h. We define a function $\varphi : g_1H \to g_2H$ by setting $\varphi(g_1h) = g_2h$.

In order to check that φ is a bijection, we find an inverse function $\psi : g_2H \to g_1H$ so that the compositions $\varphi \circ \psi$ and $\psi \circ \varphi$ are the identities on their respective cosets. We define ψ by setting $\psi(g_2h) = g_1h$. It is easy to check that φ and ψ are inverses of each other. Hence there is a bijection between g_1H and g_2H, so these two cosets have the same number of elements. ■

Since every element $g \in G$ is in the coset gH, the union of all the cosets of H is equal to all of G. Consequently we have this picture: The group G can be partitioned into a collection of sets, the cosets of H in G, and these sets are disjoint from each other. If this reminds you of equivalence relations, that's because we can indeed rephrase some of what we have done in the language of equivalence relations!

Theorem 7.7 *The relation $\sim$ on G, described by $g_1 \sim g_2$ if and only if $g_1 \in g_2H$, is an equivalence relation on G.*

Proof We must show that $\sim$ is reflexive, symmetric, and transitive. There are many ways of doing this, so let us demonstrate a few different techniques for the different parts.

Reflexivity: We need to show that $g \in gH$. But this is clear, because $e \in H$.

Symmetry: Suppose $g_1 \in g_2H$. Then we can write $g_1 = g_2h$ for some $h \in H$. Hence $g_2 = g_1h^{-1}$. But since H is a subgroup, $h^{-1} \in H$, so $g_2 \in g_1H$.

Transitivity: Suppose $g_1 \in g_2H$ and $g_2 \in g_3H$. By Theorem 7.4, since $g_1H \cap g_2H \neq \varnothing$, we must have $g_1H = g_2H$. Similarly, since $g_2H \cap g_3H \neq \varnothing$, we must again have $g_2H = g_3H$. Hence $g_1H = g_3H$, so in particular $g_1 \in g_3H$.

■

7.2 Lagrange's Theorem and Its Consequences

Cosets offer us a convenient way of proving what is probably the first interesting theorem in finite group theory.

Theorem 7.8 (Lagrange) *Let G be a finite group of order n, and let $H \leq G$ be a subgroup of order m. Then m divides n.*

Proof One way of proving that the size of one set S divides the size of some other set T is to divide T into several disjoint subsets, each of the same size as S, in such a way that each $t \in T$ is contained in exactly one of these sets. Cosets provide a natural way of doing so in this case: G is the union of the cosets gH, and by Proposition 7.6, they all have the same size as each other, and hence as H. Thus m divides n. ■

Corollary 7.9 *The order of any element of G divides the order of G.*

Proof Let $g \in G$ be any element, and let $H = \langle g \rangle$ be the subgroup generated by g. Then apply Lagrange's Theorem on G and H. ■

This Corollary offers us a simple way of proving an important result in number theory.

Theorem 7.10 (Fermat's Little Theorem) *Let p be a prime, and let a be an integer not divisible by p. Then $a^{p-1} \equiv 1 \pmod{p}$.*

Proof The nonzero elements of $\mathbb{Z}/p\mathbb{Z}$ form a group denoted $(\mathbb{Z}/p\mathbb{Z})^\times$ of order $p-1$. Since a is not divisible by p, it is represented by some element of $(\mathbb{Z}/p\mathbb{Z})^\times$, say b. Let m be the order of b in $(\mathbb{Z}/p\mathbb{Z})^\times$. By Corollary 7.9, m divides $p-1$, so $b^{p-1} = e$ in $(\mathbb{Z}/p\mathbb{Z})^\times$, which means that $a^{p-1} \equiv 1 \pmod{p}$. ■

Exercise 7.11 Modify this proof to prove Euler's Theorem: If a is relatively prime to n, then $a^{\phi(n)} \equiv 1 \pmod{n}$. Here ϕ is the so-called *totient function*, which counts the number of positive integers less than or equal to n and relatively prime to n.

Remark 7.12 It is tempting to suspect that the order of any element of S_n will be at most n. However, Lagrange's Theorem only tells us that an element has order dividing $n!$, and indeed we can easily come up with examples of elements of S_n whose orders are larger than n. For example, in S_5, the element $(12)(345)$ has order 6, which you can verify. (Exercise: What is the largest possible order of an element of S_{20}?)

7.3 Coset Spaces and Quotient Groups

In this section we will consider an abstract mathematical object: the *set of all cosets*. That is to say, we take a group G and a subgroup H, and consider all possible cosets gH. Then we collect them into a set! Let's call it $\mathcal{S} := \{gH : g \in G\}$. The elements of this set are the various cosets gH, and of course each coset is a subset of G—so $\mathcal{S}$ is a set of sets!

We would like to understand if $\mathcal{S}$ possesses any interesting mathematical structure. A natural question to ask is if it is possible to define a *multiplication* in $\mathcal{S}$, i.e. if it is possible to multiply two cosets. Let's see how this might work. Let g_1H and g_2H be two cosets. What could the product $g_1H \cdot g_2H$ be? Presumably it must be the set

$$\begin{aligned} g_1H \cdot g_2H &= \{k_1k_2 : k_1 \in g_1H \text{ and } k_2 \in g_2H\} \\ &= \{g_1hg_2h' : h, h' \in H\}. \end{aligned}$$

If the operation $\cdot$ on cosets defined above is to make any sense as an operation on $\mathcal{S}$, then it must be the case that $g_1H \cdot g_2H$ is itself a coset of H in G. So we would need to find $g \in G$ so that $g_1hg_2h' = gh''$ for some $h'' \in H$, no matter what choice of g_1, g_2, h, h' we start with. Here is a condition which guarantees that this will happen.

Definition 7.13 A subgroup H of a group G is said to be *normal* if, for any $g \in G$ and $h \in H$, $g^{-1}hg \in H$. When H is a normal subgroup of G, we write $H \lhd G$.

Exercise 7.14 The condition for normality is sometimes written as $g^{-1}Hg = H$, or as $gH = Hg$. Show that these are equivalent to the definition above. So we find that the right and left cosets of H in G agree exactly when H is a normal subgroup of G!

As a consequence, if $H \lhd G$ then we have $hg_2 = g_2\tilde{h}$ for some potentially different element $\tilde{h} \in H$. Thus $g_1hg_2h' = g_1g_2\tilde{h}h' = g_1g_2h''$, where $h'' = \tilde{h}h' \in H$ because H is a subgroup. Therefore the product of the coset g_1H and g_2H is unambiguously the coset g_1g_2H.

Example The subgroup $3\mathbb{Z}$ is normal in $\mathbb{Z}$ because $\mathbb{Z}$ is abelian, and all subgroups of abelian groups are normal (a fact we will prove in the next section, in Proposition 7.23). Now we can add cosets: for example $(1 + 3\mathbb{Z}) + (2 + 3\mathbb{Z}) = (1 + 2) + 3\mathbb{Z} = 3 + 3\mathbb{Z} = 3\mathbb{Z}$ and so on. Note that there are only three cosets in $\mathcal{S}$ in this case: $3\mathbb{Z}$, $1 + 3\mathbb{Z}$, $2 + 3\mathbb{Z}$, and that they form a group of order three possessing the same multiplicative properties as the group $\mathbb{Z}/3\mathbb{Z}$ of integers modulo three. This is no accident!

We formalize the above discussion with the following theorem.

Theorem 7.15 *If $H \lhd G$, then the space of cosets of H in G forms a group. We call this group the quotient group of G by H, and we write it as G/H.*

Proof We have endowed G/H (which used to be called $\mathcal{S}$ above) with a binary operation, and it remains for us to show that it is actually a *group* operation. To do so, we must establish the three required properties of this operation. First, the operation is associative because multiplication in G is. Second, the identity is the coset $H = eH$ itself. Third, the inverse of the coset gH is the coset $g^{-1}H$. ■

Remark 7.16 This is the origin of the possibly mysterious notation $\mathbb{Z}/n\mathbb{Z}$ for the integers modulo n: it is the group of cosets of $n\mathbb{Z}$ in $\mathbb{Z}$.

Remark 7.17 Even when H is not a normal subgroup of G, the space of left cosets G/H is sometimes still a useful object. It has the structure of a *pointed set*, i.e. a set together with a distinguished element. The distinguished element is the coset containing the identity element.

Definition 7.18 The number of cosets of H in G is called the *index* of H in G. We write this number as $[G : H]$.

Example

- If G is any group, then the trivial subgroup $\{e\}$ and the entire group G are both normal subgroups of G. We call any other normal subgroup a *nontrivial* normal subgroup.
- $[\mathbb{Z} : n\mathbb{Z}] = n$.
- $[S_n : S_{n-1}] = n$, where we think of S_{n-1} as being the subgroup of S_n consisting of all $\sigma \in S_n$ with $\sigma(n) = n$.
- Let $G = S_n$ and H be the subgroup in the first section of this chapter, consisting of those σ for which $\sigma(1) = 1$ and $\sigma(2) = 2$. Then $[G : H] = n(n-1)$.

When $H \lhd G$, then we have $[G : H] = |G/H|$, the size of the quotient group. Note that, if G is a *finite* group, then there are $[G : H]$ cosets, each with $|H|$ elements. Since every element of G is contained in a unique coset, we have

$$|G| = [G : H]|H|.$$

Note the relationship with Lagrange's Theorem.

7.4 Properties and Examples of Normal Subgroups

Now that we have seen the importance of normal subgroups, we would like to have a natural source for them. We will in fact be able to characterize normal subgroups completely. We begin with a key result.

Theorem 7.19 *Let $\phi : G \to G'$ be a homomorphism between groups. Then* $\ker(\phi) \lhd G$.

Proof Suppose $g \in G$ and $h \in H = \ker(\phi)$. Then we must show that $g^{-1}hg \in H$ as well. We have

$$\begin{aligned}\phi(g^{-1}hg) &= \phi(g^{-1})\phi(h)\phi(g)\\ &= \phi(g)^{-1}\phi(h)\phi(g)\\ &= \phi(g)^{-1}\phi(g)\\ &= e_H,\end{aligned}$$

so $g^{-1}hg \in \ker(\phi)$. ■

We can extend this result as follows.

Theorem 7.20 *A subgroup $H \leq G$ is normal if and only if there exists a group K and a homomorphism $\phi : G \to K$ so that $H = \ker(\phi)$.*

Proof Suppose that $H \lhd G$. We need to find a group K and a homomorphism $\phi : G \to K$ so that $H = \ker(\phi)$. We can take $K = G/H$ and define a map ϕ that sends g to gH. We now show that ϕ is a homomorphism: $\phi(g_1g_2) = g_1g_2H = g_1H \cdot g_2H = \phi(g_1) \cdot \phi(g_2)$. Now $g \in \ker(\phi)$ if and only if $gH = H$, i.e. if $g \in H$. Hence $H = \ker(\phi)$. ■

Definition 7.21 We call the map $G \to G/H$ a *quotient map* or a *canonical projection*.

Remark 7.22 The quotient map $G \to G/H$ is very important and will be used all over the place for the rest of your life. We think of this homomorphism as remembering certain information and forgetting other information. For example, the homomorphism $\mathbb{Z} \to \mathbb{Z}/3\mathbb{Z}$ remembers the remainder when some number n is divided by 3, but it forgets what the actual number was. We can describe any such quotient map in a similar manner.

There are some more specific results that give us conditions under which subgroups are normal. We give two such results, the first of which is obvious.

Proposition 7.23 *If G is abelian and $H \leq G$, then $H \lhd G$.*

Proof If $g \in G$ and $h \in H$, then $g^{-1}hg = h \in H$. ■

Theorem 7.24 *If $H \leq G$ and $[G : H] = 2$, then $H \lhd G$.*

Proof Let $g \in G$ and $h \in H$, where $[G : H] = 2$. Since there are only two cosets, and the cosets cover all of G, either $g^{-1}hg \in H$, or else it's in the other coset. Let us break the problem down into two cases:

Case 1: $g \in H$. In this case, g^{-1}, h, and g are all in H, so their product is as well.

Case 2: $g \notin H$. Suppose that $g \notin H$, and that $g^{-1}hg = g' \notin H$. Then $g(g')^{-1} \in H$ (because there are only two cosets). We can then rewrite the equation $g^{-1}hg = g'$ as $g = hg(g')^{-1} = h(g(g')^{-1})$, which is the product of two elements of H. Hence $g \in H$ as well, contradicting our assumption. ■

Remark 7.25 In fact, we can do better. If G is a *finite* group, p is the smallest prime dividing $|G|$, and $H \leq G$ is a subgroup with $[G : H] = p$, then $H \lhd G$. However, the proof of this is a little bit harder.

Example Since $[S_n : A_n] = 2$, $A_n \lhd S_n$. The quotient is $S_n/A_n \cong \mathbb{Z}/2\mathbb{Z}$.

7.5 Coset Representatives

The proper way of thinking of the quotient group G/H is as the set of cosets gH. However, this is a bit unwieldy at times: for example, we like to think of $\mathbb{Z}/3\mathbb{Z}$ as $\{0, 1, 2\}$ with a suitable addition law, and not as $\{3\mathbb{Z}, 1 + 3\mathbb{Z}, 2 + 3\mathbb{Z}\}$. We can do something similar in general, as follows:

Definition 7.26 Let $H \lhd G$, and let G/H be the quotient group. Let $\mathcal{A} = \{a_i : i \in I\} \subset G$ be a set of elements with the following property: for every $g \in G$, there is a unique $i \in I$ for which $g \in a_i H$. Then we say that $\mathcal{A}$ is a *set of coset representatives* for H in G.

Note that coset representatives are not unique. For example, we can take $\{0, 1, 2\}$ to be a set of coset representatives for $\mathbb{Z}/3\mathbb{Z}$, but we could also take $\{36, -11, 5\}$. In general, there is no "preferred" choice of coset representatives: any choice works equally well.

It is also worth noting that coset representatives do not usually form a group themselves—although they occasionally do, in exceptional circumstances, and it says something interesting when it does happen. For example $\mathcal{A} = \{0, 1, 2\} \subset \mathbb{Z}$ does not form a group, because $1 + 2 = 3 \notin \mathcal{A}$.

Remark 7.27 One might wonder whether it is *possible* to find a set of coset representatives for G/H that do form a group. In general, the answer is no: for example, there is no set of coset representatives for $\mathbb{Z}/3\mathbb{Z}$ in $\mathbb{Z}$ which form a group. But we can state a precise condition that allows us to find such a set of representatives. Let $p : G \to G/H$ be the canonical projection. Then we can find a set of coset representatives $\mathcal{A} \subset G$ that form a group if and only if there is a homomorphism $\iota : G/H \to G$ so that the composition $p \circ \iota : G/H \to G/H$ is the identity map; we call ι a *section* of p. If we have such a section, then $\mathrm{im}(\iota)$ is a set of coset representatives that forms a group.

7.6 A Quotient of a Dihedral Group

In this section, we will look carefully at an example of a normal subgroup and the corresponding quotient group of the dihedral group D_3. Recall that D_3 has 6 elements,

$$e, \rho, \rho^2, \sigma, \rho\sigma, \rho^2\sigma,$$

where ρ denotes a counterclockwise rotation by $2\pi/3$, and σ denotes a reflection about the y-axis. Let us recall the multiplication table for D_3.

D_3	e	ρ	ρ^2	σ	$\rho\sigma$	$\rho^2\sigma$
e	e	ρ	ρ^2	σ	$\rho\sigma$	$\rho^2\sigma$
ρ	ρ	ρ^2	e	$\rho\sigma$	$\rho^2\sigma$	σ
ρ^2	ρ^2	e	ρ	$\rho^2\sigma$	σ	$\rho\sigma$
σ	σ	$\rho^2\sigma$	$\rho\sigma$	e	ρ^2	ρ
$\rho\sigma$	$\rho\sigma$	σ	$\rho^2\sigma$	ρ	e	ρ^2
$\rho^2\sigma$	$\rho^2\sigma$	$\rho\sigma$	σ	ρ^2	ρ	e

D_3 has a normal subgroup H, which consists of the elements $\{e, \rho, \rho^2\}$. We could verify this directly by writing out a multiplication table, but we can do it more directly with some geometric thinking. We need to check that, for every $g \in D_3$ and $h \in H$, $g^{-1}hg \in H$. If $g \in H$, then we're multiplying together three elements in H, so the result is still in H. If $g \notin H$, then g contains a reflection and hence reverses orientations. But g^{-1} also contains a reflection, so it also reverses orientation. But h doesn't reverse orientation, so $g^{-1}hg$ reverses orientation exactly twice. If we reverse orientation an even number of times, then we have preserved the original orientation. The only elements of D_3 that preserve orientation are the rotations e, ρ, ρ^2. Hence $g^{-1}hg \in H$. (More abstractly, this follows from Theorem 7.24.)

So, now we understand that D_3 has a normal subgroup H of order 3. What is the quotient group? Since the order of the quotient group D_3/H is equal to the order of D_3 divided by that of H, we know that D_3/H has order 2 and thus must be isomorphic to $\mathbb{Z}/2\mathbb{Z}$. But let us work this out more explicitly. Let us make a multiplication table for the cosets.

D_3/H	$\{e, \rho, \rho^2\}$	$\{\sigma, \rho\sigma, \rho^2\sigma\}$
$\{e, \rho, \rho^2\}$	$\{e, \rho, \rho^2\}$	$\{\sigma, \rho\sigma, \rho^2\sigma\}$
$\{\sigma, \rho\sigma, \rho^2\sigma\}$	$\{\sigma, \rho\sigma, \rho^2\sigma\}$	$\{e, \rho, \rho^2\}$

We can write this multiplication table in a less cluttered form, as follows.

D_3/H	H	$H\sigma$
H	H	$H\sigma$
$H\sigma$	$H\sigma$	H

We can now write down an isomorphism $\phi : D_3/H \to \mathbb{Z}/2\mathbb{Z}$ as follows: let $\phi(\{e, \rho, \rho^2\}) = \phi(H) = 0$ and $\phi(\{\sigma, \rho\sigma, \rho^2\sigma\}) = \phi(H\sigma) = 1$. (Exercise: Verify that this is actually an isomorphism.)

7.7 Building up Finite Groups

One reason we find normal subgroups to be particularly useful is that they allow us to break a complicated group into less complicated pieces. That is, if $H \lhd G$, then G is somehow "built up" of the smaller groups H and G/H. These groups can be glued

in some way to reconstruct G. Thus, if we want to understand all (finite) groups, then a good starting point is to understand the basic building blocks—those groups that have no nontrivial normal subgroups.

Definition 7.28 We say a nontrivial group G is *simple* if its only normal subgroups are the trivial subgroup and the entire group G.

Example

- If p is a prime, then $\mathbb{Z}/p\mathbb{Z}$ is a simple group.
- If $n \geq 5$, then the alternating group A_n is simple. (This is not obvious, or especially easy. For a proof, see [Rot95, Chapter 3].)

One of the most remarkable achievements of twentieth-century mathematics was to give a complete classification of the finite simple groups. This was achieved over the course of hundreds of papers, spanning more than 10000 pages of difficult mathematics. Here are some of the highlights of that program:

Theorem 7.29 (Feit–Thompson [FT63]) *If G is a simple group and $|G|$ is odd, then $G \cong \mathbb{Z}/p\mathbb{Z}$ for some odd prime p.*

Theorem 7.30 (Classification of Finite Simple Groups) *The finite simple groups fall into 18 explicitly described infinite families, plus 26 extra "sporadic" groups.*

See [Wil09] for a book all about the simple groups and their descriptions.

Fortunately, however, this is not the end of the story. (We say "fortunately" because it is always a good thing to have more fascinating problems to work on!) The Classification of Finite Simple Groups tells us what all the building blocks are, but we still don't understand the glue used to stick them together perfectly. In general, if we know what H and G/H are, there are still several possibilities for G.

Example Suppose that we know that $H = \mathbb{Z}/2\mathbb{Z}$ and $G/H = \mathbb{Z}/2\mathbb{Z}$. What can G be? It turns out that G can be either $\mathbb{Z}/4\mathbb{Z}$ or $(\mathbb{Z}/2\mathbb{Z}) \times (\mathbb{Z}/2\mathbb{Z})$. Let us spell this out explicitly. If $G = \mathbb{Z}/4\mathbb{Z}$, which we'll think of as being the elements $\{0, 1, 2, 3\}$, and we let $H = \{0, 2\}$, then G/H is represented by $\{0, 1\}$, so it is isomorphic to $\mathbb{Z}/2\mathbb{Z}$. On the other hand, suppose $G = (\mathbb{Z}/2\mathbb{Z}) \times (\mathbb{Z}/2\mathbb{Z})$, which we'll write as the elements $\{(0, 0), (0, 1), (1, 0), (1, 1)\}$. Let $H = \{(0, 0), (0, 1)\}$. Then G/H is represented by the elements $\{(0, 0), (1, 0)\}$, which is again isomorphic to $\mathbb{Z}/2\mathbb{Z}$.

There are techniques available to tell us the various ways we can glue together H and G/H, but this problem has not been solved in general. Nor is it ever likely to be solved. For instance, every group of order $1024 = 2^{10}$ can be built up out of 10 copies of $\mathbb{Z}/2\mathbb{Z}$, but there are 49487365422 of them (up to isomorphism). See [BEO02] for a discussion on finding all groups of a given order, or just counting them. Even the problem of listing all groups of order 16 is an interesting challenge, but an elementary discussion can be found in [Wil05].

7.8 An Isomorphism Theorem

One of the most frequently used results in group theory—and abstract algebra, in general—is the following result, which relates the kernel and the image of a homomorphism. It is sometimes called the first isomorphism theorem, and sometimes the second, but it is by far the most important of all the "isomorphism theorems."

Theorem 7.31 *Let $\phi : G \to H$ be a homomorphism. Then $G/\ker(\phi) \cong \mathrm{im}(\phi)$.*

Proof Let $K = \ker(\phi)$. First, we will come up with a homomorphism $\psi : G/K \to \mathrm{im}(\phi)$. The natural choice is to try to define $\psi(gK) = \phi(g)$. However, this might not make sense, because it might be the case that $gK = g'K$, but $\phi(g) \neq \phi(g')$. So, let us check that this does not happen, i.e. that if $g(g')^{-1} \in K$, then $\phi(g) = \phi(g')$. If we can verify this, then we'll know that we have a well-defined map ψ. So, let us suppose that $g(g')^{-1} = k \in K$. Then we have

$$e = \phi(k) = \phi(g(g')^{-1}) = \phi(g)\phi(g')^{-1},$$

so $\phi(g') = \phi(g)$, as desired. Let us now verify that this map is indeed a homomorphism. We have

$$\psi(gKg'K) = \psi(gg'K) = \phi(gg') = \phi(g)\phi(g') = \psi(gK)\psi(g'K),$$

as desired.

To conclude the proof of the theorem, we must check that ψ is an isomorphism, so we need to check that it is both injective and surjective. Let us check that it is injective. Suppose that $\psi(gK) = \psi(g'K)$. We have $\psi(gK) = \phi(g)$ and $\psi(g'K) = \phi(g')$, so $\phi(g) = \phi(g')$. Hence $\phi(g(g')^{-1}) = e$, so $g(g')^{-1} \in \ker(\phi)$. This means that $gK = g'K$, which shows injectivity.

Finally, let us show that ψ is surjective. Let $h \in \mathrm{im}(\phi)$. Then there is some $g \in G$ so that $\phi(g) = h$. But then $\psi(gK) = h$ as well. Since $h \in \mathrm{im}(\phi)$ was arbitrary, we have shown that ψ is surjective. Hence ψ is an isomorphism. ■

7.9 Problems

(1) We have stated (as part of the classification theorem of finite abelian groups) that $\mathbb{Z}/105\mathbb{Z}$ can be written as a product of cyclic groups whose orders are powers of primes.

(a) Find three prime numbers p_1, p_2, and p_3 such that

$$\mathbb{Z}/105\mathbb{Z} \cong \mathbb{Z}/p_1\mathbb{Z} \times \mathbb{Z}/p_2\mathbb{Z} \times \mathbb{Z}/p_3\mathbb{Z}\,.$$

(b) Determine the possible values of $(a, b, c) \in \mathbb{Z}/p_1\mathbb{Z} \times \mathbb{Z}/p_2\mathbb{Z} \times \mathbb{Z}/p_3\mathbb{Z}$ so that the homomorphism $f : \mathbb{Z}/105\mathbb{Z} \to \mathbb{Z}/p_1\mathbb{Z} \times \mathbb{Z}/p_2\mathbb{Z} \times \mathbb{Z}/p_3\mathbb{Z}$ determined by $f(1) = (a, b, c)$ is an isomorphism.

(2) (a) Let p be a prime number. Without using the classification of finitely generated abelian groups, show that the groups $\mathbb{Z}/p\mathbb{Z} \times \mathbb{Z}/p\mathbb{Z}$ and $\mathbb{Z}/p^2\mathbb{Z}$ are not isomorphic.
(b) What are all the abelian groups of order pq up to isomorphism, where p and q are distinct primes?

(3) Consider the group $G = \mathbb{Z}/5\mathbb{Z} \times \mathbb{Z}/6\mathbb{Z} \times \mathbb{Z}/30\mathbb{Z}$. Let $H \leq G$ be the cyclic subgroup generated by the element $(1, 1, 2)$. Show that $|H| = 30$. Find a subgroup $G_1 \leq G$ such that $G = H + G_1$ and $H \cap G_1 = \{0\}$. (That is, every element of G can be written as a sum of something in H and something in G_1, and H and G_1 have trivial intersection.)

(4) Let G be a finitely generated abelian group. An *automorphism* of G is an isomorphism $f : G \to G$. Of course the identity map is always one example, but there may be others.

(a) Determine the set of possible automorphisms of $G = \mathbb{Z}/4\mathbb{Z}$, $G = \mathbb{Z}/5\mathbb{Z}$, and $G = \mathbb{Z}/12\mathbb{Z}$.
(b) Let $\mathrm{Aut}(G)$ denote the set of *all* automorphisms of G. Show that $\mathrm{Aut}(G)$ is a group.
(c) For which finitely generated abelian groups G is $\mathrm{Aut}(G)$ abelian?

(5) Let $G = \mathbb{Z}^2$, which we can think of as the set of all integer lattice points in the plane. Give a geometric description of the cosets of the following subgroups H.

(a) H is the subgroup generated by $(1, 0)$.
(b) H is the subgroup generated by $(1, 1)$.
(c) H is the subgroup generated by $(3, 3)$.

(6) Let $G = S_n$ and let $H = \{\rho \in G : \rho(n) = n\}$. In other words, each permutation in H fixes n but permutes $\{1, \ldots, n-1\}$. Let (jn) be the permutation that exchanges j and n, while leaving all other numbers fixed. Show that $H, (1n)H, (2n)H, \ldots, (n-1, n)H$ are all the cosets of H. In other words, show that $\{\mathrm{id}, (1n), (2n), \ldots, (n-1, n)\}$ is a minimal set of representatives for all the cosets of H in G.

(7) For any two elements a and b of any group G, we call the element $aba^{-1} \in G$ the conjugate of b by a. The conjugacy class of b is by definition the subset of G defined by

$$C_b = \{aba^{-1} : a \in G\}.$$

(a) Show that any two conjugacy classes C_b and $C_{b'}$ are either identical or disjoint. (Hint: If they intersect, first show that b' is conjugate to b.)
(b) What is the conjugacy class C_e?
(c) By part (a), we can divide up the elements of G into disjoint subsets that are the various conjugacy classes,

$$G = \bigcup_b C_b.$$

Do this explicitly for $G = \mathbb{Z}/n\mathbb{Z}$, $\mathbb{Z}/n\mathbb{Z} \times \mathbb{Z}/m\mathbb{Z}$, S_3, and D_4.

(8) Recall that a subgroup N in G is called *normal* if the conjugate of every element of N by any element of G is also an element of N. Suppose that $N \leq G$.

(a) Show that if every conjugate of x is in N, and if every conjugate of y is in N, then every conjugate of xy must also be in N.
(b) Show that if every conjugate of x is in N, then every conjugate of x^{-1} is in N as well.
(c) Use the above results to show that N is normal if the conjugate of every element of a generating set for N is also an element of N. (In other words, you don't have to check all elements of N, just a generating set.)
(d) Extend the above further to show that N is normal if the conjugate of every element of a generating set for N, by elements of a generating set for G that is closed under inverses, is also an element of N.

(9) (a) Using the result from the previous problem, show that $N = \{id, \rho^2, \rho^4\}$ is a normal subgroup of D_6 by computing only two conjugates.
(b) Up to isomorphism, what is the quotient D_6/N?

Chapter 8
The Fundamental Group

We have been studying groups in the past three chapters in order to lay the groundwork for introducing the *fundamental group* of a topological space S. This is a homeomorphism invariant that is associated to a topological space. Rather than being a number like the Euler characteristic $\chi(S)$ or a boolean invariant like orientability, the fundamental group associates a *group* to S, denoted $\pi_1(S)$. Furthermore if S is homeomorphic to S', then the fundamental groups $\pi_1(S)$ and $\pi_1(S')$ are isomorphic in the group-theoretic sense. In this chapter, we will build up a set of ideas for defining the fundamental group. For visualization purposes, we will phrase these ideas as if S were a surface; but everything that follows holds mostly unchanged for any topological space.

8.1 Paths and Loops on a Surface

Let S be a surface. Then a *continuous path* on S between two points $p, q \in S$ is just the easily visualized notion of an unbroken 1D curve of points connecting p to q. Formally, we define a path by a continuous mapping $\gamma : [0, 1] \to S$ that satisfies $\gamma(0) = p$ and $\gamma(1) = q$. Technically speaking, γ is a parametrization of the path, and the path itself—viewed as a geometric object—is just the range of γ, i.e. the set of points $\{\gamma(t) : t \in [0, 1]\}$. We'll often be a bit sloppy and just write γ for both the parametrization and the geometric path. Note that different parametrizations can have the same path; for instance $\gamma_1 : [0, 1] \to S$ given by $\gamma_1(t) = \gamma(t^2)$. Note also that it is not necessary to parametrize a path on the interval $[0, 1]$. For instance $\gamma_2 : [0, \frac{1}{2}] \to S$ given by $\gamma_2(t) = \gamma(2t)$ is the same path as γ_1 and γ.

When we talk about paths on a topological space, we will generally want to restrict ourselves to spaces in which there is a path connecting any two points. We call such spaces *path-connected*.

C. Bray et al., *Algebraic Topology*,
https://doi.org/10.1007/978-3-030-70608-1_8

Definition 8.1 A space S is said to be *path-connected* if, for any two points $p, q \in S$, there is a continuous path $\gamma : [0, 1] \to S$ so that $\gamma(0) = p$ and $\gamma(1) = q$.

Example

- A straight line segment in $\mathbb{R}^2$ can be parametrized as follows. Let $p \in \mathbb{R}^2$ be a starting point and $q \in \mathbb{R}^2$ an ending point. Then the vector that points from p to q is simply $v = q - p$. Then points on the line segment are given by $\gamma(t) = p + tv = (1 - t)p + tq$ for $t \in [0, 1]$.
- The unit circle in $\mathbb{R}^2$ centered at $(0, 0)$ can be parametrized by

$$\gamma(t) = (\cos(2\pi t), \sin(2\pi t))$$

for $t \in [0, 1]$.

If $p = q$, so that γ starts and ends at the same place (as in the second example above), we call γ a *loop*. If we need to single out the *basepoint* of the loop, namely the point $p = \gamma(0) = \gamma(1)$, we'll say that γ is *based at* p.

8.2 Equivalence of Paths and Loops

We will define a topological notion of equivalence for paths and loops. Let's stick to paths for now; the extension to loops is straightforward. Suppose $\gamma_0, \gamma_1 : [0, 1] \to S$ are two paths in a topological space S. We'll let γ_0 be equivalent to γ_1, denoted $\gamma_0 \sim \gamma_1$, if it is possible to continuously deform γ_0 into γ_1 while keeping the endpoints fixed. This kind of equivalence is called *homotopy*, and γ_0 is said to be *homotopic* to γ_1. A precise mathematical definition of this notion can be formulated as follows.

Definition 8.2 Two paths γ_0, γ_1 in a topological space S, starting at $p \in S$ and ending at $q \in S$, are said to be *homotopic* if there exists a continuous mapping $F : [0, 1] \times [0, 1] \to S$ such that

- $F(0, t) = \gamma_0(t)$ for all $t \in [0, 1]$,
- $F(1, t) = \gamma_1(t)$ for all $t \in [0, 1]$,
- $F(s, 0) = p$ for all $s \in [0, 1]$,
- $F(s, 1) = q$ for all $s \in [0, 1]$.

We view F as *interpolating* between γ_0 and γ_1 in S. So we should view the functions $t \mapsto F(s, t)$ for each fixed $s \in (0, 1)$ as intermediate paths connecting p to q, and we can denote these by γ_s. The function F is called a *homotopy* between γ_0 and γ_1. See Figure 8.1.

Example Let γ be any path in S, and let γ' be a reparametrization of γ that leaves the endpoints fixed. In other words, $\gamma'(t) = \gamma(g(t))$, where $g : [0, 1] \to [0, 1]$ is a homeomorphism with $g(0) = 0$ and $g(1) = 1$. Then γ and γ' are homotopic via the homotopy

$$F(s, t) = \gamma((1 - s)t + sg(t)) .$$

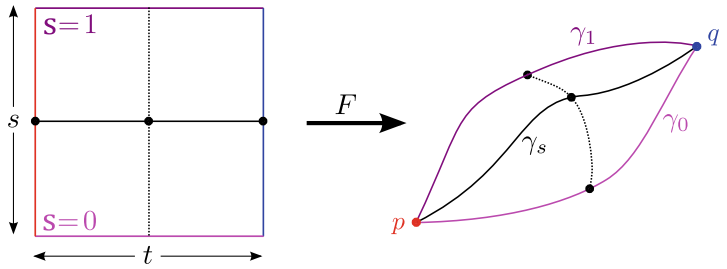


Figure 8.1 Colors match up under F. The solid black line on the left varies t, while s remains constant. This corresponds to an intermediate curve $\gamma_s(t) = F(s, t)$ on the right. The dashed black line on the left varies s, while t remains constant, and the corresponding path is shown on the right.

This example shows that homotopy is a *geometric* concept that does not depend on the way in which paths are parametrized.

Example Let γ_0 and γ_1 be *any* two paths connecting *any* pair of points p, q in $\mathbb{R}^2$. Then we can show that $\gamma_0 \sim \gamma_1$ by constructing the homotopy between γ_0 and γ_1 directly, i.e.

$$F(s, t) = (1 - s)\gamma_0(t) + s\gamma_1(t)\,.$$

This is the so-called "straight-line homotopy" between γ_1 and γ_2, because the interpolation caused by F is such that the point $F(s, t)$ lies on the line segment between $\gamma_1(t)$ and $\gamma_2(t)$.

The conclusion that we can draw from the previous example is that all paths in $\mathbb{R}^2$ with the same endpoints are equivalent to each other. But we shouldn't conclude that homotopy equivalence is a vacuous notion. In fact, a simple modification leads to a space where the straight-line homotopy argument fails to show that all curves are equivalent. This is the topological space $S = \mathbb{R}^2 \setminus \{(0, 0)\}$ in which we have removed the origin from $\mathbb{R}^2$. This time, the straight-line homotopy is not always allowed, because the basic assumption $F : [0, 1] \times [0, 1] \to S$ fails to hold. In particular, this happens when the intermediate path crosses over the origin. In the next few chapters, we'll characterize the homotopic and non-homotopic curves in S in greater detail.

8.3 Equivalence Classes of Paths and Loops

The notion of "equivalence" that we introduced in the previous section does indeed stem from an equivalence relation on the set of paths in S.

Proposition 8.3 *The relation $\sim$ on paths in S from $p \in S$ to $q \in S$ is an equivalence relation.*

Proof The relation $\sim$ is reflexive, because $F(s,t) = \gamma(t)$ for all $s \in [0,1]$ is a homotopy from γ to itself. It is symmetric because, given a homotopy $F(s,t)$ from γ_0 to γ_1, the function $F(1-s,t)$ is a homotopy from γ_1 to γ_0. Finally, if $F(s,t)$ is a homotopy from γ_0 to γ_1 and $G(s,t)$ is a homotopy from γ_1 to γ_2, then the function

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

is a homotopy from γ_0 to γ_2. (Exercise: Why is H continuous?) ■

Consequently, we can think of *the space of all paths in S starting at $p \in S$ and ending at $q \in S$* as being partitioned into a union of—possibly infinitely many, possibly uncountably many—equivalence classes of paths.

Notation We'll denote the equivalence class, or *homotopy class*, of a path γ by $[\gamma]$. Note that $[\gamma] = [\gamma']$ whenever $\gamma' \sim \gamma$, so a homotopy class can have many representatives.

8.4 Multiplication of Path and Loop Classes

A natural geometric operation on two paths γ, γ' is called *concatenation* and is simply adjoining the second path to the first. The concatenated path $\gamma * \gamma'$ is the path obtained by following γ for half of the parameter interval, then following γ' for the rest of the parameter interval. In order for $\gamma * \gamma'$ to be a continuous path, the endpoint of γ must coincide with the starting point of γ'. We make this concept rigorous with the following definition.

Definition 8.4 Let γ be a path from $p \in S$ to $q \in S$ and let γ' be a path from $q \in S$ to $r \in S$. Then $\gamma * \gamma'$ is the path from p to r defined by

$$\gamma * \gamma'(t) = \begin{cases} \gamma(2t) & t \in [0, \frac{1}{2}] \\ \gamma'(2t-1) & t \in [\frac{1}{2}, 1]. \end{cases}$$

Note that if γ, γ' are loops based at the same point, then $p = q = r$ and concatenation works. A crucial fact is that concatenation preserves homotopy classes of paths or loops.

Theorem 8.5 *Let γ_0, γ_0' be two paths with compatible start and end points, and suppose $\gamma_0 \sim \gamma_1$ and $\gamma_0' \sim \gamma_1'$. Then $\gamma_0 * \gamma_0' \sim \gamma_1 * \gamma_1'$.*

Proof Just as we can concatenate paths, we can also concatenate the homotopies between the paths. That is, let F be a homotopy from γ_0 to γ_1 and let F' be a homotopy from γ_0' to γ_1'. Then we can show that

$$H(s,t) = \begin{cases} F(s,2t) & t \in [0, \frac{1}{2}] \\ F'(s,2t-1) & t \in [\frac{1}{2}, 1] \end{cases}$$

is a homotopy between $\gamma_0 * \gamma_0'$ and $\gamma_1 * \gamma_1'$. First, $H(0,t) = F(0,2t) = \gamma(2t)$ for $t \in [0, \frac{1}{2}]$, and $H(0,t) = F'(0,2t-1) = \gamma'(2t-1)$ for $t \in [\frac{1}{2}, 1]$. Thus $H(0,t) = \gamma_0 * \gamma_0'(t)$. Similarly, $H(1,t) = \gamma_1 * \gamma_1'(t)$. Next, $H(s,0) = F(s,0) = p$ for all $s \in [0,1]$, and also $H(s,1) = F'(s,1) = p$ for all $s \in [0,1]$. Finally, H is continuous because at the transition time $t = \frac{1}{2}$ we have $F(s, 2 \cdot \frac{1}{2}) = F(s,1) = p = F'(s,0) = F'(s, 2 \cdot \frac{1}{2} - 1)$ for all s. ■

A consequence of this fortuitous property is that it is now possible to define an operation of *multiplication* on equivalence classes of paths with compatible endpoints. If $[\gamma]$ and $[\gamma']$ are two such equivalence classes, then we define their product by

$$[\gamma] \cdot [\gamma'] = [\gamma * \gamma']\,. \tag{8.1}$$

This is the "natural" definition, of course. But here is something that might have gone wrong: since representatives of equivalence classes are not unique, we can represent $[\gamma]$ as $[\tau]$ and $[\gamma']$ as $[\tau']$ for perhaps different paths τ and τ'. So now we'd hope that our definition gives us $[\gamma] \cdot [\gamma'] = [\tau] \cdot [\tau']$. But it might be the case that $\gamma * \gamma'$ and $\tau * \tau'$ are not homotopic for some reason. Luckily, Theorem 8.5 tells us that indeed $\gamma * \gamma' \sim \tau * \tau'$, so we can be assured that no matter what representatives for $[\gamma]$ and $[\gamma']$ we choose, their concatenations all lie in the homotopy class $[\gamma * \gamma']$. Thus our multiplication (8.1) is well-defined.

An important technical result is to establish the associativity of the multiplication of paths with compatible endpoints.

Theorem 8.6 *Let $\gamma_1, \gamma_2, \gamma_3$ be three paths with compatible endpoints. Then*

$$\big([\gamma_1] \cdot [\gamma_2]\big) \cdot [\gamma_3] = [\gamma_1] \cdot \big([\gamma_2] \cdot [\gamma_3]\big)\,.$$

Proof By re-writing path class multiplication in terms of path concatenation, the desired formula is equivalent to $\big[(\gamma_1 * \gamma_2) * \gamma_3\big] = \big[\gamma_1 * (\gamma_2 * \gamma_3)\big]$ or just $(\gamma_1 * \gamma_2) * \gamma_3 \sim \gamma_1 * (\gamma_2 * \gamma_3)$. So let us try to construct a homotopy between $(\gamma_1 * \gamma_2) * \gamma_3$ and $\gamma_1 * (\gamma_2 * \gamma_3)$. This is not trivial, because after we invoke the definition of $*$, we have

$$\begin{aligned} (\gamma_1 * \gamma_2) * \gamma_3 &= \begin{cases} \gamma_1 * \gamma_2(2t) & t \in [0, \frac{1}{2}] \\ \gamma_3(2t-1) & t \in [\frac{1}{2}, 1] \end{cases} \\ &= \begin{cases} \gamma_1(4t) & t \in [0, \frac{1}{4}] \\ \gamma_2(4t-1) & t \in [\frac{1}{4}, \frac{1}{2}] \\ \gamma_3(2t-1) & t \in [\frac{1}{2}, 1] \end{cases} \end{aligned}$$

and

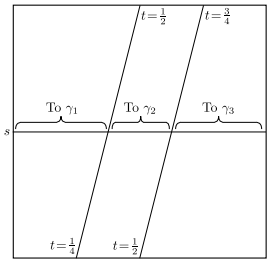


Figure 8.2 The intermediate curves change the length of time spent moving along γ_1, γ_2, and γ_3.

$$\gamma_1 * (\gamma_2 * \gamma_3) = \begin{cases} \gamma_1(2t) & t \in [0, \frac{1}{2}] \\ \gamma_2 * \gamma_3(2t-1) & t \in [\frac{1}{2}, 1] \end{cases}$$
$$= \begin{cases} \gamma_1(2t) & t \in [0, \frac{1}{2}] \\ \gamma_2(4t-2) & t \in [\frac{1}{2}, \frac{3}{4}] \\ \gamma_3(4t-3) & t \in [\frac{3}{4}, 1], \end{cases}$$

which are different. But we can construct a homotopy between them as suggested in Figure 8.2. (Exercises: Does the function shown in Figure 8.2 satisfy the properties of a homotopy? What does an intermediate curve look like? Can you convert the picture into an explicit function?) ■

8.5 Definition of the Fundamental Group

Henceforth we will consider loops based at a point p in a surface S, i.e. paths γ in S such that $\gamma(0) = p = \gamma(1)$. We can concatenate any two such loops because the endpoints are guaranteed to be compatible. We now collect all equivalence classes of all loops based at p into one set.

Definition 8.7 Let S be a topological space with $p \in S$. The *fundamental group of S with basepoint p* is defined as

$$\pi_1(S, p) = \{[\gamma] : \gamma \text{ is a loop based at } p\}.$$

The first fundamental result about the fundamental group is that it is a group!

Theorem 8.8 *Let S be a topological space with $p \in S$. Then $\pi_1(S, p)$ is a group under multiplication of homotopy classes of paths.*

Proof We already know that multiplication of homotopy classes is a well-defined, associative operation. We still have to show the existence of an identity element and the existence of inverses.

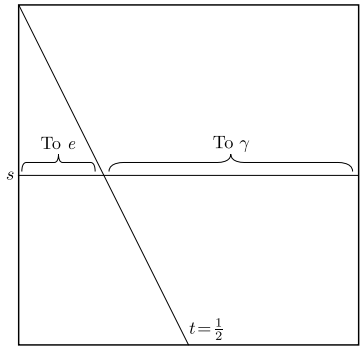


Figure 8.3 The homotopy for $e * \gamma$.

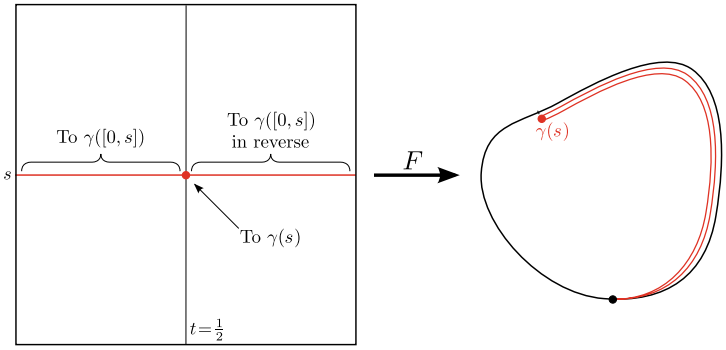


Figure 8.4 The homotopy for $\gamma * \bar{\gamma}$.

For the identity element, we first define a special path $e : [0, 1] \to S$ by $e(t) = p$ for all $t \in [0, 1]$. Next, we claim that $[e]$ is the identity in $\pi_1(S, p)$, or that $[e] \cdot [\gamma] = [\gamma]$ for all $[\gamma] \in \pi_1(S, p)$. In other words, $e * \gamma \sim \gamma$ for all loops γ. To verify this, we compute

$$e * \gamma(t) = \begin{cases} p & t \in [0, \frac{1}{2}] \\ \gamma(2t - 1) & t \in [\frac{1}{2}, 1]\,. \end{cases}$$

Therefore the following homotopy does the trick (see Figure 8.3):

$$F(s, t) = \begin{cases} p & t \in [0, \frac{1-s}{2}] \\ \gamma\left(\frac{2t+s-1}{s+1}\right) & t \in [\frac{1-s}{2}, 1]\,. \end{cases}$$

(Exercise: Double-check that F has all the desired properties!)

We leave inverses for Problem 4; see Figure 8.4. ■

In the definition of $\pi_1(S, p)$ above, we had to choose a basepoint to "anchor" our loops somewhere. This ingredient will play an important role in the future because it

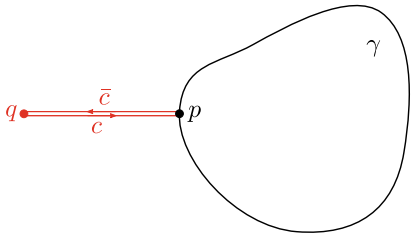


Figure 8.5 Changing basepoints.

will make many proofs simpler. But really, the choice of basepoint above was arbitrary and it would be nice if it didn't actually matter. The next theorem establishes this.

Theorem 8.9 *Let S be a* path-connected *topological space with $p, q \in S$. Then $\pi_1(S, p)$ and $\pi_1(S, q)$ are isomorphic in the sense of groups.*

Proof We will construct an isomorphism between $\pi_1(S, p)$ and $\pi_1(S, q)$. To this end, let $c : [0, 1] \to S$ be a path connecting q with p, i.e. c is continuous and $c(0) = q$ and $c(1) = p$. (This path exists because we have assumed that S is path-connected, meaning any pair of points in S can be connected by a path.) Define $\bar{c} : [0, 1] \to S$ by $\bar{c}(t) = c(1 - t)$, which traces out the path of c in reverse. Now if γ is a loop based at p, then $c * \gamma * \bar{c}$ is a loop based at q. See Figure 8.5. Finally, define $\phi : \pi_1(S, p) \to \pi_1(S, q)$ by $\phi([\gamma]) = [c * \gamma * \bar{c}]$. This ϕ is well-defined because if $\gamma \sim \gamma'$, then we have already shown (thanks to Theorem 8.5) that $c * \gamma * \bar{c} \sim c * \gamma' * \bar{c}$.

To show that ϕ is a homomorphism, we must show that $\phi([\gamma] \cdot [\tau]) = \phi([\gamma]) \cdot \phi([\tau])$. This is in fact an easy task, except for the fact that there's a lot of notation in the way. Thus to proceed, we first "unpack" the notation a bit. By applying the definition of ϕ and of homotopy class multiplication, the left-hand side becomes

$$\phi([\gamma] \cdot [\tau]) = \phi([\gamma * \tau]) = [c * \gamma * \tau * \bar{c}],$$

while the right-hand side becomes

$$\phi([\gamma]) \cdot \phi([\tau]) = [c * \gamma * \bar{c}] \cdot [c * \tau * \bar{c}] = [c * \gamma * \bar{c} * c * \tau * \bar{c}]\,.$$

Therefore, what we really need to show is $[c * \gamma * \bar{c} * c * \tau * \bar{c}] = [c * \gamma * \tau * \bar{c}]$, or more simply that $c * \gamma * \bar{c} * c * \tau * \bar{c} \sim c * \gamma * \tau * \bar{c}$. And now we can see that the homomorphism property holds if we can show that $\bar{c} * c \sim e$. And we've done this before! This is essentially the same as what you will do in Problem 4 when showing that the fundamental group is closed under inverses—i.e. is actually a group.

Finally, we show that ϕ is bijective by constructing an inverse for ϕ. For this purpose, we propose the mapping $\psi : \pi_1(S, q) \to \pi_1(S, p)$ given by $\psi([\gamma]) = [\bar{c} * \gamma * c]$. The mapping ψ is well-defined and is a homomorphism by the same reasoning as for ϕ. Also,

$$\phi \circ \psi([\gamma]) = \phi([\bar{c} * \gamma * c]) = [c * \bar{c} * \gamma * c * \bar{c}] = [c * \bar{c}] \cdot [\gamma] \cdot [c * \bar{c}] = [\gamma],$$

because $[c * \bar{c}] = [e]$. Therefore $\phi \circ \psi = e$, and so ψ is indeed the inverse of ϕ. ■

Remark 8.10 There is a major subtlety here: Although the fundamental groups based at two different points are isomorphic, there is not generally a *preferred choice* of isomorphism. If we had chosen another path c' instead of c in the above proof—where c' is not homotopic to c—we might have ended up with a different isomorphism between the two fundamental groups.

8.6 Problems

(1) Consider the homomorphism $f : F(x, y) \to F^{\text{ab}}(p, q)$ determined by $f(x) = p$ and $f(y) = q$. What are the images of the following words under the homomorphism f?

   (a) $x^2yx^3y^{-2}$
   (b) $xyx^{-3}y^2$
   (c) $x^2y^5x^{-2}y^{-2}xy^{-3}$
   (d) $x^2y^{-4}x^7y^8x^{-8}y^4$

(2) Compute the following products in the free product $F(a, b) * F^{\text{ab}}(x, y)$:

   (a) $\left(axybax^2y^3a\right) \cdot \left(bxybaxy^3a\right)$
   (b) $\left(axybax^2y^3\right) \cdot \left(x^3ybax^2y^3a\right)$
   (c) $(ab) \cdot (bax)$
   (d) $(ab) \cdot (xba)$

(3) Let γ_1 be a path from p_0 to p_1, let γ_2 be a path from p_1 to p_2, and let γ_3 be a path from p_2 to p_3. Prove that $\gamma_1 * (\gamma_2 * \gamma_3) \sim (\gamma_1 * \gamma_2) * \gamma_3$. Construct the homotopy explicitly and draw a representative picture in the (s, t)-square.

(4) Let γ be a loop based at a point x. Define $\bar{\gamma}(t) := \gamma(1 - t)$ for all $t \in [0, 1]$. Show that $\gamma * \bar{\gamma} \sim e$, where e is the loop defined by $e(t) := x$ for all $t \in [0, 1]$. Construct the homotopy explicitly, and draw a representative picture in the (s, t)-square.

(5) Show that $\pi_1(\mathbb{R}, 0) \cong \{e\}$.

(6) The fundamental group π_1 is just one of a family of groups associated to a space. For a space X with basepoint $x \in X$, define $\pi_n(X, x)$ to be the set of homotopy classes of maps from $[0, 1]^n$ to X, so that all points in $[0, 1]^n$ with at least one coordinate equal to 0 or 1 get mapped to x.

   (a) Show that $\pi_n(X, x)$ is a group for $n \geq 1$.
   (b) Show that $\pi_n(X, x)$ is abelian for $n \geq 2$.
   (c) There is also a notion of π_0. What do you expect it to mean? What topological property does it capture? (For this part, think of π_n as homotopy classes of based maps from $\mathbb{S}^n$ to X.)

Chapter 9
Computing the Fundamental Group

9.1 Homotopies of Maps and Spaces

In the last chapter, we discussed homotopies of maps between $[0, 1]$ and a topological space X. We can generalize this to maps between two arbitrary topological spaces X and Y. We say that two maps $f, g : X \to Y$ are *homotopic* if we can continuously deform one into the other. We can express this notion more formally, in a similar manner to how we defined homotopies of maps between $[0, 1]$ and X:

Definition 9.1 Suppose X and Y are two topological spaces, and $f, g : X \to Y$ are two continuous maps. Then a *homotopy* between f and g is a continuous map $H : [0, 1] \times X \to Y$ satisfying the following properties:

- $H(0, x) = f(x)$ for all $x \in X$,
- $H(1, x) = g(x)$ for all $x \in X$.

If there is a homotopy between f and g, then we say that f and g are *homotopic*. We write $f \sim g$ when f and g are homotopic.

Note that this notion of homotopy is a little bit weaker than the one we saw in the last chapter. A homotopy H between two paths f and g starting at p and ending at q must satisfy $H(s, 0) = p$ and $H(s, 1) = q$, i.e. the starting and ending points of all intermediate paths must be the same as those of f and g. In this new version of homotopy, this isn't required. Indeed, there aren't any obvious starting and ending points in sight.

For us, it will be most useful to talk about two maps from one space to itself being homotopic—especially when one of the maps is the identity map. The reason for that is that we're interested in the following notion, that of homotopies of spaces.

Definition 9.2 Suppose that X and Y are two topological spaces. We say that X and Y are *homotopy equivalent* if there are continuous maps $f : X \to Y$ and $g : Y \to X$ so that $g \circ f \sim id_X$ and $f \circ g \sim id_Y$. We call f and g *homotopy equivalences*.

At this stage, it is useful to look at some examples of homotopy equivalent spaces.

C. Bray et al., *Algebraic Topology*,
https://doi.org/10.1007/978-3-030-70608-1_9

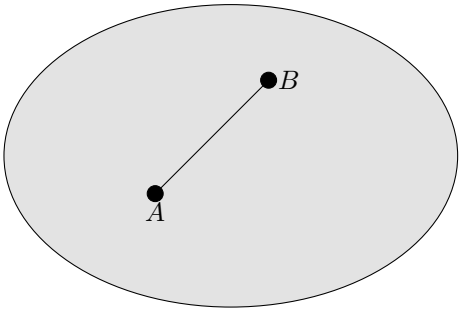


Figure 9.1 A convex set: the segment connecting any two points in the set is entirely contained in the set, as illustrated with the points labeled A and B.

Example Let X be the interval $[0, 1]$, and let Y be the single point 0. Then X and Y are homotopy equivalent. To see this, we need to define maps $f : X \to Y$ and $g : Y \to X$. We define $f(x) = 0$ for all $x \in X$, and $g(0) = 0$ (for the only point 0 in Y). Then $(g \circ f)(x) = 0$ for all $x \in X$. To see that this is homotopic to the identity map $h(x) = x$, we need to construct a homotopy $H : [0, 1] \times X \to X$ between them. Our homotopy will be defined by $H(s, x) = sx$. Then we have $H(0, x) = 0 = (g \circ f)(x)$, and $H(1, x) = x = h(x)$. So this is a homotopy between $(g \circ f)(x)$ and the identity function on X.

Now we have to show that $f \circ g$ is homotopic to the identity function on Y. But this is easier, because both functions are the same function that sends the only point in Y to itself. The homotopy J between them is defined by $J(s, x) = 0$.

This is our first example of homotopy equivalent spaces—and in this case, one of those spaces is a point. We have a word for this phenomenon: *contractible*.

Definition 9.3 We say that a space X is *contractible* if X is homotopy equivalent to a point.

Hence, the above example shows that the interval is contractible. There are many other examples of contractible spaces, and the following describes a general class of them.

Definition 9.4 A subset $X \subset \mathbb{R}^n$ of Euclidean space is called *convex* if, for any two points $x, y \in X$, the segment between x and y is also contained in X.

See Figure 9.1 for a picture of a convex set.

Proposition 9.5 *Any convex set is contractible.*

The proof is very similar to the argument above that shows that the interval is contractible. See if you can work out how to prove this proposition before reading on.

Proof Let X be a convex set, and let x be any point in X. We will find maps $f : X \to \{x\}$ and $g : \{x\} \to X$ so that the compositions are homotopic to the identities on X

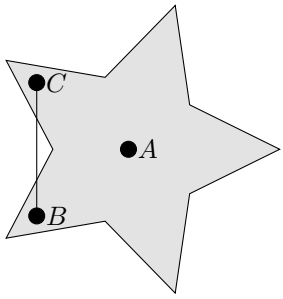


Figure 9.2 A star-shaped, but not convex, set. Every point is visible from point A, but the line connecting points B and C is not contained in the set.

and $\{x\}$. There is really only one way to define the maps: let $f(y) = x$ for all $y \in X$, and let $g(x) = x$. Then the composition $f \circ g : \{x\} \to \{x\}$ is equal to the identity map, whereas $g \circ f : X \to X$ is the map that sends every point in X to x. We now need to construct a homotopy $H : [0, 1] \times X \to X$ so that $H(0, y) = y$ for all $y \in X$, and $H(1, y) = x$ for all $y \in X$. We define H to be the "straight-line homotopy" that we saw in Chapter 8: We set $H(s, y) = (1 - s)y + sx$. The homotopy in the other direction is simply the constant map. ■

In fact, convexity was really a stronger hypothesis than we needed in the above proposition: we didn't need that the segment connecting *any* two points is in X, only that the segment connecting any point to x is in X.

Definition 9.6 We call a subset $X \subset \mathbb{R}^n$ *star-shaped* if there is some point $x \in X$ so that, for any point $y \in X$, the segment connecting y to x is contained in X.

See Figure 9.2 for a picture of a star-shaped set.

Proposition 9.7 *Star-shaped sets are contractible.*

However, there are sets which are not star-shaped that are nonetheless contractible. In fact, sometimes the question of whether a set is contractible can be rather difficult: a famous example is Bing's House with Two Rooms, pictured in Figure 9.3. Let's take a look at why this space is contractible. To do this, instead of homotoping it to a point, we'll start with another contractible space. We'll take a solid cylinder and homotope that to Bing's House with Two Rooms. This is sufficient thanks to Theorem 9.12.

Let's start with a solid cylinder $\overline{B_1(0, 0)} \times [0, 1]$ made of modeling clay, and let's consider the middle layer $\overline{B_1(0, 0)} \times \{\frac{1}{2}\}$. Now poke your finger through the bottom, say at $(\frac{1}{2}, 0, 0)$ until it passes above the middle layer. Then poke another finger through the top, say at $(-\frac{1}{2}, 0, 1)$ until it passes below the middle layer. Now, use the hole coming from the bottom to hollow out the top part, except for the hole your top finger made and a wall connecting it to the edge. Similarly, use the hole coming from the top to hollow out the bottom part, again except the hole made by your bottom finger and a wall connecting it to the edge. What's left is Bing's House with Two Rooms, or perhaps a slightly thickened version of it which is still homotopy equivalent to it. See [Bak10] for lots of pictures.

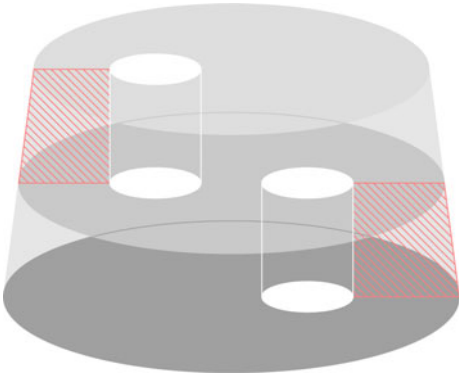


Figure 9.3 Bing's House with Two Rooms: a contractible space that doesn't "look" contractible.

So, now we've seen some examples of contractible sets, but we don't (provably) know any examples of non-contractible sets yet. However, there are many examples.

Example Let X be the set consisting of two points, say $\{0, 1\} \subset \mathbb{R}$. Then X is not contractible. To see this, suppose that it were contractible. That would mean that for a one-point set $\{x\}$, there would be maps $f : X \to \{x\}$ and $g : \{x\} \to X$ so that $g \circ f$ is homotopic to the identity. Let us suppose that $g(x) = 0$, without loss of generality. Then we would need to have some homotopy $H : [0, 1] \times X \to X$ satisfying $H(0, y) = y$ for $y = 0, 1$, and $H(1, y) = 0$ for $y = 0, 1$. Let us look at the function $h(s) = H(s, 1)$. This must be a continuous function, with $h(s) \in X$ for all $s \in [0, 1]$, i.e., $h(s)$ is always either equal to 0 or 1, and $h(0) = 1$ and $h(1) = 0$, so h must contain a "jump" somewhere. However, such a jump function cannot be continuous. To see this, suppose that h were continuous. Then let $I_1 = h^{-1}((-1/2, 1/2))$ and $I_2 = h^{-1}((1/2, 3/2))$ be the preimages of two intervals. (Hence I_1 is the preimage of 0, and I_2 is the preimage of 1.) Then I_1 and I_2 are two disjoint sets. If h is continuous, then both I_1 and I_2 would be open, as they are preimages of open sets. But then we would have expressed $[0, 1]$ as the union of two disjoint nonempty open sets, which is impossible.

But even many sets that are connected are not contractible. It will take us some time to see this, but we can start by relating contractibility to the fundamental group.

Theorem 9.8 *A contractible space has trivial fundamental group.*

We shall give a proof here which is morally correct (see [Che04] for a discussion about what that means) but which nonetheless contains a small gap. In Problem 9, you will turn this argument into a full proof.

Sketch of the Proof Let X be a contractible space. This means we have a homotopy $H : [0, 1] \times X \to X$ so that $H(1, y)$ is some single point $x \in X$. We now let that point x be the basepoint for the fundamental group, and we let γ be any loop in X based at x. We must show that we can deform γ continuously down to the trivial

loop. To do this, we deform γ via H: let $F : [0, 1] \times [0, 1] \to X$ be the map defined by $F(s, t) = H(s, \gamma(t))$. This is a homotopy from γ to the trivial loop, as $F(0, t) = H(0, \gamma(t)) = \gamma(t)$, and $F(1, t) = H(1, \gamma(t)) = x$. ■

Exercise 9.9 Find a minor flaw in this proof. How can you fix it?

As a consequence of Theorem 9.8, any space with non-trivial fundamental group is not contractible. While we haven't yet proven that there are spaces with non-trivial fundamental group, it is certainly plausible that many of our favorite spaces, such as the circle, annulus, and torus, have non-trivial fundamental group. This will turn out to be true.

However, the sphere has trivial fundamental group but is still not contractible. We can't quite prove this yet, but we can get close by reducing the statement to a plausible-looking claim that we will be able to prove once we have studied homology.

Theorem 9.10 *The sphere $\mathbb{S}^2$ is not contractible.*

Most of the Proof Suppose that $\mathbb{S}^2$ were contractible. Then we would have a homotopy H between the identity on $\mathbb{S}^2$ and a constant function, sending every point to the south pole p (say). Hence we have $H(0, x) = x$ for all $x \in \mathbb{S}^2$, and $H(1, x) = p$ for all $x \in \mathbb{S}^2$. We can use this homotopy to construct a rather strange map r from the solid sphere B to the sphere $\mathbb{S}^2$, so that the restriction of this map to the boundary sphere is the identity. We can write down a formula for r:

$$r(x) = \begin{cases} p & \|x\| = 0, \\ H\left(1 - \|x\|, \frac{x}{\|x\|}\right) & \|x\| > 0. \end{cases}$$

(Exercise: Check that r is actually a continuous map from B to $\mathbb{S}^2$, and that $r(x) = x$ for all $x \in \mathbb{S}^2$.) Once we discuss homology, we shall be able to see that such an r cannot exist. ■

Shockingly, however, the *infinite-dimensional* sphere $\mathbb{S}^\infty$ is contractible! An infinite-dimensional sphere has a slightly strange definition: it is the set of points in infinite-dimensional space at distance 1 from the origin, but a point in infinite-dimensional space can only have finitely many nonzero coordinates. Hence we have

$$\mathbb{S}^\infty = \{(x_1, x_2, \ldots) : x_1^2 + x_2^2 + \cdots = 1, \\ \text{and only finitely many } x_i\text{'s are nonzero}\}.$$

It seems rather remarkable that the infinite-dimensional sphere should be contractible when the circle and sphere are not, so let us think about why this is reasonable. Although a circle is not contractible, if we look at one particular circle on a sphere (say, the equator), then we can deform it to a point by deforming it through the north (or south) hemisphere. Let us be more precise about this: we can give a homotopy from the equator $\{(x, y, 0) : x^2 + y^2 = 1\}$ to the north pole $p = (0, 0, 1)$ that stays entirely on the sphere: one such map is

$$H(s,x) = \frac{sp + (1-s)x}{\|sp + (1-s)x\|}.$$

The interpretation of this is that it wants to be the straight line homotopy from the equator to the north pole, but that doesn't lie on the sphere. We fix this by normalizing so that it does stay on the sphere. Since the line doesn't pass through the center of the sphere (which would cause the denominator $\|sp + (1-s)x\|$ to vanish), this can be done in a continuous and unambiguous manner.

Similarly, while the sphere isn't contractible, if we put it inside a 3-dimensional sphere, we can deform it down to a point in much the same way. In general, we can always deform an $(n-1)$-dimensional sphere to a point by putting it into an n-dimensional sphere. So, stated in a very non-rigorous manner, what we do to show that the ∞-dimensional sphere is contractible is to start by finding a homotopy to an "$(\infty - 1)$-dimensional sphere" inside the full ∞-dimensional sphere. Then, we deform that to a point inside the full ∞-dimensional sphere. We now have everything we need to prove the following theorem.

Theorem 9.11 *The infinite-dimensional sphere $\mathbb{S}^\infty$ is contractible.*

Proof We construct a homotopy to a point in two steps: first, we homotope it to an "$(\infty - 1)$-dimensional subsphere," and then we deform that subsphere to a point. Let us define the first part of the map, the homotopy to the subsphere. We will define a map $H(t,x)$ from $\mathbb{S}^\infty$ to the point $(1,0,0,\ldots)$. For a point $x=(x_1,x_2,\ldots) \in \mathbb{S}^\infty$, let $T(x)=(0,x_1,x_2,\ldots) \in \mathbb{S}^\infty$. We will send x to $T(x)$. The homotopy that does this wants to go along a straight line from x to $T(x)$, except that this doesn't lie on the sphere. However, the line connecting x to $T(x)$ doesn't pass through the origin, so we can normalize every point to lie on the sphere. If we want an actual formula, we can write

$$H(s,x) = \frac{2sT(x) + (1-2s)x}{\|2sT(x) + (1-2s)x\|},$$

where $\|x\|$ is the length of x. Hence $H(0,x) = x$ and $H(1/2,x) = T(x)$.

We now work out the other part of the homotopy, which sends the image of T to the point $p = (1,0,0,\ldots)$. We do this in the same sort of way, following a straight-line homotopy from $T(x)$ to p and normalizing so that the path lines on the sphere. We thus take

$$H(s,x) = \frac{(2s-1)p + (2-2s)T(x)}{\|(2s-1)p + (2-2s)T(x)\|}$$

for this part of the homotopy, so that $H(1/2,x) = T(x)$ and $(1,x) = p$. Putting this together, we have the full homotopy:

$$H(s,x) = \begin{cases} \dfrac{2sT(x) + (1-2s)x}{\|2sT(x) + (1-2s)x\|} & 0 \le s \le 1/2, \\ \dfrac{(2s-1)p + (2-2s)T(x)}{\|(2s-1)p + (2-2s)T(x)\|} & 1/2 \le s \le 1. \end{cases}$$

Thus, we have shown that $\mathbb{S}^\infty$ is contractible. ■

Figure 9.4 Dumbbell.

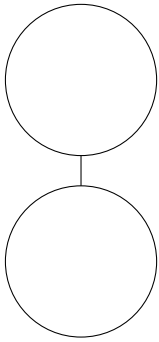
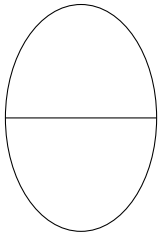


Figure 9.5 Theta.



Example Let us look at an example of two spaces that are homotopy equivalent, but which are not (or at least, to be safe, do not appear to be) contractible. The dumbbell (Figure 9.4) and the theta (Figure 9.5) are homotopy equivalent. It will be painful to try to write down equations for everything, so let us instead describe the maps in words. We can construct a map from the dumbbell to the theta by first squashing the vertical line to obtain a figure-eight. Then we push up the top of the bottom circle into a line. Finally, we push the bottom of the top circle down to the same line.

Similarly, we can describe a map from the theta to the dumbbell. We first collapse the horizontal line to obtain a figure-eight. Then, to create a vertical line, we push in a bit of the bottom of the top circle and the top of the bottom circle, and that gives us the dumbbell.

We should now check that these maps are homotopy equivalences. That is, we should look at the composition of the maps and show that this composition is homotopic to the identity. Let us look at the composition that takes the dumbbell to the theta and then back again. What does this map do? All it does is to move points on and near the vertical line around a little bit: it pushes all the points on the vertical line and in a neighborhood of the vertical line to the center point on this line, and it sends a slightly larger neighborhood to the vertical line and a neighborhood around it. To find a homotopy from the identity on the dumbbell, to this map from the dumbbell to itself, we just imagine slowly pushing the points from the identity to where they are supposed to go. This is a homotopy. The homotopy on the theta is described similarly.

Example After we classified compact surfaces, we wondered about how to classify noncompact surfaces, such as surfaces with punctures. The sphere with three punctures and the torus with one puncture are homotopy equivalent. (Exercise: Can you see why?) There is a deep connection between these two spaces and their topology and geometry; see for instance [HS09].

Homotopy equivalence is an important criterion for classifying topological spaces. It is similar in many ways to classification up to homeomorphism.

Theorem 9.12 *Homotopy equivalence is an equivalence relation.*

Proof As usual, we have to check that homotopy equivalence is reflexive, symmetric, and transitive.

Reflexive: We need to show that a space X is homotopy equivalent to itself; that is, there are maps $f, g : X \to X$ so that $f \circ g$ and $g \circ f$ are homotopic to the identity. The obvious choice here is to let both f and g *be* the identity maps, so that their composition is as well. Clearly, the identity is homotopic to the identity, so homotopy equivalence is reflexive.

Symmetric: We need to show that if X is homotopy equivalent to Y, then Y is also homotopy equivalent to X. But this is essentially built into the definition of homotopy equivalence, because the definition requires that we can find maps $f : X \to Y$ and $g : Y \to X$ so that *both* $f \circ g$ and $g \circ f$ are homotopic to their respective identities.

Transitive: Suppose we have homotopy equivalences $f_1 : X \to Y$, $f_2 : Y \to Z$, $g_1 : Y \to X$, and $g_2 : Z \to Y$. We want to show that $f_2 \circ f_1 : X \to Z$ and $g_1 \circ g_2 : Z \to X$ are homotopy equivalences as well. We need to show that $(g_1 \circ g_2 \circ f_2 \circ f_1)$ and $(f_2 \circ f_1 \circ g_1 \circ g_2)$ are homotopic to their respective identities; we will do only the first of these, as the second is similar. Since f_1 and g_1 are homotopy equivalences, there is some map $H_1 : [0, 1] \times X \to X$ homotoping $g_1 \circ f_1$ to the identity. Similarly, there is a map $H_2 : [0, 1] \times Y \to Y$ homotoping $g_2 \circ f_2$ to the identity. From this, we can define a map $H : [0, 1] \times X \to X$ homotoping $(g_1 \circ g_2 \circ f_2 \circ f_1)$ to the identity. We define H by

$$H(s, x) = \begin{cases} H_1(2s, x) & s \in [0, \frac{1}{2}] \\ g_1\big(H_2(2s - 1, f_1(x))\big) & s \in [\frac{1}{2}, 1]\,. \end{cases}$$

We can check that this is indeed a homotopy, and this shows that homotopy equivalence is an equivalence relation. ■

Remark 9.13 Although at the beginning of the book we stated that a goal of topology is to classify spaces up to homeomorphism, this goal is actually frequently beyond the power of algebraic topology: the tools of algebraic topology are rarely sufficient for distinguishing between two spaces which are homotopy equivalent but not homeomorphic, such as the dumbbell and the theta, or between the sphere with three punctures and the torus with one puncture.

The above theorem and remark suggest a generalization of Theorem 9.8.

Theorem 9.14 *If X and Y are two path-connected spaces that are homotopy equivalent, then they have isomorphic fundamental groups.*

The proof of this theorem is nearly identical to that of Theorem 9.8, so we will not present it here.

9.2 Computing the Fundamental Group of a Circle

So far, it is not yet clear whether the fundamental group is an interesting invariant—that is, does it ever distinguish spaces? Are there any spaces at all with nontrivial fundamental group? In case the name didn't give it away, here's a spoiler: yes! We will show that the circle has nontrivial fundamental group.

Before we do this, let us see intuitively why we ought to believe that the circle has nontrivial fundamental group. Suppose our circle is the set $\mathbb{S}^1 = \{(x, y) : x^2 + y^2 = 1\} \subset \mathbb{R}^2$. Let us pick as our basepoint the point $p = (1, 0)$. Let us consider the loop α on the circle; α is a map $\alpha : [0, 1] \to \mathbb{S}^1$ so that $\alpha(0) = \alpha(1) = p$, and we will choose it to be the loop $\alpha(t) = (\cos 2\pi t, \sin 2\pi t)$, so it is a loop of constant speed that goes around the circle once in the counterclockwise direction.

This loop appears not to be homotopic to the trivial loop: it seems that this loop goes around once, and the trivial loop goes around 0 times. But how can we *prove* that, by doing some clever homotopy, we can't shrink it down to a point?

There are several ways of proving this, and the different techniques highlight different properties of fundamental groups. In this section, we'll see a way to do it using a first example of covering spaces, while in the next chapter we'll see a different proof. We won't talk more about covering spaces in general in this book, but the procedure we employ here to compute fundamental groups is very general and can be used to compute the fundamental group of any reasonably nice space.

The outline of the proof is the following: We want to start with a loop on the circle, lift it up to some other space, and see what the lifted version of the loop looks like.

Theorem 9.15 *The fundamental group of the circle is isomorphic to the group $\mathbb{Z}$ of integers.*

Proof We consider the map $f : \mathbb{R} \to \mathbb{S}^1$, given by $f(t) = (\cos 2\pi t, \sin 2\pi t)$, which wraps the real line around the circle infinitely many times. Let us fix a preimage of $p = (1, 0)$ in $\mathbb{R}$, say $0 \in \mathbb{R}$. Now, let $\gamma : [0, 1] \to \mathbb{S}^1$ be a loop in $\mathbb{S}^1$ based at p. We can lift γ to a path $\widetilde{\gamma}$ in $\mathbb{R}$ so that $\widetilde{\gamma}(0) = 0$; this means that $f(\widetilde{\gamma}(t)) = \gamma(t)$; in fact, there is exactly one such path $\widetilde{\gamma}$. To lift a path from $\mathbb{S}^1$ to $\mathbb{R}$, we simply lift it a bit at a time: each point has infinitely many preimages, but only (at most) one of them is very close to any given point, so we can easily figure out which one to use. This can be proven rigorously, but we will not do so here. (For a proof, see [Mas91, Chapter V, Lemma 3.1].)

Now, while γ was a *loop* in $\mathbb{S}^1$, $\widetilde{\gamma}$ will not necessarily be a loop in $\mathbb{R}$; that is, it might have different starting and ending points. Suppose that $r \in \mathbb{R}$ is the ending point for $\widetilde{\gamma}$; that is, $r = \widetilde{\gamma}(1)$. Since $f(\widetilde{\gamma}(t)) = \gamma(t)$, we must have $f(r) = p$, so r must be *some* preimage of p. The preimages of p under f are exactly the integers. Hence, r is some integer.

This now tells us how we should define the map $\phi : \pi_1(\mathbb{S}^1, p) \to \mathbb{Z}$. That is, for any equivalence class of loops $[\gamma] \in \pi_1(\mathbb{S}^1, p)$, we define $\phi([\gamma]) = \widetilde{\gamma}(1) = r$.

There are many things we need to check about this map. First, we need to check that it is a well-defined map. What could go wrong? Remember that $[\gamma]$ is an *equivalence class* of loops in $\mathbb{S}^1$ based at p, whereas we have defined $\phi([\gamma])$ in terms of the lift of a *representative* of the equivalence class, namely the loop γ. But, of course, the class $[\gamma]$ can have many representatives; i.e. we have $[\gamma] = [\gamma']$ whenever γ' is homotopic to γ. So, we have to make sure that if we define $\phi([\gamma])$ by lifting γ or by lifting γ', we get the same answer. In order to show this, we observe that we can lift an entire homotopy from one loop to another, to a homotopy of the lifted versions in $\mathbb{R}$, so that the endpoints stay fixed throughout the homotopy. This is in much the same way that we can lift a path; see [Mas91, Chapter V, Lemma 3.3]. The conclusion we draw is that, because the homotopy doesn't move the endpoints, two homotopic loops lift to paths in $\mathbb{R}$ with the same endpoints, which is just what we need for $\phi([\gamma])$ to be well-defined.

So, now we've seen that ϕ is actually a well-defined map from $\pi_1(\mathbb{S}^1, p)$ to $\mathbb{Z}$. We must now check that it is a homomorphism. Let us take two equivalence classes of loops, $[\gamma_1]$ and $[\gamma_2]$, in $\mathbb{S}^1$ based at p. Let us suppose that $\phi([\gamma_1]) = a$ and $\phi([\gamma_2]) = b$. What is $\phi([\gamma_1][\gamma_2])$? In order to work this out, because $[\gamma_1][\gamma_2] = [\gamma_1 * \gamma_2]$, we want to figure out how to lift $\gamma_1 * \gamma_2$, given that we know how to lift γ_1 and γ_2. The lift of $\gamma_1 * \gamma_2$ will look like this:

$$\widetilde{\gamma_1 * \gamma_2}(t) = \begin{cases} \widetilde{\gamma}_1(2t) & 0 \le t \le 1/2, \\ \widetilde{\gamma}_2(2t-1) + a & 1/2 \le t \le 1. \end{cases}$$

(Exercise: Why is this the right definition?)

Now we can check that $\phi([\gamma_1][\gamma_2]) = \phi([\gamma_1]) + \phi([\gamma_2])$. Clearly $\phi([\gamma_1]) + \phi([\gamma_2]) = a + b$, and

$$\phi([\gamma_1][\gamma_2]) = \phi([\gamma_1 * \gamma_2]) = \widetilde{\gamma_1 * \gamma_2}(1) = \gamma_2(1) + a = a + b.$$

Hence ϕ is a homomorphism.

We wanted to check that ϕ is an *iso*morphism, so we must check that it is surjective and injective. Let us check that it is injective. Suppose $\phi([\gamma]) = 0$, so that $\widetilde{\gamma}(1) = 0$. In that case, $\widetilde{\gamma}$ is a *loop*, and not just a path, because $\widetilde{\gamma}(0) = \widetilde{\gamma}(1)$. But we know that $\mathbb{R}$ is contractible, so this means that there is a homotopy $H : [0, 1] \times [0, 1] \to \mathbb{R}$ which deforms $\widetilde{\gamma}$ to the constant loop. But then $f(H(s, t))$ is a homotopy between $\gamma(t)$ and the constant loop in $\mathbb{S}^1$. Thus γ is homotopic to the constant loop, so its equivalence class is trivial in $\pi_1(\mathbb{S}^1, p)$.

Finally, we need to show that ϕ is surjective, i.e. given any integer n, we can find a loop γ in $\mathbb{S}^1$ so that $\phi([\gamma_n]) = n$. To do this, we just write down such a loop γ_n, and it is a loop that goes around the circle n times in the counterclockwise direction (or $-n$ times in the clockwise direction when $n < 0$). We define γ_n to be the loop given by

$$\gamma_n(t) = (\cos 2\pi nt, \sin 2\pi nt).$$

Thus we have checked everything we needed, and we have shown that $\pi_1(\mathbb{S}^1, p) \cong \mathbb{Z}$. ■

9.3 Problems

(1) Show that a space X is contractible if and only if, for each point $x_0 \in X$, there is a homotopy $H : [0, 1] \times X \to X$ so that $H(0, x) = x$ and $H(1, x) = x_0$ for all $x \in X$.
(2) Show that homeomorphisms are homotopy equivalences.
(3) Show that a punctured torus is homotopy equivalent to the theta or the dumbbell.
(4) Let g_1 and g_2 be nonnegative integers, and let n_1 and n_2 be (strictly) positive integers. Show that an orientable compact genus g_1 surface with n_1 punctures is homotopy equivalent to an orientable compact genus g_2 surface with n_2 punctures if and only if $2g_1 + n_1 = 2g_2 + n_2$.
(5) Let $f : \mathbb{S}^1 \to \mathbb{S}^1$ be a continuous map that is not homotopic to the identity. Show that there is a point $x \in \mathbb{S}^1$ so that $f(x) = -x$; that is, $f(x)$ is diametrically opposite of x.
(6) Let γ_0 and γ_1 be two paths in $\mathbb{S}^2$ starting at $p \in \mathbb{S}^2$ and ending at $q \in \mathbb{S}^2$. Here, we can view $\mathbb{S}^2$ as the unit sphere in $\mathbb{R}^3$. Explain why there is a homotopy from γ_0 to γ_1. Note that the intermediate curves γ_s must all lie on $\mathbb{S}^2$ and connect p to q.
(7) Let S be the annular region in the plane lying between the circle of radius 1/2 and the circle of radius 2. Let $p = (1, 0)$ and $q = (-1, 0)$. Let γ_0^+ be the half-circular arc from p to q in the counterclockwise sense, and let γ_0^- be the half-circular arc from p to q in the clockwise sense. Let γ_1 be some other path in S from p to q. Explain some of the issues involved in constructing a homotopy from γ_1 to either of $\gamma_0^{\pm}$. Can you always do it? When can you not do it? If you think you can do it, how would you write down a formula for the homotopy? Note that γ_1 can be an arbitrarily "bad" path from p to q! Is γ_0^+ homotopic to γ_0^-?
(8) Mimic the computation of $\pi_1(\mathbb{S}^1)$ to compute $\pi_1(\mathbb{RP}^2)$. (Hint: Think of a natural space to lift to.)

(9) We showed in the text that a contractible space S has trivial fundamental group. We did this by showing that if γ is a loop in S and $F : [0, 1] \times S \to S$ is the contraction homotopy, then $f : [0, 1] \times [0, 1] \to S$ given by $f(s, t) = F(s, \gamma(t))$ is the homotopy of the loop to a point. However, our argument isn't quite rigorous because $f(s, 0)$ is not necessarily fixed as s varies, which the definition of loop homotopy requires. Make the necessary modifications to this argument so that it becomes completely rigorous.

Chapter 10
Tools for Fundamental Groups

10.1 More Fundamental Groups

We have worked quite hard to find a space whose fundamental group is non-trivial. We should capitalize on this result and see if we can find other, related spaces whose fundamental groups can now be computed easily as a result of our hard work. An example where this approach is successful is for *product spaces*.

We must first make a small digression and attempt to put the notion of the *product* of two topological spaces X and Y on a slightly more rigorous footing. Just as we defined the product of two groups, let us define the product space as

$$X \times Y := \{(x, y) : x \in X \text{ and } y \in Y\}.$$

So far, this just defines $X \times Y$ as a *set* of points. To really turn $X \times Y$ into a topological space, we have to extend the topological notions from X and Y to $X \times Y$. We gave the precise mathematical definition of a *topological space* earlier, in Chapter 3, but let us repeat it once more.

A topological space X is a set of points together with a *topology*, which we'll loosely take to mean "a way of defining an open set." If $X \subseteq \mathbb{R}^3$ then we said that a subset $U \subseteq X$ is open if and only if, for every $x \in U$, we can find $\varepsilon > 0$ so that the open ball $B_\varepsilon(x) \subseteq \mathbb{R}^3$ satisfies $B_\varepsilon(x) \cap X \subseteq U$. Thus we use the relatively open balls $B_\varepsilon(x) \cap X$ for all $x \in X$ and $\varepsilon > 0$ to prove the openness of any subset of X. We say that the relatively open balls of X constitute a "basis" for X.

More generally, we may have some space X that is *not* a subset of $\mathbb{R}^3$—or any $\mathbb{R}^n$ for that matter—yet we still wish to consider it to be a topological space. What this means is that we need a way of deciding whether a subset of X is open or not. We allow ourselves flexibility in how this is done, but we require that certain natural properties of open sets that we have seen before still hold.

Definition 10.1 A *topological space* is a set X together with a collection $\mathcal{T}$ of subsets of X, so that

C. Bray et al., *Algebraic Topology*,
https://doi.org/10.1007/978-3-030-70608-1_10

- $\varnothing, X \in \mathscr{T}$,
- If $A_1, A_2, \ldots$ are in $\mathscr{T}$, then $\bigcup A_i \in \mathscr{T}$,
- If $A_1, A_2 \in \mathscr{T}$, then $A_1 \cap A_2 \in \mathscr{T}$.

We call $\mathscr{T}$ a *topology*, and we call the sets in $\mathscr{T}$ *open sets*.

The upshot of the above discussion is that we have to describe the open sets of $X \times Y$ in order to transform $X \times Y$ into an honest-to-goodness topological space. It is done as follows.

Definition 10.2 We call a subset $U \subseteq X \times Y$ *open* if, for every point $(x, y) \in U$, there are open sets $U_1 \subset X$ and $U_2 \subset Y$ with $x \in U_1$ and $y \in U_2$, so that the product space $U_1 \times U_2 \subseteq U$.

Remark 10.3 There is also a version of this *product topology* on a direct product of infinitely many topological spaces, but it isn't what you would first guess it to be. See [DS84] for the general definition of the product topology, with an explanation of why it is the correct definition.

Example Our simplest examples of product spaces are the higher-dimensional Euclidean spaces. That is, $\mathbb{R}^2 := \mathbb{R} \times \mathbb{R}$ and $\mathbb{R}^n$ is defined recursively as $\mathbb{R}^n := \mathbb{R}^{n-1} \times \mathbb{R}$.

Example The *two-dimensional torus* is defined as $\mathbb{T}^2 := \mathbb{S}^1 \times \mathbb{S}^1$. The n-dimensional torus is defined recursively as $\mathbb{T}^n := \mathbb{T}^{n-1} \times \mathbb{S}^1$.

Remark 10.4 Observe that the torus $\mathbb{S}^1 \times \mathbb{S}^1$ is *not* the same as the sphere $\mathbb{S}^2$.

We now return to fundamental groups. We have just developed a simple method of constructing a bigger topological space $X \times Y$ out of two smaller ones X and Y. The question is: What happens to the fundamental groups in this process? The following theorem gives the answer. For clarity, we will use (here only!) the notation $\times_{TS}$ for the product of topological spaces just described, and $\times_G$ the direct product of groups. Recall also that $\cong$ means "isomorphic as groups."

Theorem 10.5 *Let X, Y be two path-connected topological spaces. Then $\pi_1(X \times_{TS} Y) \cong \pi_1(X) \times_G \pi_1(Y)$.*

Proof Despite the lengthy lead-up to this theorem, the proof is very straightforward. A loop in $X \times Y$ can be uniquely written $\gamma(t) := (\gamma_1(t), \gamma_2(t))$, where γ_1 and γ_2 are loops in X and Y, respectively. Furthermore, a homotopy of loops $\tilde{\gamma} := (\tilde{\gamma}_1, \tilde{\gamma}_2)$ to γ can be written $F(s, t) := (F_1(s, t), F_2(s, t))$, where $F_1(s, t)$ and $F_2(s, t)$ are homotopies of $\tilde{\gamma}_1 \sim \gamma_1$ and $\tilde{\gamma}_2 \sim \gamma_2$, respectively. We conclude from this discussion that two loops are homotopic in $X \times Y$ if and only if both of the "component loops" are homotopic in X and Y, respectively.

To construct an isomorphism $\phi : \pi_1(X \times_{TS} Y) \to \pi_1(X) \times_G \pi_1(Y)$, we simply define

$$\phi([\gamma]) := \big([\gamma_1], [\gamma_2]\big),$$

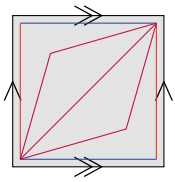


Figure 10.1 $\pi_1(\mathbb{T}^2)$ is abelian.

where $\gamma(t) := (\gamma_1(t), \gamma_2(t))$. The discussion above ensures that ϕ is well-defined (i.e. if $[\gamma] = [\tilde{\gamma}]$ or equivalently $\gamma \sim \tilde{\gamma}$, then $\phi([\gamma]) = \phi([\tilde{\gamma}])$). It remains to check that ϕ is a homomorphism and is both injective and surjective. This is an exercise for you! ■

The consequence of the theorem above, and the hard work we have done finding $\pi_1(\mathbb{S}^1)$, is that we now know other spaces with non-trivial fundamental groups: the product space which is the n-dimensional torus $\mathbb{T}^n$.

Corollary 10.6 *For every $n \in \mathbb{N}$ we have $\pi_1(\mathbb{T}^n) \cong \mathbb{Z} \overbrace{\times \cdots \times}^{n \text{ times}} \mathbb{Z}$.*

Exercise 10.7 Observe that the corollary above tells us that $\pi_1(\mathbb{T}^2)$ is an abelian group. Therefore if a, b are the generators of $\pi_1(\mathbb{T}^2)$ as a group, then $aba^{-1}b^{-1} = \mathrm{id}$. If $a = [\gamma_1]$ and $b = [\gamma_2]$ for loops $\gamma_1, \gamma_2 \subseteq \mathbb{T}^2$, then we must have $\gamma_1 * \gamma_2 * \bar{\gamma}_1 * \bar{\gamma}_2 \sim e$, where the bar denotes the reversed loop and e is the constant loop, as always. It seems to be slightly non-obvious that the two generating loops of $\pi_1(\mathbb{T}^2)$ should behave in this way, if we try to picture the loops directly on the torus. However, it's much easier to see what's going on using an ID space, as shown in Figure 10.1.

10.2 The Degree of a Loop

The purpose of this section is to introduce a tool, called the *degree*, for studying the topological properties of curves.

Let $p \in \mathbb{R}^2$ be a point, and let $\gamma : [0, 1] \to \mathbb{R}^2$ be a continuous loop that does not pass through p. We can thus view γ as a map from the circle $\gamma : \mathbb{S}^1 \to \mathbb{R}^2 \setminus \{p\}$. We would like to define the degree of γ relative to p—roughly speaking—as *the number of times γ winds around p*. To put this notion on a rigorous footing, we proceed as follows.

Definition 10.8 Let $p \in \mathbb{R}^2$, and let γ be a loop in $\mathbb{R}^2$ not passing through p. Then the *degree* of γ relative to p, denoted $\deg_p(\gamma)$, is defined via the following procedure.

1. Partition $[0, 1]$ into n disjoint sub-intervals of the form $[t_i, t_{i+1}]$, where $0 = t_0 < t_1 < \cdots < t_{n-1} < t_n = 1$. Choose these t_i so close to each other that the angle, measured counter-clockwise, between the vectors $\gamma(s) - p$ and $\gamma(s') - p$ for any $s, s' \in [t_i, t_{i+1}]$ belongs to $(-\pi/2, \pi/2)$.

2. Let θ_i be the angle, measured counter-clockwise, between $\gamma(t_{i+1}) - p$ and $\gamma(t_i) - p$.
3. Define $\deg_p(\gamma) := \frac{1}{2\pi} \sum_{i=1}^{n-1} \theta_i$.

Remark 10.9 The purpose of Steps 1–2 is to remove the ambiguity involved in deciding what the angle is between two vectors—really this is a number that is defined only up to an arbitrary integer multiple of 2π. But we can be unambiguous if the vectors point in directions that are sufficiently close to each other. This should remind you of what we did while lifting paths from $\mathbb{S}^1$ to $\mathbb{R}$ when computing $\pi_1(\mathbb{S}^1)$: each point has many preimages, but if we lift just a little bit at a time, we know how to choose the right one.

Remark 10.10 It is not entirely trivial to show that a partition of the kind described in Step 1 always exists. Suffice it to say that the existence of the required partition follows from the continuity of γ and the compactness of $\mathbb{S}^1$.

We now prove three basic results about the degree.

Lemma 10.11 *The degree is an integer.*

Proof We can measure each angle θ_i, at least up to an ambiguity equal to an integer multiple of 2π, as follows. Let α_{1i} be the angle between $\gamma(t_{i+1}) - p$ and the x-axis. Let α_{2i} be the angle between $\gamma(t_i) - p$ and the x-axis. Then $\theta_i \equiv \alpha_{1i} - \alpha_{2i} \pmod{2\pi}$. Hence $2\pi \cdot \deg_p \equiv \sum_i (\alpha_{1i} - \alpha_{2i}) \pmod{2\pi}$. But this sum is telescoping and itself equals an integer multiple of 2π, because $\gamma(0) = \gamma(1)$. ■

Lemma 10.12 *The degree is independent of the choice of partition of* $[0, 1]$.

Proof Two different partitions of $[0, 1]$ satisfying the requirements of Step 1 always have a *common refinement* that also satisfies these requirements. Moreover, each of the input partitions can be transformed into the common refined partition by adding several additional points one at a time (and the requirements of Step 1 are met at each step). We can thus prove the partition-independence of the degree, if the degree is unchanged whenever one additional point is added to a partition that satisfies the requirements of Step 1.

To this end, suppose the additional point t' is added in the sub-interval $[t_i, t_{i+1}]$. Let θ' be the angle between $\gamma(t') - p$ and $\gamma(t_i) - p$. Let θ'' be the angle between $\gamma(t_{i+1}) - p$ and $\gamma(t') - p$. Using an argument similar to the one from the previous lemma, we can say that $\theta_i \equiv \theta' + \theta'' \pmod{2\pi}$. But since $\theta', \theta'' \in (-\pi/2, \pi/2)$, it must be the case that $\theta_i = \theta' + \theta''$. Consequently the formula for degree is unchanged. ■

Lemma 10.13 *The degree* $\deg_p$ *is a homotopy invariant of curves* $\gamma : \mathbb{S}^1 \to \mathbb{R}^2 \setminus \{p\}$.

Proof Suppose $\gamma \sim \tilde{\gamma}$ are homotopic curves in $\mathbb{R}^2 \setminus \{p\}$ and let $H : [0, 1] \times [0, 1] \to \mathbb{R}^2 \setminus \{p\}$ be a homotopy between them. Introduce partitions $0 = s_0 < s_1 < \cdots <$

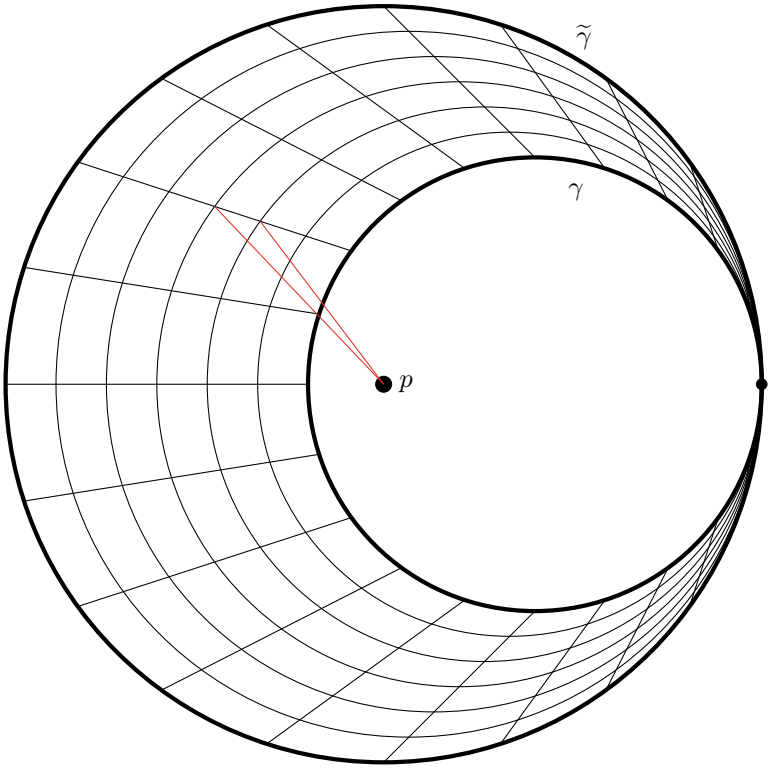


Figure 10.2 One of the θ'_{ij}'s, marked in red.

$s_m = 1$ and $0 = t_0 < t_1 < \cdots < t_n = 1$ for each of the $[0, 1]$ factors. By continuity and compactness, we can choose these s and t values so close together that the angle, measured counter-clockwise, between $H(s, t) - p$ and $H(s', t') - p$ belongs to $(-\pi/4, \pi/4)$ for all $s, s' \in [s_i, s_{i+1}]$ and $t, t' \in [t_j, t_{j+1}]$. Now define the following angles, all of which are measured counter-clockwise and relative to p:

$$\begin{aligned}
\theta_{ij} &:= \text{angle between } H(s_i, t_j) \text{ and } H(s_i, t_{j+1}) \\
\theta'_{ij} &:= \text{angle between } H(s_i, t_j) \text{ and } H(s_{i+1}, t_j) \\
\theta''_{ij} &:= \text{angle between } H(s_i, t_{j+1}) \text{ and } H(s_{i+1}, t_{j+1}) \\
\theta'''_{ij} &:= \text{angle between } H(s_{i+1}, t_j) \text{ and } H(s_{i+1}, t_{j+1}) \,.
\end{aligned}$$

One of the θ_{ij}'s is pictured in Figure 10.2.

By using arguments similar to those of Lemma 10.11, we can show that $\theta_{ij} \equiv \theta'_{ij} + \theta'''_{ij} - \theta''_{ij} \pmod{2\pi}$. But since all these angles belong to $(-\pi/4, \pi/4)$, this equality holds true without the $\pmod{2\pi}$. To conclude, we simply compute:

$$\begin{aligned} 2\pi \cdot \deg_p(\gamma) &= \sum_{j=0}^{n-1} \theta_{0j} \\ &= \sum_{j=0}^{n-1} \left(\theta'_{0j} + \theta'''_{0j} - \theta''_{0j}\right) \\ &= \sum_{j=0}^{n-1} \theta'''_{0j} \\ &= \sum_{j=0}^{n-1} \theta_{1j} \,. \end{aligned}$$

The third equation follows from the second because $\theta''_{0j} = \theta'_{0j+1}$, and thus the pair of sums telescopes to zero; we also use the fact that the angles repeat as we go around the circle, too. The fourth equation then follows because $\theta'''_{0j} = \theta_{1j}$ by definition. If we repeat the above process another $n-1$ times, we get $2\pi \cdot \deg_p(\gamma) = \sum_{j=0}^{n-1} \theta_{nj} = 2\pi \cdot \deg_p(\tilde{\gamma})$ as required. ■

Example Let $\gamma_n(t) := (\cos(2\pi nt), \sin(2\pi nt))$. We have used this curve before in Chapter 9; it is the curve that winds n times around the unit circle. We can thus view γ_n as a map $\gamma_n : \mathbb{S}^1 \to \mathbb{R}^2 \setminus \{(0,0)\}$. Because γ_n winds monotonically around the circle, it is easy to see that $\deg_{(0,0)}(\gamma_n) = n$.

10.3 Fundamental Group of a Circle—Redux

As a first application of the degree of a loop, we will prove again that $\pi_1(\mathbb{S}^1) \cong \mathbb{Z}$. Our proof here will be slightly more "hands on" than our previous proof. But you will definitely notice the similarities, because the ideas behind both proofs are essentially the same.

To begin, let α be the path that goes around the circle once: $\alpha(t) := (\cos 2\pi t, \sin 2\pi t)$. We divide the proof that $\pi_1(\mathbb{S}^1) \cong \mathbb{Z}$ into two propositions. First we will prove that the group is generated by $[\alpha]$, and then we will prove that $[\alpha]$ has infinite order. It is in this second step where the degree comes in.

Proposition 10.14 *The fundamental group $\pi_1(\mathbb{S}^1)$ is a cyclic group generated by the path α.*

Proof Construct two overlapping open subsets $\mathcal{U}_1$ and $\mathcal{U}_2$ of the circle by slightly extending the top half of the circle and the bottom half of the circle, respectively. These sets are both contractible; their union is the whole circle; and their intersection consists of two disjoint open subsets containing $(1,0)$ and $(-1,0)$, respectively.

Let γ be any loop on the circle based at the point $(1, 0)$, and let $\beta = [\gamma]$ be its homotopy class. If γ stays inside $\mathcal{U}_1$ or $\mathcal{U}_2$, then because these sets are contractible, γ is homotopic to the constant map, and $\beta = [\alpha^0] = [e]$.

If γ does not stay in $\mathcal{U}_1$ or $\mathcal{U}_2$, we can find a partition of the unit interval of the form $0 = t_0 < t_1 < t_2 < \cdots < t_{n-1} < t_n = 1$ so that the following conditions hold.

- For all $t_i < t < t_{i+1}$, γ is always in $\mathcal{U}_1$ or always in $\mathcal{U}_2$.
- If γ stays inside $\mathcal{U}_1$ for times $t_i < t < t_{i+1}$, then γ stays in $\mathcal{U}_2$ for all times $t_{i+1} < t < t_{i+2}$ (and vice-versa).

We denote by β_i the part of the curve γ restricted to $[t_i, t_{i+1}]$, so that

$$\gamma = \beta_0 * \beta_1 * \cdots * \beta_{n-1}.$$

We can arrange to have all of the β_i end at either $(1, 0)$ or $(-1, 0)$. Indeed, let σ_i be the shortest path joining $\gamma(t_i)$ with either $(1, 0)$ or $(-1, 0)$, then

$$\beta = (\beta_0 * \gamma_1) * (\bar{\gamma}_1 * \beta_1 * \gamma_2) * (\bar{\gamma}_2 * \beta_2 * \gamma_3) * \cdots * (\bar{\gamma}_{n-1} * \beta_{n-1})$$

and we relabel $\beta_i' := (\bar{\gamma}_i * \beta_i * \gamma_{i+1})$ so that $\gamma = \beta_0' * \beta_1' * \cdots * \beta_{n-1}'$.

If any of the β_i' is a closed curve—i.e. a loop based at either $(1, 0)$ or $(-1, 0)$—then the contractibility of $\mathcal{U}_1$ and $\mathcal{U}_2$ implies that β_i' is trivial, so we can just drop it from our considerations. Thus we can assume that each β_i' either joins $(1, 0)$ with $(-1, 0)$ or joins $(-1, 0)$ with $(1, 0)$. Now, again because $\mathcal{U}_1$ and $\mathcal{U}_2$ are contractible spaces, we can deform each remaining β_i' into one of four possible curves: the upper half of α in the forward or backward direction, or the lower half of α in the forward or backward direction. Call these curves $\eta_1, \bar{\eta}_1, \eta_2$ and $\bar{\eta}_2$ respectively. Notice that $\eta_1 * \eta_2 = \alpha$. Therefore, we conclude:

i. γ is the constant loop, or
ii. $\gamma = \eta_1 * \eta_2 * \eta_1 * \eta_2 * \cdots * \eta_1 * \eta_2$, or
iii. $\gamma = \bar{\eta}_2 * \bar{\eta}_1 * \bar{\eta}_2 * \bar{\eta}_1 * \cdots * \bar{\eta}_2 * \bar{\eta}_1$.

In each case we have, $\beta = [\alpha^0]$, or $\beta = [\alpha^m]$ for some $m > 0$, or $\beta = [\alpha^m]$ for some $m < 0$, respectively. ■

Proposition 10.15 *The generator* $[\alpha]$ *of* $\pi_1(\mathbb{S}^1)$ *has infinite order.*

Proof Suppose $[\alpha]$ had finite order m, namely $[\alpha]^m = [e]$. Then the m-fold concatenation $\alpha * \cdots * \alpha$ is homotopic to the constant curve. But, the constant curve has degree zero while the curve $\alpha * \cdots * \alpha$ has degree m. This is a contradiction, so $[\alpha]$ must have infinite order. ■

The consequence of these two propositions is that $\pi_1(\mathbb{S}^1) = \langle [\alpha] \rangle \cong \mathbb{Z}$.

10.4 The Induced Homomorphism on Fundamental Groups

As we have seen on several occasions already, an important theme in modern mathematics is that objects are best viewed not in isolation but in terms of how they relate to similar objects. We have been following this philosophy when we looked not just at topological spaces, but also at continuous maps between them. Similarly, we looked not just at groups, but also at homomorphisms between them. In this section, we will connect the two, by studying the following question: Suppose X and Y are two topological spaces, and $f : X \to Y$ is a continuous map. How do $\pi_1(X)$ and $\pi_1(Y)$ relate to each other?

The answer is that there is a natural homomorphism between $\pi_1(X)$ and $\pi_1(Y)$ induced by f. Actually, it is obvious that there is already a natural homomorphism between $\pi_1(X)$ and $\pi_1(Y)$, because we could be talking about the trivial homomorphism that takes every element of $\pi_1(X)$ to the identity in $\pi_1(Y)$. But we mean something much more interesting here!

First, let us introduce some terminology. When we say that (X, x) is a *based* topological space, we simply mean that X is a topological space, and one of its points $x \in X$ has been singled out for attention. A continuous map $f : (X, x) \to (Y, y)$ between two based topological spaces is simply a continuous map $f : X \to Y$ that satisfies the additional property $f(x) = y$.

Proposition 10.16 *Let $f : (X, x) \to (Y, y)$ be a continuous map between two based topological spaces. Then there is a homomorphism*[1] *$f_* : \pi_1(X) \to \pi_1(Y)$, defined as follows: If γ is a loop in X based at x, then $f_*([\gamma]) := [f \circ \gamma]$. We call f_* the* induced homomorphism *of f.*

Let us explain what f_* does in slightly more verbose language. If γ is a loop in X based at x, then we can use f to take γ to the image of γ under f, namely $f \circ \gamma$, which is a loop in Y based at y. Since we can do this for any loop, we can define a similar operation on *classes* of loops. So f_* takes a class of loops in X based at x to a class of loops in Y based at y by forming $f \circ \gamma$ to every representative γ of the class. In other words f_* takes the homotopy class of γ to the homotopy class of the image of γ under f.

Proof There are two things we need to do here: checking that f_* is well-defined, and then checking that it is a homomorphism. Let us begin by checking that it is well-defined.

Suppose γ, γ' are both representatives of the same homotopy class in (X, x), i.e. they are homotopic loops in X based at x. We must show that $f \circ \gamma$ and $f \circ \gamma'$ are in the same homotopy class in (Y, y), i.e. they are homotopic loops in Y based at y. To see this, note that since γ and γ' are homotopic, there is some homotopy $H : [0, 1] \times [0, 1] \to X$ so that $H(0, t) = \gamma(t)$, $H(1, t) = \gamma'(t)$, and $H(s, 0) = H(s, 1) = x$.

[1] This homomorphism is sometimes also written $\pi_1(f)$.

Now, we claim that $f \circ H : [0,1] \times [0,1] \to Y$ is a homotopy between $f \circ \gamma$ and $f \circ \gamma'$. This is because $f \circ H$ is a composition of two continuous functions and is thus continuous, and we have $f \circ H(0,t) = f(\gamma(t))$, $f \circ H(1,t) = f(\gamma'(t))$, and $f \circ H(s,0) = f \circ H(s,1) = y$. So, this shows that $f \circ \gamma$ and $f \circ \gamma'$ are homotopic. Hence f_* is well-defined.

Now, we must show that f_* is a homomorphism. In other words, if γ and γ' are two loops in X based at x, we must show that $f_*([\gamma * \gamma']) = f_*([\gamma]) \cdot f_*([\gamma'])$. But this is clear, since $(f \circ \gamma) * (f \circ \gamma') = f \circ (\gamma * \gamma')$. (You should check this for yourself using the definition of path concatenation $*$. You'll find that they are *exactly* the same, not just the same up to homotopy. This is quite a rare event in algebraic topology!) Hence f_* is a homomorphism. ■

Let us look at some examples of what f_* does.

Example

(1) If $X = Y$ and $x = y$, and f is the identity map, then f_* is the identity homomorphism.
(2) If $X = Y = \mathbb{S}^1$ and f is the map given by $f(e^{i\theta}) = e^{2i\theta}$ (i.e. the map that wraps the circle around itself twice), then f_* is multiplication by 2. (That is, it sends an element of $\pi_1(\mathbb{S}^1) = \mathbb{Z}$ to its double.)
(3) Suppose $X = \mathbb{T}^2 = \mathbb{S}^1 \times \mathbb{S}^1$ and $Y = \mathbb{S}^1$, and f sends a point $(e^{i\theta}, e^{i\phi})$ to $e^{i\theta}$. Then f_* sends $(a,b) \in \mathbb{Z}^2 = \pi_1(\mathbb{T})$ to $a \in \mathbb{Z} = \pi_1(\mathbb{S}^1)$.

We now derive a further elementary property of the induced homomorphism. First, if we have three spaces X, Y, and Z, and maps $f : X \to Y$ and $g : Y \to Z$, then the induced homomorphism of the composed map $(g \circ f)_*$ can be related to the induced homomorphisms f_* and g_*. In fact, the relation is exactly what you (probably) expect!

Proposition 10.17 *If $f : (X,x) \to (Y,y)$ and $g : (Y,y) \to (Z,z)$ are continuous maps, then $(g \circ f)_* = g_* \circ f_*$.*

Proof Let $[\gamma]$ be a homotopy class of loops in X based at some x. Then $(g \circ f)_*([\gamma]) = [g \circ f \circ \gamma] = g_*([f \circ \gamma]) = g_*(f_*([\gamma]))$. Since $[\gamma]$ was arbitrary, this proves the claim. ■

We saw earlier that if $f : X \to X$ is the identity map, then f_* is the identity homomorphism. We can replace f by any homeomorphism, and almost the same result still holds.

Proposition 10.18 *Suppose $f : (X,x) \to (Y,y)$ is a homeomorphism. Then f_* is an isomorphism.*

Proof Let $g : (Y,y) \to (X,x)$ be the inverse homeomorphism. Then $g \circ f$ is the identity map on X, and $f \circ g$ is the identity map on Y. Hence $(g \circ f)_*$ is the identity map on $\pi_1(X)$, and $(f \circ g)_*$ is the identity map on $\pi_1(Y)$. But, using Proposition 10.17, we have that $g_* \circ f_*$ and $f_* \circ g_*$ are the identity maps, and so g_* and f_* are inverses. Hence, f_* is an isomorphism. ■

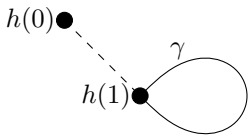


Figure 10.3 The map β_h turns γ into $h * \gamma * h^{-1}$.

As a very simple yet far-reaching consequence, we have shown that π_1 is a *homeomorphism invariant*: if (X, x) and (Y, y) are homeomorphic, then $\pi_1(X, x) \cong \pi_1(Y, y)$. Of course, we already proved this as well as the stronger statement that if (X, x) and (Y, y) are homotopy equivalent then $\pi_1(X, x) \cong \pi_1(Y, y)$ in Chapter 9, in a different way.

Note the following possible pitfall: if $f : X \to X$ is a homeomorphism from X to itself, then f_* is an isomorphism. But it is not necessarily the identity map, as the following example shows.

Example Let $X = \mathbb{S}^1$, and let $f : X \to X$ be the map given by $f(e^{2\pi it}) = e^{-2\pi it}$. Then f is a homeomorphism, but the induced homomorphism on fundamental groups is $f_* : \mathbb{Z} \to \mathbb{Z}$, given by $f(n) = -n$.

In fact, we can do quite a bit better. Not only is f_* an isomorphism when f is a homeomorphism; it's still an isomorphism even if f is only a homotopy equivalence:

Theorem 10.19 *Suppose that $f : (X, x) \to (Y, y)$ is a homotopy equivalence. Then f_* is an isomorphism.*

In order to prove this, we first need a lemma.

Lemma 10.20 *Let $x_0 \in X$ be a basepoint. Suppose $H : [0, 1] \times X \to Y$ is a homotopy of maps from X to Y, and $h : [0, 1] \to Y$ is the path given by $h(t) = H(t, x_0)$. Let $\beta_h : \pi_1(Y, h(1)) \to \pi_1(Y, h(0))$ be the map on fundamental groups given by concatenating a loop in Y based at $h(1)$ with h and $\bar{h}$. (See Figure 10.3.) Let $H_{1*} : \pi_1(X, x_0) \to \pi_1(Y, h(1))$ and $H_{0*} : \pi_1(X, x_0) \to \pi_1(Y, h(0))$ be the induced maps. Then*

$$H_{0*} = \beta_h \circ H_{1*}.$$

Proof Let h_t be the restriction of h to $[0, t]$, reparametrized so that its domain is $[0, 1]$, i.e. $h_t(s) = h(st)$. Let γ be a loop in X based at x_0. Then the concatenation $h_t * H(t, f(x)) * \bar{h}_t$ is a homotopy of loops based at $h(0)$. Restricting to $t = 0$ and $t = 1$ says that

$$H(0, \gamma(t)) \sim \beta_h(H(1, \gamma(t))),$$

as desired. ■

We can now prove Theorem 10.19.

Proof of Theorem 10.19 Let $f : (X, x) \to (Y, y)$ be a homotopy equivalence. Let $g : (Y, y) \to (X, x)$ be a homotopy equivalence in the reverse direction, so that $f \circ g$ and $g \circ f$ are homotopic to their respective identities. Since $g \circ f$ is homotopic to the identity, Lemma 10.20 implies that there is some path h, giving rise to a suitable β_h, for which $g_* \circ f_* = \beta_h$. (Here, we are taking H_{1*} to be the identity and H_{0*} to be $(g \circ f)_* = g_* \circ f_*$.) Since β_h is an isomorphism, f_* is injective. Repeating the argument with the other composition similarly shows that f_* is surjective. Hence f_* is an isomorphism. ■

10.5 Retracts

One reason to study the induced homomorphism is that it allows us to pick up on topological features of the spaces and the map between them that are not readily available on their own. For example, certain types of maps between spaces induce injective or surjective maps on fundamental groups, and we can use this fact to learn more about the topology of the spaces.

Earlier, we were trying to show that $\mathbb{S}^2$ is not contractible, and we were almost able to do it. However, we were missing one key ingredient: that there is no continuous map r from the solid sphere B to the boundary sphere $\mathbb{S}^2$ such that the restriction of r to the boundary is the identity. At the moment, we still won't be able to prove that no such r exists exactly. But using the induced map, we will be able to do this in one dimension lower: There is no retract from the disk to the boundary circle $\mathbb{S}^1$.

Definition 10.21 Suppose that X is a topological space, and A is a subspace of X. Then a continuous map $r : X \to A$ is called a *retract* if the restriction of r to A is the identity map.

We will also find the following special type of retract important.

Definition 10.22 A retract $r : X \to A$ is called a *deformation retract* of X if r is homotopic to the identity map on X.

Remark 10.23 Observe a slight sloppiness in notation here: The codomain of r is A, but we are asking for it to be homotopic to a map whose codomain is X. If we were to be more pedantic, we would treat r as a map from $X \to X$ whose image is A, rather than a map from X to A.

We focus on retracts and deformation retracts because they behave nicely with respect to the induced map on fundamental groups.

Theorem 10.24 *Let $r : X \to A$ be a retract, let $\iota : A \hookrightarrow X$ be the inclusion map from A into X, and let $x \in A$ be a basepoint. Then ι_* is injective. If r is a deformation retract, then ι_* is an isomorphism.*

Proof If r is a retract, then $r \circ \iota$ is the identity map on A. Hence

$$r_* \circ \iota_* = (r \circ \iota)_* = \mathrm{id}_*$$

is the identity. This means that r_* is surjective and ι_* is injective by elementary properties of injectivity and surjectivity.[2] Now, suppose that r is a deformation retract. We can show that ι_* is also surjective as follows. In this case, we have a homotopy H between r and the identity on X. Suppose that γ is a loop in X based at $x \in A$. We want to show that γ is homotopic to some loop γ' in A, which would show that $[\gamma] = \iota_*[\gamma']$. To show this, we simply compose γ with H, which gives a homotopy from γ to a loop in A. ■

As a consequence, we can prove that there is no retract from a disk to a circle.

Theorem 10.25 *Let D be a disk and $\mathbb{S}^1$ its boundary circle. There is no retract from D to $\mathbb{S}^1$.*

Proof Suppose there were such a retract $r : D \to \mathbb{S}^1$, and let $\iota : \mathbb{S}^1 \hookrightarrow D$ be the inclusion map. Then, by Theorem 10.24, ι_* would be an injection. However, $\pi_1(\mathbb{S}^1) \cong \mathbb{Z}$, and $\pi_1(D) = \{0\}$, so there is no injection from $\pi_1(\mathbb{S}^1)$ to $\pi_1(D)$, and hence no retract $D \to \mathbb{S}^1$. ■

So, we see that the group theory can tell us about existence of retracts. We can give a more general group-theoretic statement about the existence (or non-existence) of retracts. If A is a retract of X, then we have the following maps on spaces:

$$A \overset{\iota}{\hookrightarrow} X \overset{r}{\to} A,$$

where the composition is the identity. These maps induce maps on fundamental groups:

$$\pi_1(A) \overset{\iota_*}{\hookrightarrow} \pi_1(X) \overset{r_*}{\to} \pi_1(A),$$

where again the composition is the identity. Hence, we can think of $\pi_1(A)$ as being a subgroup of $\pi_1(X)$. If $\pi_1(A)$ is normal in $\pi_1(X)$, then $\pi_1(X)$ can be written as a direct product of $\pi_1(A)$ and the kernel of r_*. If $\pi_1(A)$ is not normal, then $\pi_1(X)$ is a *semidirect product* of $\pi_1(A)$ and the kernel of r_*. Being a direct product or a semidirect product is quite a special and unusual thing in group theory!

Deformation retracts are especially nice because, as they are defined, they are homotopy equivalences. Hence, they induce isomorphisms on fundamental groups. As a result, we now have a wider class of spaces whose fundamental groups we can compute.

---

[2]Reminder: Suppose that $f : A \to B$ and $g : B \to A$ are two maps such that $f \circ g = \mathrm{id}$. Now, if $g(x) = g(y)$, then applying f to both sides yields $x = y$, hence g is injective. Also, in order to solve the equation $f(x) = y$ given y, we simply use $x := g(y)$. Hence f is surjective.

Example

- The fundamental group of an annulus (for example, the set $\{(x, y) \in \mathbb{R}^2 : 1/2 \leq x^2 + y^2 \leq 2\}$) has fundamental group $\mathbb{Z}$, because it deformation retracts onto a circle.
- The solid torus also deformation retracts onto a circle (namely the circle in the center of the solid torus with the same axis of rotational symmetry as the solid torus), so its fundamental group is also $\mathbb{Z}$.

10.6 Problems

(1) Compute the fundamental groups of the following spaces:

   (a) $\mathbb{S}^1 \times \mathbb{S}^1 \times \mathbb{S}^1 \times \mathbb{S}^1 \times \mathbb{S}^1$
   (b) $\mathbb{S}^1 \times \mathbb{S}^1 \times \mathbb{S}^1 \times \mathbb{S}^2$
   (c) $\mathbb{S}^1 \times \mathbb{S}^2 \times \mathbb{S}^2$

(2) Let $\gamma : [0, 1] \to \mathbb{S}^1$ be a loop and let $f : \mathbb{S}^1 \to \mathbb{S}^1$ be a homeomorphism. What can you say about $\deg_{(0,0)}(f \circ \gamma)$?

(3) How can you describe or visualize—or otherwise develop some intuition for—the space $\mathbb{S}^1 \times \mathbb{S}^1 \times \mathbb{S}^1$?

(4) In each of the questions below, build explicit homotopies.

   (a) Show that a line segment in $\mathbb{R}^n$ is contractible.
   (b) Show that the image of a path $\gamma : [0, 1] \to \mathbb{R}^n$ without self-intersections is contractible.

(5) A connected space is called *simply connected* if its fundamental group is trivial. Show that a space S is simply connected if and only if, for all x_0 and x_1 in S, all paths joining x_0 and x_1 are homotopic.

(6) Suppose S is a contractible space and let $F : S \to S'$ be a homeomorphism. Show that S' is contractible.

(7) Answer any three of the following questions.

   (a) Give a formula for a deformation retraction from $\mathbb{R}^3$ to the z-axis.
   (b) Give a formula for a deformation retraction from the solid torus in $\mathbb{R}^3$ obtained by rotating the circle $(x - 2)^2 + z^2 \leq 1$ in the xz-plane around the z-axis, to the annulus in the xy-plane.
   (c) Give a formula for a deformation retraction from the solid torus in $\mathbb{R}^3$ to its core circle.
   (d) Is the unit circle in the xy-plane a deformation retract for $\mathbb{R}^2$? If so, write down the retraction; if not, prove it. Answer the same question for $\mathbb{R}^2 \setminus \{(0, 0)\}$.
   (e) Is the unit circle in the xy-plane a deformation retract for $\mathbb{R}^3 \setminus z$-axis? If so, write down the retraction; if not, prove it.

(8) Find a space S and a subset $A \subseteq S$ that is a retract of S but not a deformation retract.

(9) Let $S = \overline{B_1(0)} \setminus \{0\}$ in $\mathbb{R}^2$, and let $A \subseteq S$ be the unit circle. Let $\phi : S \to A$ be given by $\phi(x) = x/\|x\|$ and let $\psi : A \to S$ be given by $\psi(x) = x$. (In other words, ψ is the inclusion map.) Verify that ϕ and ψ are homotopy equivalences—namely that $\phi \circ \psi : A \to A$ and $\psi \circ \phi : S \to S$ are homotopic to the identity mappings $\mathrm{id}_A : A \to A$ and $\mathrm{id}_S : S \to S$, respectively.

(10) Let $\phi, \phi' : S \to S'$ be two maps such that there is a subset $A \subseteq S$ on which ϕ and ϕ' are identical, i.e. $\phi(a) = \phi'(a)$ for all $a \in A$. We say ϕ and ϕ' are *homotopic relative to* A if all the intermediate maps in the homotopy are identical on A, i.e. if there is a homotopy $H(s, x)$ such that $H(s, a) = \phi(a)$ for all $s \in [0, 1]$ and for all $a \in A$.

(a) Find a pair of (different) maps from the complement of the open unit ball in $\mathbb{R}^2$ to the unit circle in $\mathbb{R}^2$ that are homotopic relative to the unit circle in $\mathbb{R}^2$.

(b) Show that if ϕ_0 and ϕ_1 are continuous maps $X \to Y$ with $\phi_0(x_0) = \phi_1(x_0) = y_0$, and if these two maps are homotopic relative to the subset $\{x_0\} \subset X$, then the two induced homomorphisms $\phi_{0*}, \phi_{1*} : \pi_1(X, x_0) \to \pi_1(Y, y_0)$ are identical.

(11) (a) Let X be the space obtained by forming the connected sum of two infinite cylinders. Show that X is homotopy equivalent to a union of two circles that intersect at two points (just like, for example, two different great circles on the sphere).

(b) Suppose instead that X is obtained by taking the connected sum of two Möbius strips (where the disks we cut out of these Möbius strips do not touch the boundary). Is X homotopy equivalent to some union of curves, as in the first part of this problem? If so, describe it.

Chapter 11
Applications of Fundamental Groups

In this chapter, we will see several applications of the formalism we have developed so far. The first is a proof of an extremely important result that you already know—the Fundamental Theorem of Algebra! The second is a result known as the Borsuk–Ulam Theorem, a corollary of which is the amusingly named "Avocado Sandwich Theorem." The third is a result known as the Brouwer Fixed-Point Theorem.

11.1 The Fundamental Theorem of Algebra

Complex Numbers. We begin with a quick review of some facts about complex numbers. Complex numbers are numbers of the form $z = x + iy$, where $x, y \in \mathbb{R}$ and i satisfies $i^2 = -1$. In other words, i and also $-i$ are roots of the polynomial $p(z) = z^2 + 1$.

- We can add two complex numbers: $(x_1 + iy_1) + (x_2 + iy_2) = (x_1 + x_2) + i(y_1 + y_2)$.
- We can multiply a complex number by a real number: $c(x + iy) = cx + icy$.

Thus, in this respect, complex numbers behave just like points in $\mathbb{R}^2$—the complex number $x + iy$ becomes the point (x, y) and now addition and real number multiplication of complex numbers become vector addition and scalar multiplication in $\mathbb{R}^2$.

- We can also multiply one complex number by another. The product $(x_1 + iy_1)(x_2 + iy_2)$ is found by fully multiplying these two brackets out, and replacing i^2 by -1 when it occurs. The answer is $(x_1x_2 - y_1y_2) + i(y_1x_2 + y_2x_1)$.

In $\mathbb{R}^2$, we can use polar coordinates to represent points. We represent $(x, y) \in \mathbb{R}^2$ by its distance from the origin $r = \sqrt{x^2 + y^2}$, and the angle θ made by the line

C. Bray et al., *Algebraic Topology*,
https://doi.org/10.1007/978-3-030-70608-1_11

connecting (x, y) to $(0, 0)$, so that $\tan(\theta) = y/x$. Now $(x, y) = (r\cos(\theta), r\sin(\theta))$. We can thus also use polar coordinates to describe complex numbers.

- The length of a complex number $z = x + iy$ is denoted $|z| = \sqrt{x^2 + y^2}$.
- The polar angle of z is denoted $\arg(z)$, and $\tan(\arg(z)) = y/x$.
- Now $z = |z|\big(\cos(\arg(z)) + i\sin(\arg(z))\big)$.
- De Moivre's Theorem states that $(\cos(\theta) + i\sin(\theta))^n = \cos(n\theta) + i\sin(n\theta)$. This formula can be proven easily by induction on n, using the angle-sum formulae for cosine and sine. It follows that $z^n = |z|^n\big(\cos(n\arg(z)) + i\sin(n\arg(z))\big)$.
- We therefore define $e^{i\theta} = \cos(\theta) + i\sin(\theta)$, because it has the analogous property $\big(e^{i\theta}\big)^n = e^{in\theta}$. (There's more to this, but we won't get into it here.)
- Note that $|e^{i\theta}| = 1$ and that as $\theta \in [0, 2\pi]$ advances, the curve $\alpha(\theta) = e^{i\theta}$ traces out the unit circle in $\mathbb{C}$ when it is viewed as $\mathbb{R}^2$.
- As a consequence, the curve $\gamma(\theta) = e^{in\theta}$ traces out n windings of the unit circle for $\theta \in [0, 2\pi]$.

Polynomials. A polynomial of degree n is a very familiar object. A real polynomial is a function $p : \mathbb{R} \to \mathbb{R}$ of the form

$$p(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0,$$

with the a_i's in $\mathbb{R}$. A real root of a polynomial is any number $x_0 \in \mathbb{R}$ such that $p(x_0) = 0$. If a polynomial has a root, then it can be factored as $p(x) = (x - x_0)q(x)$, where q is a polynomial of degree $n - 1$. There are real polynomials that do not have roots, such as $p(x) = x^2 + 1$. But when we expand the domain of this polynomial from the real numbers to the complex numbers, we see that it has two complex roots $\pm i$.

It is in fact very fruitful to view a polynomial as a function $p : \mathbb{C} \to \mathbb{C}$ by allowing complex inputs to be used; we simultaneously also allow the coefficients a_i to lie in $\mathbb{C}$. The output is thus a complex number as well. Note that we now can also view p as a function $p : \mathbb{R}^2 \to \mathbb{R}^2$ by making the association of complex numbers to points in the plane as described earlier.

When do polynomials have roots in $\mathbb{C}$? In other words, we are looking for a point $z_0 \in \mathbb{C}$ so that $p(z_0) = 0$. The answer to this question is nearly always!

Theorem 11.1 (Fundamental Theorem of Algebra) *Every nonconstant polynomial has at least one root in* $\mathbb{C}$.

Applying this theorem to an arbitrary polynomial p of degree n and using factorization, we can now write $p = (z - z_0)q$, where q is a polynomial of degree $n - 1$. Hence we can apply the theorem to factor q further. Therefore the Fundamental Theorem of Algebra actually tells us that the polynomial p always has n roots and can always be written as a product of n factors of degree one.

A Preliminary Topological Result. Before moving on to the proof of the Fundamental Theorem of Algebra, we derive a preliminary result. This is where topology and induced homomorphisms come in!

Lemma 11.2 *Let $\phi : \overline{B_1(0,0)} \to \mathbb{R}^2 \setminus \{(0,0)\}$ be a continuous mapping, and let γ denote the boundary circle of $\overline{B_1(0,0)}$. Then $\phi(\gamma)$ is trivial in $\pi_1(\mathbb{R}^2 \setminus \{(0,0)\})$.*

This result says that if the image of $\overline{B_1(0,0)}$ under ϕ is in $\mathbb{R}^2 \setminus \{(0,0)\}$, then the image of the boundary of $\overline{B_1(0,0)}$ can't wrap around the origin in a nontrivial manner. Draw some pictures to convince yourself that this is an "obvious" result. But, as is often the case, the "obvious" results are very hard to prove using elementary techniques and can only be tackled successfully using more sophisticated tools. See if you can do it without invoking the fundamental group!

Proof We have the mappings

$$\gamma \xrightarrow{i} \overline{B_1(0,0)} \xrightarrow{\phi} \mathbb{R}^2 \setminus \{(0,0)\},$$

where $i : \gamma \to \overline{B_1(0,0)}$ is the inclusion. We thus have the induced mappings

$$\pi_1(\gamma) \xrightarrow{i_*} \pi_1(\overline{B_1(0,0)}) \xrightarrow{\phi_*} \pi_1(\mathbb{R}^2 \setminus \{(0,0)\})$$

between fundamental groups. Because $\pi_1(\overline{B_1(0,0)}) = \{e\}$, all these mappings must be trivial. Consequently $\phi_*([\gamma]) = e$. By definition, this says that $[\phi \circ \gamma] = [\text{const}]$. This in turn means that $\phi \circ \gamma$ is homotopic to the trivial loop. ■

Proof of the Fundamental Theorem of Algebra. Let $p(z) = z^n + a_{n-1}z^{n-1} + \cdots + a_1 z + a_0$. Note that from the point of view of proving the existence of a root, it is sufficient to assume that the leading coefficient of p is 1. Assume that p has no roots. Hence, as a mapping of topological spaces, we have $p : \mathbb{R}^2 \to \mathbb{R}^2 \setminus \{(0,0)\}$. Assume also that $|a_0| + |a_1| + \cdots + |a_{n-1}| \leq \frac{1}{2}$. We'll remove this assumption later.

Consider the map $F : [0,1] \times \mathbb{R}^2 \to \mathbb{R}^2$ given by $F(s,z) = z^n + s\big(a_{n-1}z^{n-1} + \cdots + a_0\big)$. This is a homotopy of maps from p to the very simple polynomial $q(z) = z^n$. Let us investigate the behaviour of this homotopy of maps *restricted* to the unit circle in $\mathbb{R}^2$. This simply means that we must consider the map $\widehat{F} : [0,1] \times \gamma \to \mathbb{R}^2$ given by

$$\widehat{F}(s,\theta) = F(s, z = e^{i\theta}).$$

We claim that in fact $F : [0,1] \times \gamma \to \mathbb{R}^2 \setminus \{(0,0)\}$. To prove this claim, note first that it is already true for $\widehat{F}(1,\theta) = p(e^{i\theta})$, because we know that p maps all of $\mathbb{R}^2$ into $\mathbb{R}^2 \setminus \{(0,0)\}$. It is also true for $\widehat{F}(0,\theta) = e^{in\theta}$, because this is an n-fold winding of the unit circle. For intermediate values of s, we compute a sequence of inequalities from which we derive the conclusion. These inequalities are:

$$\begin{aligned}
|\widehat{F}(s,\theta)| &= |F(s, z = e^{i\theta})| \\
&= \left|z^n + s(a_{n-1}z^{n-1} + \cdots + a_1 z + a_0)\right| \\
&\geq \left|z^n\right| - s\left|a_{n-1}z^{n-1} + \cdots + a_1 z + a_0\right| \\
&\geq |z|^n - s\big(|a_{n-1}||z|^{n-1} + \cdots + |a_1||z| + |a_0|\big) \\
&= 1 - s(|a_{n-1}| + \cdots + |a_1| + |a_0|) \\
&\geq 1 - s/2 \\
&\geq 1/2\,.
\end{aligned}$$

The first inequality is a version of the *triangle inequality*, namely $|A + B| \geq |A| - |B|$. The second inequality is another version of the triangle inequality, namely $|A_1 + A_2 + \cdots + A_N| \leq |A_1| + |A_2| + \cdots + |A_N|$. We've also used De Moivre's Theorem to say $|z^n| = |z|^n$. The third equality follows because $|z| = 1$ when $z = e^{i\theta}$. The fourth inequality follows from our assumption about the size of the coefficients of p. The final inequality follows since $s \in (0, 1)$. As a consequence of the entire sequence of inequalities, we can now say that $|\widehat{F}(s,\theta)| \geq 1/2$. This means that the distance of $\widehat{F}(s,\theta)$ from the origin is always greater than $1/2$. Therefore we can assert that $\widehat{F}(s,\theta)$ is a homotopy between $\widehat{F}(0,\theta)$ and $\widehat{F}(1,\theta)$ that is entirely within $\mathbb{R}^2 \setminus \{0\}$.

Now what do we have? We know that $\pi_1(\mathbb{R}^2 \setminus \{(0,0)\}) \cong \mathbb{Z}$. On the one hand, the curve $\widehat{F}(0,\theta)$ is homotopically nontrivial in $\pi_1(\mathbb{R}^2 \setminus \{(0,0)\})$, because it wraps around the origin n times. However, the curve $\widehat{F}(1,\theta)$ must be homotopically trivial because of the Preliminary Topological Result: the curve $\widehat{F}(1,\theta)$ is the image of the boundary of the unit circle under p, and p maps the whole unit ball to $\mathbb{R}^2 \setminus \{(0,0)\}$. Hence we have a homotopy connecting a nontrivial curve to a trivial one. This is a contradiction.

We have thus almost proved the Fundamental Theorem of Algebra. All that remains is to remove the assumption on the size of the coefficients of p. To this end, let p now be a polynomial with coefficients as large as we want. Let C be a large real number. Then the polynomial $q(z) = p(Cz)/C^n$ is a new polynomial with the property that $q(z) = 0$ if and only if $p(z/C) = 0$. Thus, if we demonstrate the existence of a root of q, then we have a root of p. But

$$q(z) = z^n + \frac{a_{n-1}}{C}z^{n-1} + \cdots + \frac{a_1}{C^{n-1}}z + \frac{a_0}{C^n}\,.$$

Hence, by choosing C to be sufficiently large, we can make the coefficients of q be as small as we need. In this way we can reduce the case of a polynomial with general coefficients to the special case discussed above.

11.2 Further Applications of the Fundamental Group

The Borsuk–Ulam Theorem. The Borsuk–Ulam Theorem is a classical result in topology that asserts the existence of a special kind of point (the solution of an equation) based on very minimal assumptions!

Theorem 11.3 (Borsuk–Ulam) *Suppose $f : \mathbb{S}^2 \to \mathbb{R}^2$ is a continuous function from the sphere to the plane. Then there exists $x \in \mathbb{S}^2$ so that $f(x) = f(-x)$.*

Therefore, no matter the function, there exist two antipodal points on the sphere with identical function values. A surprising, silly application of this theorem (which everyone is contractually obligated to mention when first discussing the Borsuk–Ulam Theorem) is that there exists a pair of antipodal points on the surface of the Earth (or a perfectly spherical version of the Earth, at least, so that we can define antipodes) where the temperature and atmospheric pressure are exactly the same.

Proof Suppose the Borsuk–Ulam Theorem is false. Define a new function $\widehat{f} : \mathbb{S}^2 \to \mathbb{R}^2$ by $\widehat{f}(x) := f(x) - f(-x)$, which by our assumption on the falsehood of the Borsuk–Ulam Theorem is actually a function $\widehat{f} : \mathbb{S}^2 \to \mathbb{R}^2 \setminus \{(0,0)\}$. Note that $\widehat{f}$ is an *odd function* because $\widehat{f}(-x) = -\widehat{f}(x)$. Next, let $\alpha : [0, 2\pi] \to \mathbb{S}^2$ be the curve that winds once around the equator, i.e. $\alpha(s) := (\cos(s), \sin(s), 0)$. The curve $\widehat{f} \circ \alpha : [0, 2\pi] \to \mathbb{R}^2 \setminus \{(0,0)\}$ now winds a certain number of times around the origin $(0,0)$; this number is $\deg_{(0,0)}(\widehat{f} \circ \alpha)$. This number *has* to be zero because we can easily construct a homotopy of the curve α to a point by sliding it upwards to the north pole of $\mathbb{S}^2$. Hence this homotopy must also allow the curve $\widehat{f} \circ \alpha$ to shrink continuously to a point inside $\mathbb{R}^2 \setminus \{(0,0)\}$. Consequently, the degree of the curve $\widehat{f} \circ \alpha$, as defined in Chapter 10, is zero.

We can now reach a contradiction, because we can actually show that the degree of $\widehat{f} \circ \alpha$ has to be an odd number. This is due to the following calculation:

$$\begin{aligned}
\widehat{f} \circ \alpha(s + \pi) &= \widehat{f}(\cos(s + \pi), \sin(s + \pi), 0) \\
&= \widehat{f}(-\cos(s), -\sin(s), 0) \\
&= -\widehat{f}(\cos(s), \sin(s), 0) \\
&= -\widehat{f} \circ \alpha(s).
\end{aligned}$$

Therefore there is an integer m such that the first half of the interval $[0, \pi]$ is wound $m + \frac{1}{2}$ times around the origin by $\widehat{f} \circ \alpha$, while the second half of the interval is also wound $m + \frac{1}{2}$ times around the origin. In total, $\widehat{f} \circ \alpha$ winds $2m + 1$ times around the origin. This is our contradiction; hence the Borsuk–Ulam Theorem must be true. ■

Exercise 11.4 Make the final argument in the preceding proof more formal using the definition of the degree of a curve from Chapter 10.

Remark 11.5 Thereis also an n-dimensional version of the Borsuk–Ulam Theorem, which says that if $f : \mathbb{S}^n \to \mathbb{R}^n$ is continuous, then there is some point $x \in \mathbb{S}^n$ such that $f(x) = f(-x)$. The proof technique given above does not work, because the "equatorial" $\mathbb{S}^{n-1}$ on $\mathbb{S}^n$ has trivial fundamental group. However, the higher homotopy groups and homology groups can be used to give an analogous proof.

A corollary of the Borsuk–Ulam theorem is the so-called "Avocado Sandwich Theorem" This states that it is always possible to cut a sandwich (consisting of two slices of bread on either side of delicious contents) into two pieces, and each half contains equal volumes of bread from each of the slices and of the avocado. More formally:

Theorem 11.6 (Avocado Sandwich Theorem) *Let A, B, C be compact subsets of $\mathbb{R}^3$. Then there is a plane that simultaneously divides each of A, B, C into two pieces of equal volume.*

Here we should think of A and C as the two slices of bread, B as the avocado, and the plane as the direction in which the knife moves as it cuts the sandwich.

Proof We're going to define a function $f : \mathbb{S}^2 \to \mathbb{R}^2$ to which we'll apply the Borsuk–Ulam Theorem. To this end, pick $x \in \mathbb{S}^2$ and view it as a vector in $\mathbb{R}^3$. Let P_x be a plane normal to x that cuts A into two halves of equal volume, and let P_x^+ and P_x^- denote the half spaces *above* and *below* P_x, respectively—i.e. the regions into which the vector x points into and away from, respectively. Now, there may be several such planes, for instance if A is disconnected; if that is the case, then the union of all these planes is a direct product of $\mathbb{R}^2$ with an interval—an interval's worth of planes. When this happens, choose P_x to be the central plane in this interval of planes, i.e. the one corresponding to the midpoint of the interval. Now define

$$f(x) := \begin{pmatrix} \text{Volume of} \\ P_x^+ \cap B \end{pmatrix}, \begin{pmatrix} \text{Volume of} \\ P_x^+ \cap C \end{pmatrix}.$$

We omit the proof that f is continuous. However, this is true. (Can you argue why?) So we can apply the Borsuk–Ulam Theorem to find an x so that $f(x) = f(-x)$. But

$$\begin{aligned} f(-x) &= \begin{pmatrix} \text{Volume of} \\ P_{-x}^+ \cap B \end{pmatrix}, \begin{pmatrix} \text{Volume of} \\ P_{-x}^+ \cap C \end{pmatrix} \\ &= \begin{pmatrix} \text{Volume of} \\ P_x^- \cap B \end{pmatrix}, \begin{pmatrix} \text{Volume of} \\ P_x^- \cap C \end{pmatrix} \end{aligned}$$

because the region *above* the plane P_{-x} is the same as the region *below* the plane P_x. Therefore we have

$$\begin{aligned} \begin{pmatrix} \text{Volume of} \\ P_x^+ \cap B \end{pmatrix} &= \begin{pmatrix} \text{Volume of} \\ P_x^- \cap B \end{pmatrix}, \\ \begin{pmatrix} \text{Volume of} \\ P_x^+ \cap C \end{pmatrix} &= \begin{pmatrix} \text{Volume of} \\ P_x^- \cap C \end{pmatrix}, \end{aligned}$$

which proves the Avocado Sandwich Theorem. ■

Remark 11.7 Just as in the case of the Borsuk–Ulam Theorem, there is also a higher-dimensional Avocado Sandwich Theorem. It says that if $A_1, \ldots, A_n$ are n compact sets in $\mathbb{R}^n$, then there is a hyperplane dividing each one into two pieces of equal volume.

Another consequence of the Borsuk–Ulam Theorem is the Lyusternik–Shnirel'man Theorem.

Theorem 11.8 (Lyusternik–Shnirel'man) *Suppose that $A, B, C \subset \mathbb{S}^2$ are sets covering $\mathbb{S}^2$, i.e. $A \cup B \cup C = \mathbb{S}^2$. Suppose that A and B are either open or closed.*[1] *Then one of A, B, and C contains a pair $x^*, -x^*$ of antipodal points on $\mathbb{S}^2$.*

Remark 11.9 In the original Lyusternik–Shnirel'man Theorem, A, B, and C are required to be closed. Another popular version requires them all to be open. More recently, Greene in [Gre02] showed that the conclusion still holds if we only require that each one is either open or closed. But we don't even need that, because we make no assumption on C.

Proof Assume that none of A, B, and C contains any antipodal points. Let us write $d(x, A)$ for the distance from x to A, and similarly for the others. Let us define a function $f : \mathbb{S}^2 \to \mathbb{R}^2$ by

$$f(x) = (d(x, A), d(x, B)).$$

The function f is continuous, so by the Borsuk–Ulam Theorem, there are antipodal points x^* and $-x^*$ so that $f(x^*) = f(-x^*)$. Because C does not contain any pair of antipodal points, at least one of x^* and $-x^*$ is not in C. Say $x^* \notin C$ without loss of generality. Then x^* must be in one of the other two sets (i.e. A or B), say A without loss of generality. Since $f(x^*) = f(-x^*)$, this implies that $d(-x^*, A) = 0$, because certainly $d(x^*, A) = 0$.

Now, is $-x^* \in A$? If A is closed, then $d(-x^*, A) = 0$ implies that $-x^* \in A$. Thus A contains the antipodal points x^* and $-x^*$. On the other hand, if A is open, then $-x^*$ lies in the closure $\overline{A}$ of A. Since A does not contain antipodal points, $A \cap -A = \varnothing$, so $A \subset \mathbb{S}^2 \setminus -A$, and the latter is a closed set. Thus $\overline{A} \subset \mathbb{S}^2 \setminus -A$, because $\overline{A}$ is the smallest closed set containing A. Since $-x^* \in \overline{A}$, it follows that $-x^* \in \mathbb{S}^2 \setminus -A$, so $-x^* \notin -A$ and therefore $x^* \notin A$. But we selected A so that $x^* \in A$, so this is a contradiction. ∎

Remark 11.10 Here, too, there is an n-dimensional version. If $A_1, A_2, \ldots, A_{n+1} \subseteq \mathbb{S}^n$ cover $\mathbb{S}^n$, and $A_1, \ldots, A_n$ are open or closed, then one of the A_i's contains a pair of antipodal points.

Brouwer Fixed-Point Theorem. The Brouwer Fixed-Point Theorem is another classical result in topology that asserts the existence of a solution of a different kind of

[1] C does not need to be either open or closed. Also, it is allowed for one of A and B to be open and the other to be closed.

equation, again based on very minimal assumptions. It has wide-ranging implications, even in places where you would least expect it. For example, it is possible to use the Brouwer Fixed-Point Theorem to prove that the game of HEX cannot end in a draw; see [Gal79] for details.[2] For another application, one can prove that multiplayer games have Nash equilibria using the Brouwer Fixed-Point Theorem, as Nash did in [Nas51]. This is the most important theorem in economics.

Theorem 11.11 (Brouwer) *Let* $f : \overline{B_1(0,0)} \to \overline{B_1(0,0)}$ *be a continuous function of the closed unit disk in* $\mathbb{R}^2$ *to itself. Then* f *has a* fixed point*, i.e. there is an* $x \in \overline{B_1(0,0)}$ *so that* $f(x) = x$.

Proof Let us suppose that the Brouwer Fixed-Point Theorem is false and f has no fixed point. We can now reach a contradiction because, as we'll see momentarily, we'll be able to construct a retraction of $\overline{B_1(0,0)}$ onto its boundary $\mathbb{S}^1$. Of course this can't happen, because $\pi_1(\overline{B_1(0,0)})$ is trivial while $\pi_1(\mathbb{S}^1) \cong \mathbb{Z}$.

To construct the retraction, we proceed as follows. Since for every $x \in \overline{B_1(0,0)}$ we have $x \neq f(x)$, there is a well-defined line, let's call it L_x, between x and $f(x)$, which intersects the boundary of $\mathbb{S}^1$ in exactly two places. Define $r : \overline{B_1(0,0)} \to \mathbb{S}^1$ by

$$r(x) := \big(\text{the point in } L_x \cap \mathbb{S}^1 \text{ closer to } x \text{ than to } f(x)\big).$$

This function is continuous, as we can easily show by writing down an explicit formula for $r(x)$. (Do this!) Also, if $x \in \mathbb{S}^1$, then $r(x) = x$ because $f(x)$ is somewhere else in $\overline{B_1(0,0)}$, and so the closest point on $\mathbb{S}^1$ to x is just x itself. As a result, r satisfies the definition of a retraction. And we know that there is no such map, yielding the contradiction, so the Brouwer Fixed-Point Theorem must be true. ■

Remark 11.12 Unsurprisingly, there is also an n-dimensional version of the Brouwer Fixed-Point Theorem.

Remark 11.13 Both the Borsuk–Ulam Theorem and the Brouwer Fixed-Point Theorem can also be proved without using the fundamental group, but with the help of combinatorial lemmas called Tucker's Lemma and Sperner's Lemma, respectively. The combinatorial versions are more suited for *finding*—or at least approximating—the antipodal points and the fixed point guaranteed by the Borsuk–Ulam Theorem and the Brouwer Fixed-Point Theorem, respectively. See [Mat03] for Borsuk–Ulam and [AZ14, Chapter 27] and [Su99] for Brouwer. The combinatorial proofs also prove the n-dimensional versions without significant modification.

[2] Actually, if one knows that HEX cannot end in a draw, one can use that to prove the Brouwer Fixed-Point Theorem reasonably quickly as well! In this sense, the Brouwer Fixed-Point Theorem is *equivalent* to the fact that HEX cannot end in a draw.

11.3 Problems

(1) Find the fundamental groups of the following spaces by showing that they have the same homotopy type as spaces we are more familiar with.

(a) $\mathbb{R}^2 \setminus \{(0, 0)\}$
(b) $\mathbb{R}^3 \setminus \{(0, 0, 0)\}$
(c) $\mathbb{R}^3 \setminus z$-axis
(d) $\mathbb{R}^3 \setminus z$-axis and unit circle in the xy-plane
(e) $\mathbb{S}^2 \setminus \{\text{any two distinct points on } \mathbb{S}^2\}$

(2) (a) The Brouwer Fixed-Point Theorem states that every continuous mapping f from the disk to itself has a fixed point, i.e. there exists an x so that $f(x) = x$. Present a succinct yet complete version of the proof in more or less your own words.
(b) Let $f(x) = x + \varepsilon e^x$. Use the Brouwer Fixed-Point Theorem in the 1-dimensional case to prove that there is some open interval I containing 0, so that if $\varepsilon \in I$, then there exists x_ε so that $f(x_\varepsilon) = 0$. (Note: this result is easy to prove using the Intermediate Value Theorem, but you should not use the IVT in your solution to this problem.)

(3) (a) Let $A \subseteq \mathbb{R}^2$ be a compact subset. Give $\mathbb{R}^2$ the usual (x, y) coordinates and define the half-space $H_t := \{(x, y) \in \mathbb{R}^2 : y \geq t\}$. Define a function $f : \mathbb{R} \to \mathbb{R}$ by $f(t) = \text{Area}(A \cap H_t)$. Will f always be continuous?
(b) Show/discuss the continuity of the mapping of the Avocado Sandwich Theorem.

(4) (a) Is there a Borsuk–Ulam Theorem for the torus? That is, if $f : \mathbb{S}^1 \times \mathbb{S}^1 \to \mathbb{R}^2$ is a continuous map, must there be some $(x, y) \in \mathbb{S}^1 \times \mathbb{S}^1$ for which $f(x, y) = f(-x, -y)$?
(b) Does the Brouwer Fixed-Point Theorem hold for a torus? How about a sphere? Your favorite surface?

(5) Suppose that the wind is blowing on the surface of the earth in a constant and continuous fashion. Suppose also that at every point on the equator, the wind is blowing directly east, thus ensuring that the wind never blows anything from one hemisphere to the other. Show that there must exist some point in the northern hemisphere N such that a feather dropped at that point will return to its original location after exactly one minute. Assume that the equator is considered to be part of the northern hemisphere.

(6) Let X be the infinite cylinder, and Y be the punctured plane $\mathbb{R}^2 \setminus \{(0, 0)\}$. Find formulas for maps $f : X \to Y$ and $g : Y \to X$ that are homotopy equivalences, and find the homotopies to verify that the required compositions are homotopic to the identity mappings on X and Y respectively.

(7) Two thieves steal a necklace consisting of several jewels on a string. (The string is a straight line, not circular.) There are two types of jewels, and the number of jewels of each type is even. The thieves would like to split up the necklace so that

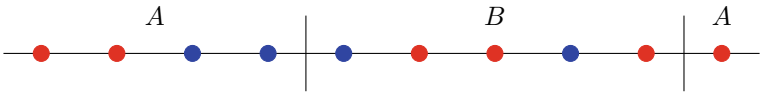


Figure 11.1 A 10-jewel necklace divided among two people with two cuts.

each thief receives the same number of jewels of each type, and they would like to do so by cutting the string in the smallest number of places. (See Figure 11.1.)

(a) Use the Borsuk–Ulam Theorem to show that there is always a way of splitting up the string using two cuts so that both people get the same number of jewels of each type.
(b) Using the Borsuk–Ulam Theorem in higher dimensions, prove that there is a way of splitting a string consisting of n types of jewels using n cuts.

Chapter 12
The Seifert–Van Kampen Theorem

12.1 The Fundamental Group of a Wedge of Circles

So far, we have learned how to compute fundamental groups for a few spaces, such as the circle, the sphere, the torus, and the annulus. But there are many more spaces whose fundamental groups we would like to know. In order to work them out, we will try to build them up from spaces whose fundamental groups we already know. Before we introduce the general theorem, let us look at an example, that of the wedge of two circles, meaning two circles that intersect at exactly one point (see Figure 12.1).

What should we expect the fundamental group of the wedge of two circles to be? Let us assume the basepoint is the point of intersection of the two circles. Then it seems we can go around the left circle as many times as we want, then around the right loop as many times as we want, then around the left loop as many times as we want, and so forth. Let a denote the loop that travels around the left loop once in the counterclockwise direction, and let b be the loop that travels around the right loop once in the counterclockwise direction. Then a typical element of the fundamental group will be $a^{i_1}b^{j_1}a^{i_2}b^{j_2}\cdots a^{i_r}b^{i_r}$, where the i_k's and j_k's are nonzero integers. (There are also other cases, such as starting with the a loop and ending with the a loop rather than the b loop, and so forth.) Furthermore, none of these loops seem as though they should be homotopic to each other. If we go around a bunch of a's and b's in various orders, that should never be homotopic to the constant loop that just stays at the basepoint. Recall from Chapter 5 that what we have just described is simply is the free group F_2 on two generators.

This intuition is correct, and now we wish to formalize it. To do this, let us call the wedge of two circles X, call the basepoint x, and call the left loop a and the right loop b.

Theorem 12.1 $\pi_1(X, x) \cong F_2$.

Proof We define a map $\phi : F_2 \to \pi_1(X, x)$. Let α and β be the generators of F_2. Because F_2 is a free group, we can define ϕ just by saying what $\phi(\alpha)$ and $\phi(\beta)$ are. So, we define $\phi(\alpha)$ to be the loop that goes around a once, and we define $\phi(\beta)$ to be

C. Bray et al., *Algebraic Topology*,
https://doi.org/10.1007/978-3-030-70608-1_12

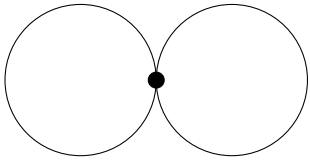


Figure 12.1 A wedge of two circles.

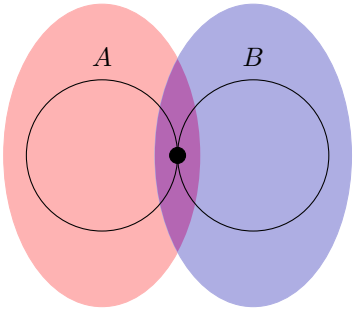


Figure 12.2 Breaking the wedge of two circles into two open sets.

the loop that goes around b once—in both cases in the counterclockwise direction. This map extends to a map from F_2 to $\pi_1(X, x)$; for example, $\phi(\alpha^3\beta^{-4}\alpha^2)$ is the loop that goes around a three times in the counterclockwise direction, then around b four times in the *clockwise* direction, then around a twice in the counterclockwise direction.

We need to show that ϕ is an isomorphism; that is, we need to show that it is surjective and injective. Let us start with surjectivity. First, we partition X into two open sets A and B: A is an open set slightly larger than the a circle, and B is an open set slightly larger than the b circle. (See Figure 12.2.) Let γ be a loop in X based at x. We think of γ as a map $[0, 1] \to X$ with $\gamma(0) = \gamma(1) = x$. We divide the interval $[0, 1]$ up into finitely many subintervals $[s_i, s_{i+1}]$, where $s_0 = 0$, $s_n = 1$, and for each i with $0 \leq i \leq n - 1$, $\gamma(t)$ restricted to $s_i \leq t \leq s_{i+1}$ is either entirely contained in A or entirely contained in B, so that they alternate: if $\gamma(t)$ restricted to $s_i \leq t \leq s_{i+1}$ is contained entirely in A, then $\gamma(t)$ restricted to $s_{i+1} \leq t \leq s_{i+2}$ is entirely contained in B, but not entirely contained in A, and vice versa. Let us also arrange things so that $\gamma(s_i) = x$ for all i. Note that we can break up the interval into only *finitely many* pieces in this way, because if the curve were to bounce back and forth between $A \setminus B$ and $B \setminus A$ infinitely often, then we would be able to find a sequence of points $0 < x_1 < x_2 < \cdots < 1$ (or possibly in the reverse order) so that $\gamma(x_{2n}) \in A \setminus B$ and $\gamma(x_{2n-1}) \in B \setminus A$, and thus $\lim_{n\to\infty} \gamma(x_n)$ would not exist—even though $\lim_{n\to\infty} x_n$ does—which contradicts the continuity of γ.

Now, since A and B are each homotopy equivalent to $\mathbb{S}^1$, we can homotope $\gamma(t)$ restricted to $s_i \leq t \leq s_{i+1}$ to some loop γ in the relevant $\mathbb{S}^1$: it is either some number of loops around a, or some number of loops around b. Hence this path is homotopic to a^k or b^k for some k, and every a^k or b^k can be obtained from $s_i \leq t \leq s_{i+1}$ by taking the appropriate path in that interval. Hence, when we put all these pieces together, we see that ϕ is surjective.

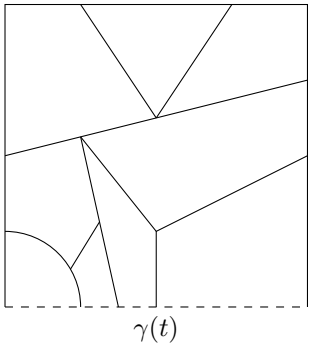


Figure 12.3 The curves drawn are the places where $F(s,t) = x$. Thus, in each region, $F(s,t)$ must either be entirely in the left loop or entirely in the right loop.

Now we must show injectivity. To do that, suppose that we have some element $w \in F_2$ so that $\phi(w)$ is the identity. We must show that w is the trivial word.[1] If $\phi(w)$ is the identity, then it means that the loop γ corresponding to w is homotopic to the constant loop in X. Let $F : [0,1] \times [0,1] \to X$ be a homotopy, so that $F(t,0) = \gamma(t)$, $F(t,1) = x$, and $F(0,s) = F(1,s) = x$. Draw curves in the square $[0,1] \times [0,1]$ where $F(s,t) = x$. These curves break up the square into several regions. In each of those regions, $F(s,t)$ is either entirely contained in the left loop or entirely contained in the right loop. Look at the regions bordering the bottom of the square. Each of these regions can be interpreted as a homotopy between a loop in either the left circle or the right circle, and the constant loop. Hence, each piece of γ contained in only one of the loops must be homotopic to the constant loop. This shows that ϕ is injective. See Figure 12.3. ■

Similarly, we can compute the fundamental group of a wedge of n circles meeting at a point. If we were to do this for $n = 3$, given the result above, we would break the wedge of three circles up into two pieces, A and B. The A piece is an open set slightly larger than the union of two of the circles, and the B piece is an open set slightly larger than the remaining circle. Thus A is homotopy equivalent to the wedge of two circles, and B is homotopy equivalent to one circle. (See Figure 12.4.) We would then run through a very similar argument to show that the fundamental group of the wedge of three circles is the free product $F_2 * \mathbb{Z} \cong F_3$: the free group on three generators. Continuing on by induction, we could show that the fundamental group of the wedge of n circles is $F_{n-1} * \mathbb{Z} \cong F_n$.

12.2 The Seifert–Van Kampen Theorem: First Version

This approach of breaking a space up into two simpler spaces, whose fundamental groups we already understand, is very useful for computing fundamental groups. We

[1] By a "word," we simply mean an element of the free group F_2, which is a string of symbols a, b, a^{-1}, and b^{-1}. The trivial word is the string with no characters in it.

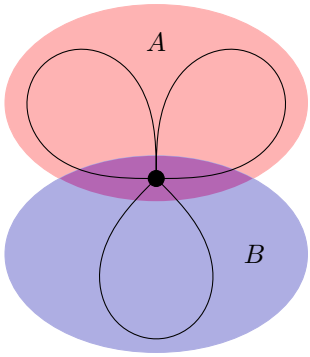


Figure 12.4 Breaking the wedge of three circles into two open sets.

will now prove a generalization of Theorem 12.1, known as the Seifert–Van Kampen Theorem. We will start by only stating and proving a special case of this theorem, but it is already fairly strong.

Theorem 12.2 (Seifert–Van Kampen Theorem, Version 1) *Let X be a topological space with $X = A \cup B$, where A and B are open sets, and $A \cap B$ is a nonempty set that is path-connected and simply connected (i.e. $\pi_1(A \cap B)$ is trivial). Let $x \in A \cap B$ be a basepoint. Then*

$$\pi_1(X, x) = \pi_1(A, x) * \pi_1(B, x).$$

Proof The idea is to follow the strategy from Theorem 12.1—but now redo everything with more generality. So, we construct a homomorphism $\phi : \pi_1(A, x) * \pi_1(B, x) \to \pi_1(X, x)$ and show that it is injective and surjective.

An element of $\pi_1(A, x) * \pi_1(B, x)$ is a word of the form

$$w = a_1 b_1 a_2 b_2 \cdots a_m b_m$$

(or possibly starting with a b or ending with an a), where each $a_i \in \pi_1(A, x)$, each $b_i \in \pi_1(B, x)$, and no a_i or b_i is equal to the identity in $\pi_1(A, x)$ or $\pi_1(B, x)$. If w is the identity element, then we consider it to be a product of length 0, consisting of no a's or b's. Thus w corresponds to a loop obtained by doing a_1 in A, then doing b_1 in B, and so forth.

Let us begin by showing surjectivity. To do this, we must show that any element of $\pi_1(X, x)$ is represented by a loop of the above form. Let γ be a loop in X based at x. Partition $[0, 1]$ into $s_0 = 0 < s_1 < s_2 < \cdots < s_n = 1$, so that for each i, $\gamma(t)$ restricted to $[s_i, s_{i+1}]$ is entirely contained either in A or in B—and they alternate. Then, for each i, $\gamma(t)$ restricted to $[s_i, s_{i+1}]$ is homotopic either to an element of $\pi_1(A, x)$ or to an element of $\pi_1(B, x)$. Suppose $\gamma(t)$ restricted to $[s_i, s_{i+1}]$ is homotopic to c_i, where either $c_i \in \pi_1(A, x)$ or $c_i \in \pi_1(B, x)$. Then γ is homotopic to the concatenation $c_0 c_1 \cdots c_{n-1}$. Hence we have written the homotopy class of γ as an element of $\pi_1(A, x) * \pi_1(B, x)$, so ϕ is surjective.

Now we must show that ϕ is injective. Suppose that $w \in \pi_1(A, x) * \pi_1(B, x)$ is a word, with $\phi(w) = e$, the identity in $\pi_1(X, x)$. We must show that w is the empty word. Let γ be the loop corresponding to w. If $\phi(w) = e$, then there is a homotopy $F : [0, 1] \times [0, 1] \to X$ that deforms γ to the constant loop at x. We cut up both the s-interval and the t-interval into sufficiently small pieces that, on any rectangle $[s_j, s_{j+1}] \times [t_i, t_{i+1}]$, the image of F lies either entirely in A or entirely in B. If two adjacent rectangles, either in the horizontal direction or in the vertical direction, both lie in A or both lie in B, then join them together. Thus we have broken up the square $[0, 1] \times [0, 1]$ into a bunch of regions bounded by horizontal and vertical line segments. Let us call the regions we have $R_1, \ldots, R_m$.

The image of the boundary of any R_i must lie in $A \cap B$ and is thus a closed loop in $A \cap B$. Since $A \cap B$ is path-connected and simply connected, the boundary of each R_i maps to a loop in X that is homotopic to the identity. So, we can modify the map slightly so as to deform the image of the boundary of each R_i to x. But this isn't *quite* right, because we cannot modify the image of the $s = 0$ part of the square, because that part *has* to be $\gamma(t)$. So we can't modify the $s = 0$ boundary of the square. But the rest of the boundaries of the R_i's can move.

Now each R_j gives a homotopy between $\gamma(t)$ restricted to $[t_i, t_{i+1}]$ and the identity. Hence each little piece of γ is homotopic to the relevant identity, so w must have been the empty word, and so ϕ is injective. ■

12.3 More Fundamental Groups

Although we have not given the most general form of the Seifert–Van Kampen Theorem, we can already use it to compute fundamental groups for quite a lot of spaces.

Example Let X and Y be two topological spaces, and let $x \in X$ and $y \in Y$ be points. We define their *wedge sum* $X \vee Y$ as follows: let $X \sqcup Y$ be the disjoint union of X and Y. Define an equivalence relation $\sim$ on $X \sqcup Y$ by declaring that $x \sim y$, but all other points are only equivalent to themselves. Then we let $X \vee Y$ be $(X \sqcup Y)/\sim$ be the quotient space. (The wedge sum of two circles $\mathbb{S}^1 \vee \mathbb{S}^1$ above is a special case of this construction.) Intuitively, this means that we glue X and Y together at one point. More generally, if $\{X_\alpha\}_{\alpha \in A}$ is any (not necessarily finite) collection of topological spaces, and we declare basepoints $x_\alpha \in X_\alpha$ for each $\alpha \in A$, then we define their wedge sum $\bigvee_{\alpha \in A} X_\alpha$ to be the quotient space of $\bigsqcup X_\alpha$, modulo the equivalence relation that sets $x_\alpha \sim x_\beta$ for all $\alpha, \beta \in A$, but all other points are only equivalent to themselves.

We also need to know the *topology* on the wedge sum, i.e. the open sets. The topology we put on the wedge sum is the quotient topology as $\bigsqcup X_\alpha/\sim$. Assuming that the X_α's are nice (in particular, Hausdorff; see Appendix A for a definition), this means the following: Around any point $x \in X_\alpha$ other than the basepoint, a small neighborhood around x in $\bigvee_{\alpha \in A} X_\alpha$ is just a neighborhood in X_α not containing

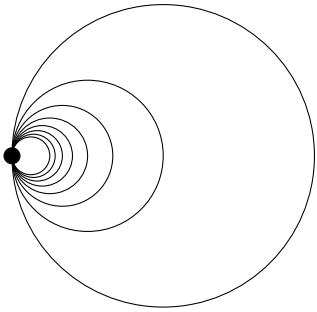


Figure 12.5 The Hawaiian earring. This is the union of circles of radius $1/n$, all meeting at a single point. There is no neighborhood of this common point that is contractible, because any neighborhood of this point must contain infinitely many circles. Its fundamental group is a very complicated object!

the basepoint x_α, considered as a subset of the wedge sum. On the other hand, a neighborhood of the basepoint in the wedge sum is the union of neighborhoods around the basepoint in each X_α.

Let X and Y be two spaces, and let x be the point connecting X and Y in $X \vee Y$. Suppose that, in both X and Y, x has a *contractible* open neighborhood. (See the Nonexample below for a situation in which this fails to happen.) Let A be an open set containing X inside $X \vee Y$, which is homotopy equivalent to X, and similarly let B be an open set containing Y inside $X \vee Y$, which is homotopy equivalent to Y. Then $\pi_1(A, x) \cong \pi_1(X, x)$ and $\pi_1(B, x) \cong \pi_1(Y, x)$, so $\pi_1(X \vee Y, x) \cong \pi_1(A, x) * \pi_1(B, x)$.

Nonexample It is worth being careful about wedge sums. The *Hawaiian earring* (see Figure 12.5) is a classic example of a space that looks like a wedge sum (of infinitely many circles). However, it isn't a wedge sum, because any open neighborhood of the basepoint must contain all but finitely many circles, whereas a neighborhood in the infinite wedge sum of circles can contain only an arc of each circle. You can find a description of the fundamental group of the Hawaiian earring in [dS92].

Unfortunately, there are still many spaces whose fundamental groups we cannot determine with this version of the Seifert–Van Kampen Theorem. We will need to use a more powerful version of the theorem in order to determine their fundamental groups. This is what we will now discuss.

12.4 The Seifert–Van Kampen Theorem: Second Version

In the previous sections, we presented a special case of the Seifert–Van Kampen Theorem in order to be able to compute fundamental groups of more complicated spaces when we are able to break these spaces down into simple "building blocks" whose fundamental groups we already understand. However, this special case is not yet sufficiently powerful to allow us compute the fundamental group of a very important topological space: the identification space of a compact surface without

boundary. For this we will need a more general version of the Seifert–Van Kampen Theorem; once we have stated and proved this theorem, we will be able to apply it to the case of identification spaces. We will therefore be able to compute $\pi_1(S)$, where S is any compact surface!

The generalization of the Seifert–Van Kampen Theorem that we will present addresses the case where $\pi_1(A \cap B)$ is non-trivial, where A and B are the building blocks whose union is the topological space of interest. The following theorem gives the result. But note that this is still not the most general version of the Seifert–Van Kampen Theorem!

Theorem 12.3 (Seifert–Van Kampen Theorem, Version 2) *Let X be a topological space with $X = A \cup B$, where A and B are open sets, and $A \cap B$ is nonempty and path-connected. Assume further that B is simply connected (i.e. $\pi_1(B)$ is trivial). Then*

$$\pi_1(X) \cong \pi_1(A)/N,$$

where N is the smallest normal subgroup containing the image of $\pi_1(A \cap B)$ under the homomorphism induced by the inclusion mapping $\iota : A \cap B \to A$.

We'll sketch the proof of this theorem at the end of this chapter, after presenting examples of how this theorem can be applied to compute fundamental groups.

12.5 The Fundamental Group of a Compact Surface

We'll start with two examples that show how the second version of the Seifert–Van Kampen Theorem can be used to compute the fundamental group of a compact surface presented as an identification space.

Example Let $\mathbb{T}$ be the torus, presented as the ID space $aba^{-1}b^{-1}$ obtained by identifying the sides of a rectangle in the usual way. Let B be contained in the interior of the rectangle, consisting of most of the rectangle not quite all the way to its boundary. Let A be the remaining part of the rectangle—extended a bit into the interior of B. Then the intersection $A \cap B$ is a "ribbon" that runs parallel to the boundary of the rectangle. See Figure 12.6.

Now B is homotopic to an open disk, which is contractible and thus has trivial fundamental group. Also $A \cap B$ is an annulus, which deformation retracts onto the circle and has fundamental group $\pi_1(A \cap B) \cong \mathbb{Z}$. What about A? By folding the rectangle up into a cylinder by gluing together the a-edge, we can see that A is homotopic to a thickened "figure eight," which deformation retracts onto the wedge of two circles and has $\pi_1(A) \cong \mathbb{Z} * \mathbb{Z}$. Concretely, we can say that $\pi_1(A)$ is the free group $F([a], [b])$.

In order to apply the Seifert–Van Kampen Theorem, we must know how $\pi_1(A \cap B)$ injects into $\pi_1(A)$ under the homomorphism induced by the inclusion map $\iota : A \cap B \to A$. To this end, observe that the curve γ pictured in Figure 12.6, whose

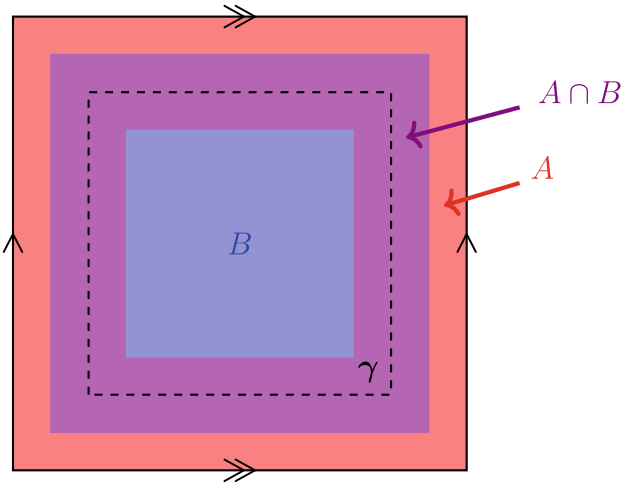


Figure 12.6 The decomposition of $\mathbb{T}$ into a pair of overlapping open sets A and B.

equivalence class generates $\pi_1(A \cap B)$, is a curve that winds once around the rectangle by following curve segments that are almost—but not quite equal to—the edges of the rectangle. In fact, we can say that $\gamma \sim a * b * \bar{a} * \bar{b}$, where we recall that $*$ is curve concatenation and the bar denotes reversed orientation. Therefore, $\iota_*[\gamma] = [a][b][a]^{-1}[b]^{-1}$.

The normal subgroup N of the Seifert–Van Kampen Theorem is therefore the subgroup of $F([a], [b])$ that contains $[a][b][a]^{-1}[b]^{-1}$ along with all elements generated by all conjugates of $[a][b][a]^{-1}[b]^{-1}$. Hence, in the quotient group, we'll have $[a][b][a]^{-1}[b]^{-1} = [e]$ or else $[a][b] = [b][a]$. The quotient group is abelian! In fact, we'll find

$$\begin{aligned}\pi_1(\mathbb{T}) &\cong F([a], [b])/N \\ &\cong \langle [a], [b] \mid [a][b] = [b][a] \rangle \\ &\cong \mathbb{Z} \times \mathbb{Z}.\end{aligned}$$

This is precisely the result we have already obtained.

To see that $G_1 := \langle [a], [b] \mid [a][b] = [b][a] \rangle$ is isomorphic to $G_2 := F([a], [b])/N$, we note that in G_1 we have $[a][b] = [b][a]$, so for any $g \in G_1$ we also have $[a][b][a]^{-1}[b]^{-1} = e$, so $g[a][b][a]^{-1}[b]^{-1}g^{-1} = e$. Hence any conjugate of $[a][b][a]^{-1}[b]^{-1}$ is in the kernel of the map $G_2 \to G_1$. Now, since G_2 is a quotient of $F([a], [b])$ having only the elements of N as relations, it must be the largest possible group determined by these generators and relations. Hence $G_1 = G_2$. Note that this is a little bit sketchy, but essentially correct. A more rigorous way of showing that $G_1 = G_2$ would be to define G_1 in terms of the properties (called universal

properties) it has with respect to other groups and then show that G_2 satisfies these properties. But we wish to avoid getting into that here.

Example Let $\mathbb{K}$ be the Klein bottle, presented as the ID space $ab^{-1}a^{-1}b^{-1}$ obtained by identifying the sides of a rectangle in the usual way. Similar arguments as in the torus example lead to $\pi_1(\mathbb{K}) \cong \langle [a], [b] \mid [a][b]^{-1} = [b][a] \rangle$. This group is non-abelian, because $[a]$ and $[b]$ do not commute with each other: $[b][a]$ is not equal to $[a][b]$, but rather to $[a][b]^{-1}$. To give a complete proof that $\pi_1(\mathbb{K})$ is not abelian, it is necessary to show that $[a][b]^{-1} \neq [a][b]$ in the group with the presentation $\langle [a], [b] \mid [a][b]^{-1} = [b][a] \rangle$. We save this task for Problem 6.

Let us now see if we can use arguments as in the torus example to compute the fundamental group of an arbitrary compact surface, presented as an ID space of the form given in the classification theorem of compact surfaces, which we proved in Chapter 4. Recall from the proof given there that we can present the ID space of a genus g surface as a polygon with $4g$ sides written in "torus order" as $a_1 b_1 a_1^{-1} b_1^{-1} \cdots a_g b_g a_g^{-1} b_g^{-1}$.

Theorem 12.4 *Let S_g be an orientable surface of genus g presented as the identification space $a_1 b_1 a_1^{-1} b_1^{-1} \ldots a_g b_g a_g^{-1} b_g^{-1}$. Then*

$$\pi_1(S_g) \cong \langle a_1, b_1, \cdots, a_g, b_g \mid a_1 b_1 a_1^{-1} b_1^{-1} \cdots a_g b_g a_g^{-1} b_g^{-1} = 1 \rangle .$$

Proof Let P be a regular polygon with $4g$ sides, which becomes homeomorphic to S_g when its edges are identified. Let B be the interior of P. Hence B is a simply connected and path-connected open set. Let A be an open neighborhood of the boundary of P. Consequently, $A \cap B$ is an open "ribbon" in the interior of P that deformation retracts onto a circle γ satisfying $\gamma = a_1 * b_1 * \bar{a}_1 * \bar{b}_1 * \cdots * a_g * b_g * \bar{a}_g * \bar{b}_g$.

After the identifications are made, and we view A and B as subsets of S_g, then B remains homeomorphic to a disk, while A is homeomorphic to a wedge of $2g$ circles equal to $a_1, b_1, \ldots, a_g, b_g$. Consequently, $\pi_1(A) \cong \mathbb{Z} * \cdots * \mathbb{Z} \cong F([a_1], [b_1], \ldots, [a_g], [b_g])$, which is the free group on $2g$ generators. Moreover, if $\iota : A \cap B \to S_g$ is the inclusion mapping, then $\iota_*[\gamma] = [a_1][b_1][a_1]^{-1}[b_1]^{-1} \cdots [a_g][b_g][a_g]^{-1}[b_g]^{-1}$. Therefore, the Seifert–Van Kampen Theorem implies that $\pi_1(S_g) \cong \pi_1(A)/N$, where N is the smallest normal subgroup of $F([a_1], [b_1], \ldots, [a_g], [b_g])$ containing the element $[a_1][b_1][a_1]^{-1}[b_1]^{-1} \cdots [a_g][b_g][a_g]^{-1}[b_g]^{-1}$. The result of this quotienting is the group given in the statement of the theorem. ■

12.6 Even More Fundamental Groups

We can also apply the technique used in these examples to topological spaces that are not compact surfaces. For instance, consider the space X obtained by identifying the three edges of an equilateral triangle in consecutive order; i.e., if we label each edge by a, then we're talking about the space aaa. This is not quite the same as the

dunce cap that we met earlier: that space is the ID space aaa^{-1}. By analogy with the dunce cap, we'll call this space X the *dance cup*.

Exercise 12.5 Why is the dance cup not a surface? Also, show that it contains an orientation-reversing path.

Example Similar arguments as in the torus example lead to $\pi_1(X) \cong \langle [a] \mid [a]^3 = [e] \rangle = \mathbb{Z}/3\mathbb{Z}$.

12.7 Proof of the Second Version of the Seifert–Van Kampen Theorem

Proof We'll show how to construct an isomorphism $\phi : \pi_1(X) \to \pi_1(A)/N$.

To begin, let γ be any loop in X based at $x_0 \in A \cap B$. (Recall that the location of the basepoint can be chosen at will.) We first claim that γ is homotopic to a loop γ' based at x_0 and contained entirely in A. The reason is as follows. We can subdivide γ into sub-curves $\gamma_1, \gamma_2, \ldots, \gamma_r$, where each γ_i is either entirely contained in A or entirely contained in B, and has its endpoints in $A \cap B$. For any sub-curve $\gamma_i \subseteq B$, let σ_i be a curve in $A \cap B$ connecting the endpoints of γ_i. Then $\gamma_i * \sigma_i$ is a homotopically trivial loop in B, because $\pi_1(B) = \{[e]\}$. Therefore we can homotope γ_i into σ_i. (Exercise: verify this!) If we apply this operation to all parts of γ entirely contained in B and concatenate the results, we get the curve γ'.

The analysis above shows that $[\gamma] = [\gamma']$, and because $\gamma' \subseteq A$ we can say $[\gamma'] \in \pi_1(A)$. Now we can define $\phi([\gamma]) := \text{ proj } ([\gamma'])$, where $\text{ proj} : \pi_1(A) \to \pi_1(A)/N$ is the natural projection homomorphism that the quotient group construction of $\pi_1(A)/N$ gives us.

We must now ask the usual set of questions about ϕ. First, is it well-defined? To answer this question, suppose that we had started with $\gamma_1 \sim \gamma$. Then, with a bit of attention to detail, we can show that $\gamma_1' \sim \gamma'$ where the "prime" denotes the operation of deforming a curve in $A \cup B$ to one in A alone, and so $\phi([\gamma_1]) = \phi([\gamma])$.

Next, is ϕ a homomorphism? To answer this question, we can write down what is required for $\phi([\gamma_1][\gamma_2]) = \phi([\gamma_1])\phi([\gamma_2])$—and it turns out that the critical step is to show that $(\gamma_1 * \gamma_2)' \sim \gamma_1' * \gamma_2'$. Again, by being careful with the construction of the "prime" operation, we can show this without too much difficulty.

Next, is ϕ surjective? To answer this question, suppose $x \in \pi_1(A)/N$ is an arbitrary element of the quotient group. Since proj is surjective, we can write $x = \text{ proj}([\gamma])$ for some $[\gamma] \in \pi_1(A)$. Clearly $[\gamma]$ can also be viewed as a class in $\pi_1(X)$ so we can say $\phi([\gamma]) = x$.

Finally, is ϕ injective? To answer this question, suppose that $\phi([\gamma]) = \text{ proj}([\gamma'])$ $= \text{id}$. Then $[\gamma'] \in N$ or in other words, there exists a class $[\sigma] \in \pi_1(A)$ and a class $[\gamma_0] \in \pi_1(A \cap B)$ so that $[\gamma'] = [\sigma][\gamma_0][\sigma]^{-1}$. In other words, at the level of loops, we have $\gamma' \sim \sigma * \gamma_0 * \bar{\sigma}$. Now, because $\gamma_0 \subseteq A \cap B \subseteq B$ and B is homotopically trivial, we can homotope γ_0 to the trivial path. Hence $\sigma * \gamma_0 * \bar{\sigma} \sim \sigma * e * \bar{\sigma} \sim e$.

Therefore γ' is homotopically trivial. This in turn implies that γ is homotopically trivial, because γ' and γ differ only by curve segments in B. Therefore $[\gamma] = [e]$, showing that ϕ is indeed injective. ■

12.8 General Seifert–Van Kampen Theorem

So far we have worked out several cases of the Seifert–Van Kampen Theorem. For a complete picture, we now present the general statement, although we will not give a proof.

Theorem 12.6 (Seifert–Van Kampen Theorem) *Let X be a path-connected space, and suppose that $X = A \cup B$, where A and B are open sets with $A \cap B$ path-connected. Let $\iota_1 : A \cap B \to A$ and $\iota_2 : A \cap B \to B$ be the inclusion mappings and $(\iota_1)_* : \pi_1(A \cap B) \to \pi_1(A)$ and $(\iota_2)_* : \pi_1(A \cap B) \to \pi_1(B)$ the induced homomorphisms. Let $N \leq \pi_1(A) * \pi_1(B)$ be the normal subgroup generated by elements of the form $(\iota_1)_*(h)(\iota_2)_*(h)^{-1}$ for $h \in \pi_1(A \cap B)$. Then $\pi_1(X) = \pi_1(A) * \pi_1(B)/N$.*

Remark 12.7 We call the quotient of the free product that appears in the general Seifert–Van Kampen Theorem an *amalgamated free product*. The amalgamated free product is a generalization of the free product: Rather than joining together two groups that have nothing to do with each other and forming strings of symbols that alternate between the two groups, the amalgamated free product identifies certain marked subgroups of each of the two groups and then makes a group out of the two groups, one that is as free of relations as it can be, subject to the constraint that it has glued together the marked subgroup of each. See [Ser03] for a book on amalgamated free products and their appearances in topology and number theory.

12.9 Groups as Fundamental Groups

Whenever we can construct a group associated to some object that occurs elsewhere in mathematics, it is tempting to ask which groups can arise from such a construction. In the case of fundamental groups, we can answer this question fully.

Theorem 12.8 *Every group occurs as the fundamental group of some topological space.*

Sketch of Proof Let G be a group. We will construct a space whose fundamental group is isomorphic to G. First, we find a presentation $G = \langle S \mid R \rangle$ for G. This can always be done, because at worst we can let S consist of all the elements of G, and we can let R consist of all relations that arise in G. (In practice, we will be able to come up with far more efficient presentations, but that won't be necessary for this proof.) Now, construct a wedge of circles, one for each element of S, and label the

circles by elements of R. Next, for each relation $r \in R$, say $r = s_1 \dots s_n$, add a disk whose boundary is the concatenation $s_1 \dots s_n$. This space X has $\pi_1(X) \cong G$. The proof is very similar to the calculation we did above with the fundamental groups of surfaces from their ID-space representations. ■

Example What would a space X with $\pi_1(X) \cong \mathbb{Q}$ look like? There are many spaces with $\pi_1(X) \cong \mathbb{Q}$, but the classic example is called a rational telescope. To construct it, start with a finite cylinder $\mathbb{S}^1 \times [0, 1]$. The fundamental group of this cylinder is $\mathbb{Z}$, generated by the loop α that goes around it once in the counterclockwise direction. Now, at the top of this cylinder, glue another cylinder, but arrange it so that the top boundary of the bottom cylinder is identified with the loop that goes around the bottom loop of the top cylinder *twice*. The fundamental group of the resulting figure is still $\mathbb{Z}$, but it is no longer generated by α, because for a loop β around the top cylinder, we have $[\beta * \beta] = [\alpha]$, so the fundamental group of the resulting figure is generated by β. Then, take a third cylinder, and glue the bottom of this cylinder to the top of the second cylinder, arranging it so that the top boundary of the third cylinder goes around the bottom circle of the second cylinder *three* times. The resulting figure still has fundamental group $\mathbb{Z}$, but again it is a different $\mathbb{Z}$, this time generated by a

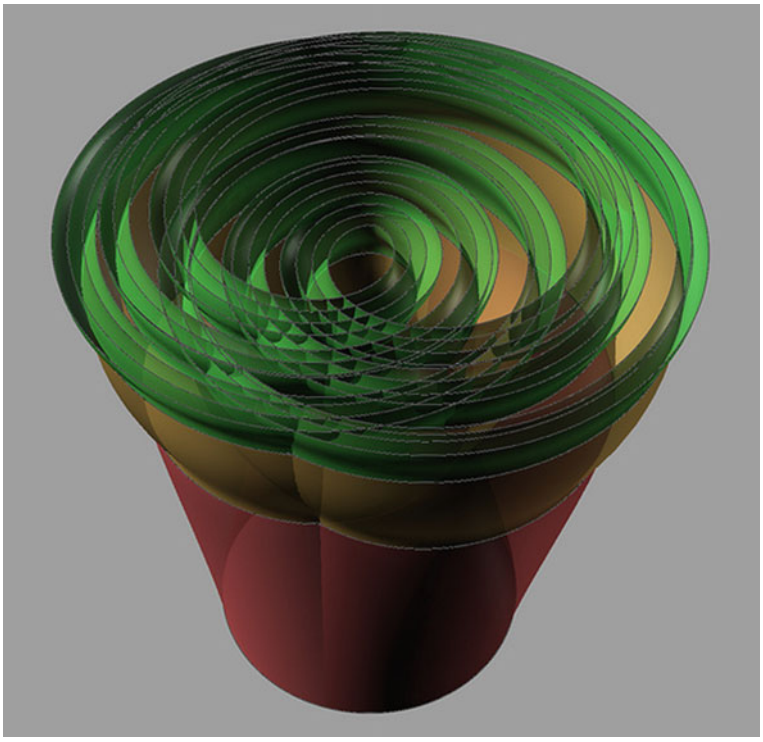


Figure 12.7 A picture of a rational telescope.

loop γ around the third cylinder, so that $[\gamma]^3 = [\beta]$. Keep doing this infinitely many times, identifying the bottom of the n^{th} cylinder with the top of the $(n-1)^{\text{st}}$ cylinder, going around n times. The final result is a space whose fundamental group is $\mathbb{Q}$. (See Figure 12.7 for an idea of what the rational telescope looks like.) This is an example of a very important class of constructions in homotopy theory. See [Bak09], which is where we got the picture, and [Bae09] for more information about the rational telescope and related spaces, and why they are important in homotopy theory.

12.10 Problems

(1) Let X be a wedge of two projective planes. Give a presentation for $\pi_1(X)$.
(2) Explain carefully why the Hawaiian earring is not homeomorphic to a wedge of countably many circles. (Remark: *Countable* means in bijection with the positive integers. Hence, a wedge of countably many circles means countably many circles of radius 1, all glued together at a single point. In terms of quotient topologies, the wedge of countably many circles consists of (x, n), where $x \in \mathbb{S}^1$ and $n \in \mathbb{N}$, modulo the relation $((1,0), m) \sim ((1,0), n)$ for all m, n; that is, we glue all the $(1, 0)$ points on the circles together.)
(3) Let X be the Hawaiian earring, and let Y be a wedge of a countable number of circles. Is there a surjective continuous function $f : X \to Y$? What about $Y \to X$?
(4) (a) Let X_3 be the complement in $\mathbb{R}^3$ of the three coordinate axes. Find a simpler space that is homotopy equivalent to X_3, and compute $\pi_1(X_3)$.
    (b) In the same vein, let X_n be the complement in $\mathbb{R}^3$ of n lines which pass through the origin. Compute $\pi_1(X_n)$.
(5) Use the Seifert–Van Kampen Theorem to calculate the fundamental group of $\mathbb{RP}^2$.
(6) We saw in the text that $\pi_1(\mathbb{K})$ is nonabelian.
    a. Find two curves in an ID-space representation of $\mathbb{K}$ whose classes don't (appear to) commute.
    b. Prove that $\pi_1(\mathbb{K})$ is nonabelian. One possible method is by finding a quotient of $\pi_1(\mathbb{K})$ that you already know to be nonabelian, because all quotients of abelian groups are abelian.
(7) The Klein bottle is either the ID-space $aba^{-1}b$ or else it is the ID-space $aabb$. This comes from the fact that $\mathbb{K}$ is homeomorphic to $\mathbb{RP}^2 \# \mathbb{RP}^2$. You can thus compute $\pi_1(\mathbb{K})$ in two ways and you get seemingly different answers. What's going on?
(8) Use the Seifert–Van Kampen Theorem to calculate the fundamental group of the nonorientable surface of genus g (i.e. the one obtained by taking the connected sum of the orientable surface of genus g with the projective plane).

(9) Let W be the space obtained by taking the union of the sphere $\mathbb{S}^2$ and a straight line (not a great circle!) that connects the north and south poles. Can you calculate $\pi_1(W)$ using the Seifert–Van Kampen Theorem? If not, can you deform W to a space where you can apply this theorem?

Chapter 13
Introduction to Homology

13.1 The Idea of Homology

We have already seen one of the key algebraic invariants for topological spaces: the fundamental group. Roughly, the fundamental group detects interesting maps from the circle $\mathbb{S}^1$ to a space X. There are higher-dimensional versions of the fundamental group, known as homotopy groups and denoted by $\pi_n(X)$; these are defined in terms of homotopy classes of maps from $\mathbb{S}^n$ to X. In computing $\pi_1(\mathbb{S}^1) \cong \mathbb{Z}$, we already found that we needed a somewhat involved argument. Nonetheless, as we learned when studying the Seifert–Van Kampen Theorem, there is a general method for computing fundamental groups of nice spaces.

No similar methods exist for the higher homotopy groups, and even the computation of higher homotopy groups of (higher-dimensional) spheres is a major topic of current research. For example, who would imagine that there are interesting ways of mapping $\mathbb{S}^3$ or $\mathbb{S}^4$ to $\mathbb{S}^2$? Or that there are exactly 2880 homotopy classes of maps from $\mathbb{S}^{14}$ to $\mathbb{S}^4$? See [Hat02, Chapter 4] for much more on this.

So, while higher homotopy groups are a nice idea, and they are very important in advanced algebraic topology, they are very tricky to compute in practice—one must rely on some technical machinery, especially spectral sequences, in order to compute them. (See [BT82] for several examples of these computations.) Instead, we must return to the land of triangulations and dig more deeply, in order to define a family of topological invariants known as the *homology groups*.

The basic idea of homology is to count n-dimensional holes. Roughly speaking, an n-dimensional hole H in X is a compact n-dimensional manifold—an n-dimensional analogue of a surface—without boundary in X. But a hole is trivial if it is *filled in*; that is, if there is a compact $(n+1)$-dimensional manifold *with boundary* in X, whose boundary is H. (This isn't quite correct, but it will suffice for the sake of intuition for now.) For example, $\mathbb{S}^2$ has no nontrivial 1-dimensional holes, because for every 1-dimensional manifold (namely, a circle) C inside $\mathbb{S}^2$, we can find a disk whose boundary is C. On the other hand, the sphere *does* have a 2-dimensional hole, because the sphere itself is not the boundary of a 3-dimensional manifold with

C. Bray et al., *Algebraic Topology*,
https://doi.org/10.1007/978-3-030-70608-1_13

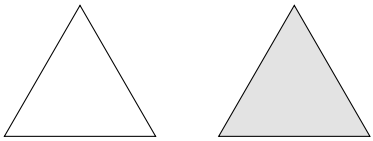


Figure 13.1 Left: an empty triangle has a nontrivial hole, whereas when we fill it in (Right), the hole becomes trivial.

boundary contained inside $\mathbb{S}^2$. (In fact, there are no 3-dimensional manifolds at all contained inside $\mathbb{S}^2$.) The homology groups measure holes "up to triviality," in a way that we will soon make precise.

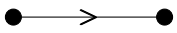
To give the simplest example of a trivial hole and a nontrivial hole, let us consider the triangle on the left in Figure 13.1, by which we mean just the edges without the interior. This triangle is a 1-dimensional hole, because we can map a circle (a 1-manifold) to it. This same hole also exists in the filled triangle on the right in Figure 13.1, except this time the hole is trivial: it's the boundary of the filled triangle. Once we learn the definition of homology and compute it, we'll find that the triangle on the left has nontrivial 1-dimensional homology, whereas the filled triangle on the right has trivial 1-dimensional homology.

Now, if we take the filled triangle on the right and glue a second filled triangle to the boundary so as to make a triangular pillow, we would create a nontrivial 2-dimensional hole formed by the two triangles: we can map an $\mathbb{S}^2$ to our space by squashing the sphere onto the two triangles. Furthermore, this hole is nontrivial, because it isn't the boundary of a 3-manifold with boundary. Thus the 2-dimensional homology of the triangular pillow is nontrivial.

13.2 Chains

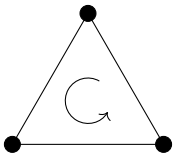
Our basic building blocks in homology—or, to be more precise, simplicial homology or Δ homology—are *simplices*.[1] Simplices are vertices, edges, faces, and their higher-dimensional analogues that we saw while working with triangulations when we discussed the Euler characteristic. A vertex is a 0-simplex, an edge is a 1-simplex, a face is a 2-simplex, and so forth; in general, an n-simplex has $n + 1$ vertices.

There is a small difference between the vertices, edges, and faces we saw before, and the simplices we are using now: in homology, it is necessary to work with *oriented simplices*. In low dimensions, we can easily visualize oriented simplices. 0-dimensional simplices, or vertices, do not need to be oriented. In the case of a 1-dimensional simplex, or edge, we can travel along the edge in either direction; we put an arrow from the start vertex to the end vertex:



[1] The singular of *simplices* is *simplex*.

In the case of a 2-dimensional simplex, we imagine traveling around the boundary of the simplex. We can travel in the clockwise direction, or in the counterclockwise direction. We draw a circular arrow to indicate our direction:



It is harder to visualize the orientation on higher-dimensional simplices in an analogous way, so instead we define the orientation more formally. Let us suppose that the vertices bounding an n-dimensional simplex S are $v_0, \ldots, v_n$. (Note that it has $n+1$ vertices, as expected.) We write $S = [v_0, \ldots, v_n]$. We can order the vertices in any of $(n+1)! = |S_{n+1}|$ ways. The orientation is the *sign* of the permutation described by the ordering. (Recall that the sign of a permutation $\pi \in S_{n+1}$ is $(-1)^{t(\pi)}$, where $t(\pi)$ is the number of transpositions in a factorization of π; $t(\pi)$ is not well-defined, but it *is* well-defined modulo 2: for a fixed permutation, there will either always be an even number of transpositions in a factorization of π, or else always an odd number.) Thus there are only two orientations for each simplex: positive and negative.

Thus, as the pictures suggest, the edge $[v_0, v_1]$ has the opposite orientation from $[v_1, v_0]$; the faces $[v_0, v_1, v_2]$ and $[v_2, v_0, v_1]$ have the same orientation, whereas $[v_0, v_2, v_1]$ has the opposite orientation.

If two orderings of the vertices of some n-simplex have the same sign, then we consider them to be the same; if they have opposite sign, then we consider them to be off by a factor of -1.

Now, suppose we have a triangulation T of a space X. In the future, we shall implicitly assume that spaces come with triangulations into simplices, without mentioning it explicitly. Pick, once and for all, an orientation on each simplex. An *n-chain* on X is a formal integer linear combination of the (oriented) n-simplices of X.

This concept is worthy of an example. Suppose the oriented n-simplices in a triangulation T of X are $T_1, \ldots, T_r$. Then a typical example of an n-chain might be $3T_1 - 5T_2 + 0T_3 + \cdots - 6T_r$. If there are *infinitely many* n-simplices, then we are only allowed to use finitely many of them in the sum; all the rest must be "multiplied by 0." All the coefficients must be integers, and they are allowed to be positive, negative, or zero.

Adding and multiplying n-simplices does not have any geometric meaning: after all, how might we interpret 3 times some simplex minus 5 times another? Rather, we think of an n-chain as being something *formal*: an algebraic object rather than a geometric one.

The n-chains of a space X form a group, denoted $C_n(X)$ (no relation to the cyclic group C_n), under addition: just add the coefficients of each of the simplices. For example, suppose the n-simplices of X are $T_1, \ldots, T_r$, and we have two n-chains $a_1T_1 + \cdots + a_rT_r$ and $b_1T_1 + \cdots + b_rT_r$. Then their sum is $(a_1 + b_1)T_1 + \cdots +$

$(a_r + b_r)T_r$. Since addition is clearly commutative, $C_n(X)$ is an *abelian* group. In fact, it is isomorphic to the *free abelian group* $\mathbb{Z}^r$.

Remark 13.1 The number of n-simplices need not be finite in general, but it will be in all our examples.

13.3 The Boundary Map

Recall that the idea of homology is to count holes that are not boundaries. Thus, we need to be able to detect when something is or is not a boundary. We start, naturally enough, by defining the boundary of a simplex.

Suppose that we have an n-simplex $[v_0, \ldots, v_n]$. We define its *boundary* to be

$$\partial_n([v_0, \ldots, v_n]) = \sum_{i=0}^{n} (-1)^i [v_0, \ldots, \widehat{v_i}, \ldots, v_n],$$

where the hatted term $\widehat{v_i}$ is omitted. For example, if $n = 2$, we have

$$\partial_2([v_0, v_1, v_2]) = [v_1, v_2] - [v_0, v_2] + [v_0, v_1].$$

Observe that the boundary of an n-simplex is an $(n-1)$-chain.

We now extend the boundary map to a homomorphism $\partial_n : C_n(X) \to C_{n-1}(X)$ in the only way possible:

$$\partial_n(a_1 T_1 + \cdots + a_r T_r) = a_1 \partial_n(T_1) + \cdots + a_r \partial_n(T_r).$$

The most important property of the boundary map is that *the boundary of a boundary is zero*, i.e. the following theorem.

Theorem 13.2 *If $A \in C_n(X)$ is any n-chain, then $\partial_{n-1} \circ \partial_n(A) = 0$.*

Proof We simply compute. It suffices to check this in the case that $A = [v_0, \ldots, v_n]$ is an n-simplex because the boundary maps are homomorphisms. Then we have

$$\begin{aligned}
\partial_{n-1} \circ \partial_n(A) &= \partial_{n-1} \left(\sum_{i=0}^{n} (-1)^i [v_0, \ldots, \widehat{v_i}, \ldots, v_n] \right) \\
&= \sum_{i=0}^{n} (-1)^i \partial_{n-1}([v_0, \ldots, \widehat{v_i}, \ldots, v_n]) \\
&= \sum_{i=0}^{n} (-1)^i \left(\sum_{j=0}^{i-1} (-1)^j [v_0, \ldots, \widehat{v_j}, \ldots, \widehat{v_i}, \ldots, v_n] \right. \\
&\qquad \left. + \sum_{j=i+1}^{n} (-1)^{j-1} [v_0, \ldots, \widehat{v_i}, \ldots, \widehat{v_j}, \ldots, v_n] \right).
\end{aligned}$$

In the final expression on the right, we see that there are exactly two terms missing both v_i and v_j. But what are the signs? Suppose that $i < j$. Then it occurs once with coefficient $(-1)^i(-1)^{j-1} = (-1)^{i+j-1}$ by removing first v_i and then v_j, and once with coefficient $(-1)^j(-1)^i = (-1)^{i+j}$ by first removing v_j and then removing v_i. The sum of these two coefficients is 0, so the coefficient of the $(n-2)$-simplex $[v_0, \ldots, \widehat{v_i}, \ldots, \widehat{v_j}, \ldots, v_n]$ is 0. This is true for each $(n-2)$-simplex, so $\partial_{n-1} \circ \partial_n(A) = 0$, as desired. ■

One way of expressing this theorem is to say that $\mathrm{im}(\partial_n) \leq \ker(\partial_{n-1})$, because $\ker(\partial_{n-1})$ consists of the $(n-1)$-chains whose boundaries are 0, and $\mathrm{im}(\partial_n)$ consists of the $(n-1)$-chains that are images of n-chains under the boundary map.

Definition 13.3 We call $\ker(\partial_n) \leq C_n(X)$ the group of *n-cycles* and denote it by $Z_n(X)$, and we call $\mathrm{im}(\partial_{n+1}) \leq C_n(X)$ the group of *n-boundaries* and denote it by $B_n(X)$.

Remark 13.4 Recall that we motivated homology by describing it as measuring n-dimensional holes, up to triviality. The cycles $Z_n(X)$ are the holes. Note that a hole, such as a loop, has trivial boundary—which is exactly what defines a cycle. The boundaries $B_n(X)$ are the trivial holes, because if $c \in B_n(X)$ is the boundary of some $a \in C_{n+1}(X)$, then the hole c is filled in by a.

The sequence of groups and maps we have here is very important—important enough to deserve its own name.

Definition 13.5 A sequence

$$\cdots \to C_{n+1} \xrightarrow{\partial_{n+1}} C_n \xrightarrow{\partial_n} C_{n-1} \to \cdots$$

of abelian groups is called a *chain complex* if $\partial_n \circ \partial_{n+1} = 0$ for all n.

Chain complexes are among the fundamental objects of study in homological algebra, a powerful algebraic formulation and generalization of the homology we study in topology.

13.4 Homology

At last, we can define homology. We know that $B_n(X) \leq Z_n(X)$; the homology is a measure of the discrepancy between these two groups.

Definition 13.6 The n^{th} *homology group* of X is the quotient group $H_n(X) = \frac{Z_n(X)}{B_n(X)}$.

It is true, although we will not prove it here, that the homology groups do not depend on the choice of triangulation of X. (See [Mun84, Section 18] for a proof.) Hence, the homology groups are a *homeomorphism invariant*. In fact, they are also a

homotopy invariant: two spaces that are homotopy equivalent have the same homology groups. Furthermore, the triangulation does not have to satisfy our strict rules for triangulations as defined in Chapter 3. Instead, simplices are allowed to meet in more complicated ways, without any ill effects. In particular, the additional possibilities we allow are that lower-dimensional faces of a simplex can be glued together, and the intersection of two simplices must be a union of their lower-dimensional faces. In addition, we are allowed to glue parts of the boundary of a simplex together, as long as they are of the same type: that is, we are allowed to glue together two vertices of a simplex, or two edges, and so forth. The official name of a triangulation with these weaker rules about intersection types is a Δ-*complex*, and perhaps we should correspondingly call our homology Δ-*homology*. But this name is less commonly used than simplicial homology, so we will stick to the term *simplicial homology*.

A very simple example of a Δ-complex that takes advantage of our looser rules for intersection types is a (filled-in) triangle with two vertices glued together, as shown in Figure 13.2. This isn't a simplicial complex or a triangulation in the usual sense, but we allow it among our Δ-complexes and could compute its homology if desired.

Let us compute the homology groups of a torus, as shown in Figure 13.3. Note that this isn't a triangulation in the sense of Chapter 3, because the two 2-simplices U and L intersect at three edges e, f, and g. But this is the sort of thing we allow in a Δ-complex. This is a huge help, because it means that we can get away with far fewer simplices than would be needed to give a triangulation in the sense of Chapter 3, and we still get the correct answer.

We need to calculate the boundary maps for the torus. For the 2-chains, we have $\partial_2(U) = -e - f + g$ and $\partial_2(L) = e + f - g$. We can see this geometrically: as we go around the edges of U (for example), in the direction indicated by the orientation

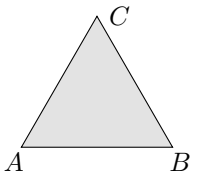


Figure 13.2 Gluing vertices A and B (and no other points) gives a Δ-complex.

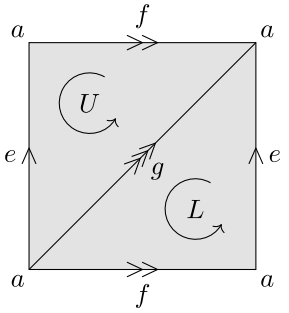


Figure 13.3 A triangulation of a torus.

of U, we go along g in the "right direction," but we go along e and f in the "wrong direction." Hence, the signs on $\partial_2(U)$ are positive for g and negative for e and f.

For the 1-chains, we have $\partial_1(e) = \partial_1(f) = \partial_1(g) = a - a = 0$: there's only one vertex a, and all three of those edges both start and end at a. All other boundary maps in dimensions other than 1 and 2 are 0.

Now let us compute the homology. We start with H_2. We have $Z_2(\mathbb{T}) = \langle U + L \rangle$, because $\partial_2(U + L) = 0$. Now, $B_2(\mathbb{T}) = 0$, because there are no 3-chains. Hence $H_2(\mathbb{T}) = \langle U + L \rangle \cong \mathbb{Z}$: the free abelian group with one generator, called $U + L$.

Now let us compute H_1. We have $Z_1(\mathbb{T}) = \langle e, f, g \rangle$,[2] because all the boundaries are 0, whereas $B_1(\mathbb{T}) = \langle e + f - g \rangle$. Calculating $H_1(\mathbb{T})$ out of this data requires a bit of finesse now: it's $Z_1(\mathbb{T})/B_1(\mathbb{T})$, but what is that as an abstract abelian group? If we quotient out by $\langle e + f - g \rangle$, that means that $e + f - g = 0$ in $H_1(\mathbb{T})$. Thus, we may "solve for g" and replace every instance of g with $e + f$. So, given a cycle $a_1 e + a_2 f + a_3 g$, we may rewrite that as $(a_1 + a_3)e + (a_2 + a_3) f$ in $H_1(\mathbb{T})$. Furthermore, every cycle involving only e's and f's is distinct in $H_1(\mathbb{T})$, so we have

$$H_1(\mathbb{T}) \cong \langle e, f \rangle \cong \mathbb{Z}^2.$$

Finally, there's H_0. We have $Z_0(\mathbb{T}) = C_0(\mathbb{T}) = \langle a \rangle \cong \mathbb{Z}$, whereas $B_0(\mathbb{T}) = 0$, because the boundary of every 1-chain is 0. Hence $H_0(\mathbb{T}) \cong \mathbb{Z}$. All other homology groups are 0. Thus we have calculated:

$$H_n(\mathbb{T}) \cong \begin{cases} \mathbb{Z} & n = 0, 2, \\ \mathbb{Z}^2 & n = 1, \\ 0 & n \geq 3. \end{cases}$$

13.5 The Zeroth Homology Group

The 0-dimensional homology group H_0 is easy to understand in general. If X is path-connected, then $H_0(X) \cong \mathbb{Z}$. Why? Pick a vertex (0-simplex) a in the triangulation of X. Then $\{na\}$ are all distinct elements in $H_0(X)$, for if ma and na were equal in $H_0(X)$, then they would have to differ by a boundary $\partial_1(c)$ for some $c \in C_1(X)$. However, the sum of the coefficients in $\partial_1(c)$ is always zero, so this cannot happen. Now, suppose that a and b are two vertices in X. Then there is some sequence of (oriented) edges $e_1, \ldots, e_r$ that starts at a and ends at b. Thus $\partial_1(e_1 + \cdots + e_r) = b - a$, so $b - a \in B_0(X)$, so it is zero in $H_0(X)$, i.e. $a = b$ in $H_0(X)$. Thus we have shown the following.

Proposition 13.7 *If X is path-connected, then $H_0(X) \cong \mathbb{Z}$.*

[2] When discussing homology, all our presentations of groups will be of *abelian* groups. That means the relations implying that the generators commute will be omitted when we write our presentations.

More generally, if X consists of k path-components, then $H_0(X) \cong \mathbb{Z}^k$ for the same reason.

13.6 Homology of the Klein Bottle

We now compute the homology of the Klein bottle, which will turn out to have a special surprise! We can use almost the same picture as a torus, but now one of the edges (say, the left edge) must switch orientation, as shown in Figure 13.4.

We now compute the boundary maps. We have

$$\begin{aligned} \partial_2(U) &= e - f + g, \\ \partial_2(L) &= e + f - g, \\ \partial_1(e) = \partial_1(f) &= \partial_1(g) = 0. \end{aligned}$$

Hence

$$\begin{aligned} Z_2(X) &= 0, \\ B_2(X) &= 0, \\ Z_1(X) &= \langle e, f, g \rangle, \\ B_1(X) &= \langle e - f + g, e + f - g \rangle. \end{aligned}$$

We don't need to compute H_0, because we already know it: The Klein bottle is path-connected, so $H_0(\mathbb{K}) \cong \mathbb{Z}$.

Clearly $H_2(\mathbb{K}) = 0$, but what about $H_1(\mathbb{K})$? We have

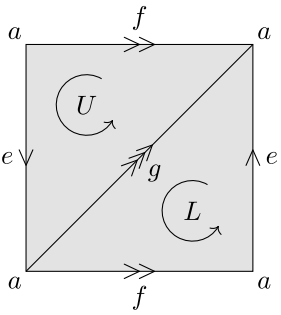


Figure 13.4 A triangulation of a Klein bottle.

$$\begin{aligned} H_1(\mathbb{K}) &= Z_1(\mathbb{K})/B_1(\mathbb{K}) \\ &= \langle e, f, g \rangle / \langle e - f + g, e + f - g \rangle \\ &= \langle e, f, g \mid e - f + g, e + f - g \rangle \\ &= \langle e, f, g \mid 2e, e + f - g \rangle \\ &= \langle e, f \mid 2e \rangle \\ &\cong \mathbb{Z} \times (\mathbb{Z}/2\mathbb{Z}). \end{aligned}$$

This means that, while e is not a boundary of any 2-dimensional submanifold, $2e$ is! What is this submanifold? It's just the Klein bottle itself: cut along e, and you have a cylinder, with two boundary components, both labeled e. But here, the boundary components have the same orientation, so the boundary consists of *both* copies of e. Contrast this with the case of the torus, where we can again cut along e to obtain a cylinder, but then the boundary components are oriented in opposite directions, so they cancel out in $Z_1(\mathbb{T}^2)$.

13.7 Homology and Euler Characteristic

Homology is a generalization of the Euler characteristic. In order to understand what that means and how it works, we must first define the *Betti numbers*. If X is a topological space that has a finite triangulation, then it has finitely many nonzero homology groups; furthermore, each homology group is a finitely generated abelian group, and hence of the form $H_i(X) \cong \mathbb{Z}^k \times \prod_{j=1}^{m} (\mathbb{Z}/p_j^{e_j}\mathbb{Z})$. We define the i^{th} *Betti number* $h_i(X)$ to be k, the number of copies of $\mathbb{Z}$, also known as the *rank* of a finitely generated abelian group.

Theorem 13.8 *The Euler characteristic is the alternating sum of the Betti numbers, i.e.*

$$\chi(X) = \sum_{i=0}^{\infty} (-1)^i h_i(X).$$

Note that this is actually a *finite* sum, because all but finitely many of the Betti numbers are zero.

Proof Recall that the Euler characteristic is the alternating sum of the number of faces of dimension i. Now, it is *not* true that the number of faces of dimension i is equal to the i^{th} Betti number, only that their alternating sums are equal. (They couldn't possibly be equal in general, because the number of faces depends on the choice of triangulation, and we have stated that the homology does not depend on the choice of triangulation.) But note that the number of faces of dimension i is the rank of $C_i(X)$. So we have to show that

$$\sum_{i=0}^{\infty}(-1)^i \operatorname{rank} C_i(X) = \sum_{i=0}^{\infty}(-1)^i \operatorname{rank} H_i(X).$$

To do this, observe that $\operatorname{rank} H_i(X) = \operatorname{rank} Z_i(X) - \operatorname{rank} B_i(X)$; each new relation decreases the number of "free" generators by one. Also, recall the isomorphism theorem for groups: if $f : G \to H$ is a homomorphism of groups, then $G/\ker(f) \cong \operatorname{im}(f)$. This implies that if G is a finitely generated abelian group, then

$$\operatorname{rank} G = \operatorname{rank} \operatorname{im}(f) + \operatorname{rank} \ker(f).$$

Now we're ready to go! By the above, using the boundary maps, we have

$$\operatorname{rank} C_i(X) = \operatorname{rank} B_{i-1}(X) + \operatorname{rank} Z_i(X).$$

So

$$\begin{aligned}
\chi(X) &= \sum_{i=0}^{\infty}(-1)^i \operatorname{rank} C_i(X) \\
&= \sum_{i=0}^{\infty}(-1)^i (\operatorname{rank} B_{i-1}(X) + \operatorname{rank} Z_i(X)) \\
&= \sum_{i=0}^{\infty}(-1)^i (\operatorname{rank} Z_i(X) - \operatorname{rank} B_i(X)) \\
&= \sum_{i=0}^{\infty}(-1)^i \operatorname{rank} H_i(X),
\end{aligned}$$

which is what we wanted to show! ■

Exercise 13.9 As a sanity check, verify that Theorem 13.8 holds for the torus and the Klein bottle.

13.8 Homology and Orientability

Let S be a compact, connected surface (without boundary). The homology detects the orientability of S, in the following way. Note that $H_2(\mathbb{T}) \cong \mathbb{Z}$, whereas $H_2(\mathbb{K}) = 0$. In general, the 2-dimensional homology of S is $\mathbb{Z}$ if S is orientable, and it's 0 if S is nonorientable. More generally, if X is a compact, connected n-dimensional manifold (without boundary), then $H_n(X) \cong \mathbb{Z}$ if X is orientable, and $H_n(X) = 0$ if X is nonorientable. We'll only prove this for surfaces since we'll work in terms of ID spaces.

Let's first suppose that S is orientable, and that we have an ID space for S, which is a polygon with edges identified in pairs. As we recall from Chapter 4, because S is orientable, the edges are Type I edges, i.e. as we traverse the boundary of the ID space polygon, the two instances of that edge appear with opposite orientations.

Now, split the ID space for S up into triangles so that we have a triangulation of S, into triangles $T_1, T_2, \ldots, T_r$. We orient each triangle T_i in the counterclockwise orientation. Then a 2-chain c is a sum $\sum_{i=1}^{r} a_i T_i$, where each $a_i \in \mathbb{Z}$. What does it mean for c to be a 2-cycle? Take two triangles that share an edge in the interior of the polygon, say T_i and T_j, which share edge e. These are the only two triangles containing that edge, so they are the only contributors to e in $\partial_2(c)$. Thus $\partial_2(a_i T_i + a_j T_j)$ must have a coefficient of 0 for e. The contribution from $a_i T_i$ is a_i, whereas the contribution from $a_j T_j$ is $-a_j$ (or the signs may both be swapped). Thus we find that a necessary condition for c to be a 2-cycle is that $a_i = a_j$. Because we can apply this argument to an arbitrary interior edge, we find that all the a_i's must be equal, i.e. it must be the case that $c = \sum_{i=1}^{r} a T_i$ for some $a \in \mathbb{Z}$.

But is such a c actually a cycle? The only thing that can go wrong is that the boundary edges of the polygon might not cancel. However, since S is assumed to be orientable, each boundary edge appears once with a positive orientation and once with a negative orientation. Thus chains of the form $\sum_{i=1}^{r} a T_i$ are indeed cycles, and they are the only cycles in S. If $a \neq 0$, then they are not boundaries, because $C_3(S) = 0$. Thus we find that $H_2(S) = \langle \sum_{i=1}^{r} T_i \rangle \cong \mathbb{Z}$.

Definition 13.10 Let S be an orientable, compact, connected surface divided into triangles $T_1, \ldots, T_r$ as above. Then the cycle $\sum_{i=1}^{r} T_i$ is called a *fundamental class* of S.

Thus a fundamental class generates $H_2(S)$. Note that we could have reflected our ID space polygon, which would flip the orientation of all the triangles and thus multiplied the fundamental class by -1. This means that there are two possible choices for a fundamental class of S.

Now, what happens if S is nonorientable? As before, we find that a *necessary* condition for c being a cycle is that it has the form $c = \sum_{i=1}^{r} a T_i$. However, now we run into a problem with the boundary. Since S is nonorientable, there is some edge e on the boundary that is oriented in the same way both times. Thus the contribution of e to $\partial_2(c)$ is $\pm 2a$. In particular, $c \notin Z_2(S)$ unless $a = 0$. Thus $Z_2(S) = 0$, so $H_2(S) = 0$ as well.

13.9 Smith Normal Form

It seems as though computing homology is easy and completely mechanical—so that the process is something that one could program a computer to do. But there is one step that is still difficult. Once we have computed $Z_i(X)$ and $B_i(X)$, we obtain *some* presentation for $H_i(X)$, but we would like to be able to identify it in a more

convenient form. If X has a finite triangulation, then $H_i(X)$ is a finitely generated abelian group, and we know what all the finitely generated abelian groups look like. But when we see a group like

$$\langle a_1, a_2, a_3, a_4 \mid 5a_1 - 2a_2 + 3a_4, 3a_1 + 2a_2 + 2a_3, 4a_3 - 2a_4, 9a_2 + 6a_3 \rangle, \tag{13.1}$$

how do we write that nicely, in the form $\mathbb{Z}^k \times$(finite group)?

Fortunately, there is a fairly simple algorithm for doing this. It will be convenient to write out the relations as a matrix. Each relation gets a row, and each generator gets a column, and the coefficients go in the matrix. Hence the matrix we get from the presentation (13.1) is

$$\begin{pmatrix} 5 & -2 & 0 & 3 \\ 3 & 2 & 2 & 0 \\ 0 & 0 & 4 & -2 \\ 0 & 9 & 6 & 0 \end{pmatrix}.$$

The goal is to find *better* generators and relations, ones that make it more obvious what the group structure is.

So, how do we find other generators? If a_1 and a_2 generate an abelian group, then a_1 and $a_1 + a_2$ generate it just as well, as do a_1 and $5a_1 + a_2$. More generally, if a_1 and a_2 are two of the generators for an abelian group (and there may be others), then a_1 and $ca_1 + a_2$, for any integer c, together with the remaining generators, also generate the same group.

What happens to the matrix when we modify the generators in this way? Replacing a_2 with $ca_1 + a_2$ means we add c times the a_1 column to the a_2 column. For example, in the matrix above, replacing a_2 by $2a_1 + a_2$ would turn the matrix into

$$\begin{pmatrix} 5 & 8 & 0 & 3 \\ 3 & 8 & 2 & 0 \\ 0 & 0 & 4 & -2 \\ 0 & 9 & 6 & 0 \end{pmatrix}.$$

Also allowable is switching the order of the generators, which amounts to switching the order of the columns.

Similarly, we can modify the relations: if r_1 and r_2 are two relations, then $cr_1 + r_2$ is also a relation, and we can replace r_2 with $cr_1 + r_2$ in the list of relations. Thus we can do the same operations to the rows as we can to the columns. Using these row and column operations, we can convert the matrix to one that is in a particularly nice form, called the Smith normal form.

Definition 13.11 A matrix $A = (a_{ij})_{\substack{1 \le i \le m \\ 1 \le j \le n}}$ is said to be in *Smith normal form* if, for some $k \le \min(m, n)$, the first k diagonal entries are nonzero, with a_{ii} dividing $a_{i+1,i+1}$, and all other entries are 0.

Thus, a matrix in Smith normal form looks like this:

$$\begin{pmatrix} a_{11} & & & & & & \\ & a_{22} & & & & & \\ & & \ddots & & & & \\ & & & a_{kk} & & & \\ & & & & 0 & & \\ & & & & & \ddots & \\ & & & & & & 0 \end{pmatrix},$$

with all other omitted entries being 0.

It turns out that we can put any integer matrix into Smith normal form; this fact is equivalent to the classification of finitely generated abelian groups, and in fact this is the idea behind the standard proof of that classification theorem. (We didn't quite prove it properly.) The details of how to get a matrix into Smith normal form are slightly outside the scope of this book (see [Mun84, Section 11] for a proof). This process involves some elementary number theory (in particular, the Euclidean algorithm), but the idea is as follows. If there is some nonzero entry in the matrix, then let g be the gcd of all the entries in the matrix. Using the Euclidean algorithm repeatedly, it is possible to perform row and column operations so that $a_{11} = g$. Then, everything else in the first row and column is a multiple of g, so we can add multiples of the first row or column to the others—so as to make all the other entries in the first row and first column 0. Now, if there are any nonzero entries left in the matrix other than a_{11}, do the same process to the matrix obtained by deleting the first row and column, making a_{22} into the gcd of all the rest of the elements, and so forth.

Once we have a matrix in Smith normal form, it is easy to read off the group structure so as to fit it into the classification. Let us say that the columns correspond to generators $g_1, \ldots, g_n$. If the i^{th} column is all 0's, then g_i contributes a factor of $\mathbb{Z}$ to the group. If there is a nonzero entry in column i (hence the entry a_{ii}), then g_i contributes a factor of $\mathbb{Z}/a_{ii}\mathbb{Z}$ to the group.

Why? Let us look at the i^{th} row. Once the matrix is in Smith normal form, there is at most one nonzero entry in that row; if all the entries in row i are 0, then that corresponds to the relation $0 = 0$, which we already knew. But if there is a nonzero entry a_{ii}, then the relation is $a_{ii}e_i = 0$, giving us a factor of $\mathbb{Z}/a_{ii}\mathbb{Z}$.

In the case of the matrix above, the Smith normal form is

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 336 \end{pmatrix},$$

meaning that the corresponding group is isomorphic to $\mathbb{Z}/336\mathbb{Z}$.

13.10 The Induced Map on Homology

Recall that, given two topological spaces X and Y, base points $x \in X$ and $y \in Y$, and a continuous function $f : X \to Y$ with $f(x) = y$, there is a homomorphism $f_* : \pi_1(X, x) \to \pi_1(Y, y)$. This is the induced homomorphism.

Similarly, in the case of homology, we also have an induced homomorphism—at least sometimes. Let us think about how to mimic the construction of the induced homomorphism on fundamental groups, in the case of homology. Remember that, given a loop γ in X based at x, we set $f_*([\gamma]) = [f \circ \gamma]$. The analogous construction in the case of homology, which would give us maps $f_* : H_n(X) \to H_n(Y)$ for each n, would be to pick an n-chain $\sum_{i=0}^{r} a_i T_i$ and define f_* *on chains* by $f_*\left(\sum_{i=0}^{r} a_i T_i\right) = \sum_{i=0}^{r} a_i f(T_i)$, where $f(T_i)$ is the image of the simplex T_i under f. That would be a map of *chains*, so a homomorphism $f_* : C_n(X) \to C_n(Y)$. The grand goal would be to show that, when we restrict f_* to $Z_n(X)$ and $B_n(X)$, we have $f_*(Z_n(X)) \leq Z_n(Y)$ and $f_*(B_n(X)) \leq B_n(Y)$.

But there is a problem: if T_i is a simplex of X, then we have no guarantee that $f(T_i)$ is a simplex of Y; it might just be some fairly arbitrary subset of Y with no nice properties. In order for everything to work out, we need to ensure that the image of every simplex of X is a simplex of Y.

Definition 13.12 Let X and Y be simplicial complexes and $f : X \to Y$ a continuous function. We say that f is a *simplicial map* if the image of every simplex of X is a simplex of Y.

Most maps are not simplicial, but for simplicial maps, the idea above for constructing an induced homomorphism on homology works perfectly.

Proposition 13.13 *Suppose $f : X \to Y$ is a simplicial map. Then, for each n, $f_*(Z_n(X)) \leq Z_n(Y)$ and $f_*(B_n(X)) \leq B_n(Y)$.*

Proof We show first that, for any n-simplex T of X, $\partial_n \circ f_*(T) = f_* \circ \partial_n(T)$. Suppose that $T = [v_0, \ldots, v_n]$ and that $f_*(T) = [w_0, \ldots, w_n]$, with $w_i = f(v_i)$. Then we have

$$\begin{aligned}
\partial_n \circ f_*(T) &= \partial_n([w_0, \ldots, w_n]) \\
&= \sum_{i=0}^{n} (-1)^i [w_0, \ldots, \widehat{w_i}, \ldots, w_n] \\
&= \sum_{i=0}^{n} (-1)^i f([v_0, \ldots, \widehat{v_i}, \ldots, v_n]) \\
&= f_*\left(\sum_{i=0}^{n} (-1)^i [v_0, \ldots, \widehat{v_i}, \ldots, v_n] \right) \\
&= f_* \circ \partial_n(T),
\end{aligned}$$

as desired.

Now, suppose that $c = \sum_{i=0}^{r} a_i T_i \in Z_n(X)$, so that $\partial_n(c) = 0$. Then

$$0 = f_* \circ \partial_n(c) = \partial_n \circ f_*(c),$$

so $f_*(c) \in Z_n(Y)$. Similarly, if $c = \partial_n(d) \in B_n(X)$, then

$$f_*(c) = f_* \circ \partial_n(d) = \partial_n \circ f_*(d),$$

so $f_*(d) \in B_n(Y)$. ■

Remark 13.14 It may be the case that $f(v_i) = f(v_j)$ for some $i \neq j$, so that $f_*(T)$ is a lower-dimensional simplex. It is okay if a vertex appears multiple times in $f_*(T)$; this means that we treat $f_*(T)$ formally as an n-dimensional simplex; nothing in our definitions ever has to "know" that $f_*(T)$ is secretly lower-dimensional.

It is immediate from the definition that f_* is a homomorphism.

Now we are ready to define the induced homomorphism on homology. Let $f : X \to Y$ be a simplicial map, let $c \in Z_n(X)$, and let $[c] = c + B_n(X)$ be its class in homology. Then we define $f_* : H_n(X) \to H_n(Y)$ by setting $f_*([c]) = [f_*(c)]$. Let us verify that f_* is well-defined. Suppose $[c] = [c']$. Then $c - c' = d$ for some $d \in B_n(X)$. We have

$$f_*(c) - f_*(c') = f_*(d) \in f_*(B_n(X)) \leq B_n(Y).$$

Thus $[f_*(c)] = [f_*(c')]$, so $f_* : H_n(X) \to H_n(Y)$ is indeed well-defined.

The induced homomorphism on homology satisfies all the same basic properties as does the induced homomorphism on fundamental groups. For example:

1. If $f : X \to Y$ and $g : Y \to Z$ are simplicial maps, then $(g \circ f)_* = g_* \circ f_*$.
2. If $f : X \to Y$ is a homotopy equivalence, then f_* is an isomorphism.

There are many types of homology. The one we have been using is called *simplicial homology* or Δ-*homology*. (Which one we're using depends on the requirements we put on the intersections of simplices.) This is one of the more computable formulations of homology. Assuming there are only finitely many simplices, then the computation is a finite problem—unlike in the case of fundamental groups, where there are potentially infinitely many loops on a space X that must be considered. One downside of using simplicial or Δ-homology, though, is that we are only able to consider simplicial maps. The theorems remain true if we replace "simplicial maps" in all our theorem statements with arbitrary continuous maps, but the proofs no longer hold. Furthermore, we have claimed—but not proven—that homology does not depend on the choice of triangulation, and the proof requires some serious work! With other formulations, especially *singular* homology (see Appendix B), it is immediately clear that homology is a homeomorphism invariant and plays nicely with all continuous maps. The cost, however, is that computation becomes very challenging and is not a finite problem—at least, not without some serious theorems. All

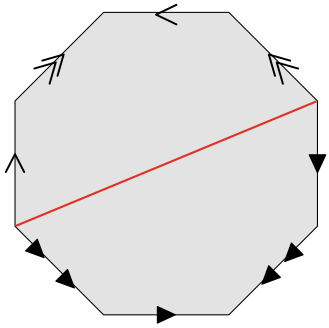


Figure 13.5 A loop (marked in red) in a genus-2 surface.

homology theories[3] give the same answers in the case of "nice" spaces, but there are pathological spaces where they may disagree.

13.11 Problems

(1) Compute the homology groups of $\mathbb{S}^2$ and $\mathbb{RP}^2$.
(2) (a) Compute the homology of a genus-2 surface.
    (b) In terms of your generators from part (a), what is the red curve in Figure 13.5?
(3) Compute the homology of the space given by the ID space aaa.
(4) (a) Compute the homology of the n-dimensional sphere $\mathbb{S}^n$.
    (b) Show that if $m \neq n$, then $\mathbb{R}^m$ is not homeomorphic to $\mathbb{R}^n$.
(5) Consider the space obtained by taking a 2-simplex with vertices v_0, v_1, v_2 and identifying the edges $[v_0, v_1]$ and $[v_1, v_2]$. Compute its homology. Is this space homeomorphic to a space you are familiar with?
(6) Consider the space obtained by taking a 2-simplex and identifying all its vertices. Compute the homology groups of this space.
(7) Using the induced map on homology, prove the Brouwer Fixed-Point Theorem in n dimensions: let $f : D_n \to D_n$ be a continuous map from the n-dimensional unit disk to itself. Then there must be a point $x \in D_n$ so that $f(x) = x$. (You may use the fact that there is an induced homomorphism associated to *any* continuous map, not just a simplicial map.)

[3] There is a set of axioms, known as the *Eilenberg–Steenrod axioms*, which all homology theories must satisfy. See [Mun84, Section 26] for a list of these axioms. Unfortunately, we cannot discuss them here as they rely on the notion of relative homology, which would be too large of a diversion.

Chapter 14
The Mayer–Vietoris Sequence

14.1 Exact Sequences

Although it is possible to compute homology directly from the definition, it is not always much fun to do so—computing the homology for a genus-g surface would require a lot of simplices and matrix manipulations! We were able to compute the *fundamental group* for an arbitrary surface using the Seifert–Van Kampen Theorem, breaking it up into smaller regions and splicing together their fundamental groups. In particular, we were able to express $\pi_1(A \cup B)$ in terms of $\pi_1(A)$, $\pi_1(B)$, $\pi_1(A \cap B)$, and some information about how they all fit together. It would be nice if we could do that for homology as well.

In fact, we can relate the homology of $A \cup B$ to the homologies of A, B, and $A \cap B$, but the relation is a bit more complicated than in the case of fundamental groups. In particular, we relate $H_n(A \cup B)$ not just to $H_n(A)$, $H_n(B)$, and $H_n(A \cap B)$, but to all the homologies of these spaces, as well as to all the other homology of $A \cup B$. Before we can state this connection, known as the Mayer–Vietoris sequence, we need to introduce the notion of an *exact sequence*.

Recall *chain complexes* from the last chapter: these are sequences

$$\cdots \to A_{n+1} \xrightarrow{f_{n+1}} A_n \xrightarrow{f_n} A_{n-1} \to \cdots$$

of abelian groups and maps between them such that composing any two consecutive maps gives the zero map, i.e. $f_n \circ f_{n+1} = 0$ for any n. This means that $\operatorname{im}(f_{n+1}) \leq \ker(f_n)$.

Definition 14.1 A sequence

$$\cdots \to A_{n+1} \xrightarrow{f_{n+1}} A_n \xrightarrow{f_n} A_{n-1} \to \cdots$$

of abelian groups and maps is said to be *exact at* A_n if $\operatorname{im}(f_{n+1}) = \ker(f_n)$. It is said to be *exact* if it is exact at all A_n.

C. Bray et al., *Algebraic Topology*,
https://doi.org/10.1007/978-3-030-70608-1_14

It is sometimes the case that a sequence does not go on forever, or perhaps it only goes on forever in one direction. In this case, we say it is exact if it is exact at all positions other than the end or ends of the sequence.

It is useful to distinguish between two types of exact sequences: short ones and long ones.

Definition 14.2 An exact sequence of the form

$$0 \to A' \to A \to A'' \to 0$$

is said to be a *short exact sequence*.

Naturally, we thus call a sequence with more than three (potentially) nonzero terms a *long exact sequence*.

Example If A and B are any two abelian groups, then we have a short exact sequence

$$0 \to A \xrightarrow{f} A \times B \xrightarrow{g} B \to 0, \tag{14.1}$$

where f is defined by $f(a) = (a, 0)$ and g is defined by $g(a, b) = b$. Let us check that this is exact at A. The image of the map $0 \to A$ is just 0. The kernel of f is 0, because if $a \neq 0$, then $(a, 0) \neq (0, 0) \in A \times B$. Now let us check exactness at B. The kernel of $B \to 0$ is all of B: everything gets mapped to 0. The image of g is also all of B because, for any $b \in B$, $g(0, b) = b$. Finally, let us check exactness at $A \times B$. The image of f is $\{(a, 0)\}$, and this is also the kernel of g. Hence the sequence is exact at each position. Short exact sequences of the form (14.1) are known as *split exact sequences*.

Note that exactness at A' means that $A' \to A$ is injective, and exactness at A'' means that $A \to A''$ is surjective.

Example If m and n are any positive integers, consider the sequence

$$0 \to \mathbb{Z}/m\mathbb{Z} \to \mathbb{Z}/mn\mathbb{Z} \to \mathbb{Z}/n\mathbb{Z} \to 0,$$

where the maps $\mathbb{Z}/m\mathbb{Z} \to \mathbb{Z}/mn\mathbb{Z}$ and $\mathbb{Z}/mn\mathbb{Z} \to \mathbb{Z}/n\mathbb{Z}$ are given by $a + m\mathbb{Z} \mapsto an + mn\mathbb{Z}$ and $b + mn\mathbb{Z} \mapsto b + n\mathbb{Z}$, respectively. This sequence is a short exact sequence. If m and n are relatively prime, then this sequence is a split exact sequence by the Chinese Remainder Theorem, but otherwise it is not. For example, if p is prime, the sequence

$$0 \to \mathbb{Z}/p\mathbb{Z} \to \mathbb{Z}/p^2\mathbb{Z} \to \mathbb{Z}/p\mathbb{Z} \to 0$$

is exact but not split exact.

In fact, we can in some sense classify all short exact sequences: If A and B are two abelian groups with $A \leq B$, then $0 \to A \to B \to B/A \to 0$ is an exact sequence,

and any exact sequence can be expressed in this way. (See [Rot95, p. 307] for an explanation of why, or work it out yourself!)

14.2 The Mayer–Vietoris Sequence

Suppose X is a topological space (implicitly with a triangulation), and A and B are subspaces with $A \cup B = X$. Furthermore, assume that A and B are each unions of simplices in X.

Proposition 14.3 *For any n, we have a short exact sequence*

$$0 \to C_n(A \cap B) \xrightarrow{\alpha} C_n(A) \times C_n(B) \xrightarrow{\beta} C_n(A \cup B) \to 0$$

of abelian groups. Here α is given by inclusion: an n-chain c of $A \cap B$ is also an n-chain of A and an n-chain of B, so we define $\alpha(c) = (c, c)$. Similarly, an n-chain of A is an n-chain of $A \cup B$, and similarly for B. We define $\beta(c, d) = c - d$.

Remark 14.4 The minus sign in the definition of β is necessary to make this sequence exact!

Proof It is clear that α is injective, so this sequence is exact at $C_n(A \cap B)$. Now, let $N = \sum_{i=1}^{r} a_i T_i$ be an n-chain of $A \cup B$. Let us suppose that $T_1, \ldots, T_k$ are in A and $T_{k+1}, \ldots, T_r$ are in B. (Each simplex must be in either A or B; some of them may be in both, so let us—arbitrarily—group them with A.) Let $c = \sum_{i=1}^{k} a_i T_i$ and $d = \sum_{i=k+1}^{r} a_i T_i$. Then $c \in C_n(A)$ and $d \in C_n(B)$, and $N = c + d = \beta(c, -d)$. Thus β is surjective.

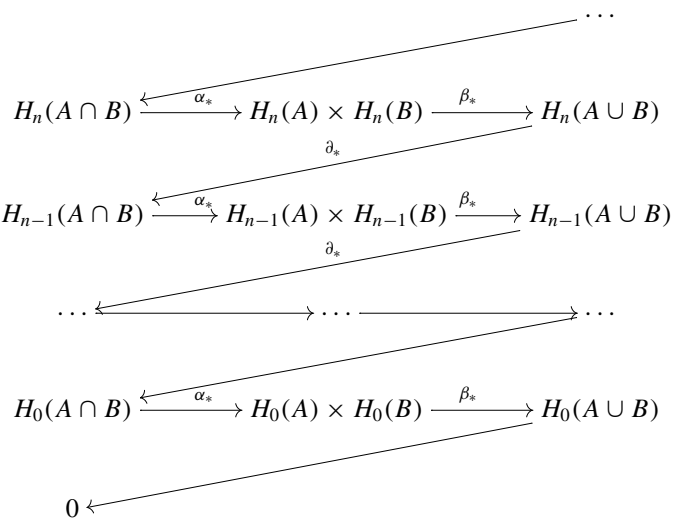
Finally, we must show that $\operatorname{im} \alpha = \ker \beta$. We have

$$\operatorname{im} \alpha = \ker \beta = \{(c, c) : c \in C_n(A \cap B)\}.$$

Thus the sequence is also exact at $C_n(A) \times C_n(B)$. ■

A general phenomenon in mathematics is that *a short exact sequence of chains induces a long exact sequence in homology*. In this case, the resulting long exact sequence is the Mayer–Vietoris sequence.

Theorem 14.5 (Mayer–Vietoris) *Let $X = A \cup B$, where X, A, and B are equipped with triangulations. Then there is a long exact sequence*



of homology groups.

It is possible to prove—for once and for all—that short exact sequences at the level of chains give rise to long exact sequences in homology. We will think in terms of homology of topological spaces, but nothing in the proof depends on that. We'll leave certain aspects of the proof for Problems 3–5.

Proof The first step in the proof is to define the maps α_*, β_*, and ∂_*. The first two are straightforward: they are the induced maps in homology coming from the maps α and β of Proposition 14.3. The map ∂_* is more complicated: for each n, we wish to construct a homomorphism $\partial_* : H_n(A \cup B) \to H_{n-1}(A \cap B)$. Observe the following diagram of abelian groups:

$$\begin{array}{ccccc} C_n(A \cap B) & \xrightarrow{\alpha_n} & C_n(A) \times C_n(B) & \xrightarrow{\beta_n} & C_n(A \cup B) \\ \downarrow{\scriptstyle \partial^{(1)}} & & \downarrow{\scriptstyle \partial^{(2)}} & & \downarrow{\scriptstyle \partial^{(3)}} \\ C_{n-1}(A \cap B) & \xrightarrow{\alpha_{n-1}} & C_{n-1}(A) \times C_{n-1}(B) & \xrightarrow{\beta_{n-1}} & C_{n-1}(A \cup B) \end{array} \tag{14.2}$$

This diagram *commutes*. This means that if we pick either square of the diagram and start from the top left corner and then take the horizontal arrow followed by the vertical arrow, we get the same result as we do if we first take the vertical one then the horizontal one: The square

$$\begin{array}{ccc} W & \xrightarrow{\phi} & X \\ \downarrow{\scriptstyle \psi} & & \downarrow{\scriptstyle \theta} \\ Y & \xrightarrow{\rho} & Z \end{array}$$

commutes if and only if $\theta \circ \phi(w) = \rho \circ \psi(w)$ for all $w \in W$.

In order to construct $\partial_* : H_n(A \cup B) \to H_{n-1}(A \cap B)$, we first *attempt* to construct a homomorphism $\delta : Z_n(A \cup B) \to Z_{n-1}(A \cap B)$ in an interesting way. We will fail. But we will fail in *exactly* the way we need in order to get a homomorphism on homology! Our attempt is as follows:

(1) Pick $x \in Z_n(A \cup B)$; this means that $\partial^{(3)}(x) = 0$.
(2) We saw earlier that β_n is surjective, so we can find some $y \in C_n(A) \times C_n(B)$ so that $\beta_n(y) = x$. (There may be many choices for y; pick one at random.)
(3) Now, look at $z = \partial^{(2)}(y)$. Because the right square of (14.2) commutes, we have

$$\partial^{(3)} \circ \beta_n(y) = \beta_{n-1} \circ \partial^{(2)}(y) = 0,$$

so $\beta_{n-1}(z) = 0$.
(4) The bottom row of (14.2) is exact and $\beta_{n-1}(z) = 0$, so there is some $w \in C_{n-1}(A \cap B)$ so that $\alpha_{n-1}(w) = z$.
(5) We wish to set $\delta(x) = w$.

We now check that $w \in Z_{n-1}(A \cap B)$. Observe that $z \in Z_{n-1}(A) \times Z_{n-1}(B)$. (In fact, $z \in B_{n-1}(A) \times B_{n-1}(B)$, which is stronger, but we will not need this at the moment.) Consider the commutative square

$$\begin{array}{ccc} C_{n-1}(A \cap B) & \xrightarrow{\alpha_{n-1}} & C_{n-1}(A) \times C_{n-1}(B) \\ \downarrow{\scriptstyle \partial^{(1)}} & & \downarrow{\scriptstyle \partial^{(2)}} \\ C_{n-2}(A \cap B) & \xrightarrow{\alpha_{n-2}} & C_{n-2}(A) \times C_{n-2}(B) \end{array}$$

Because $\partial^{(2)}(z) = \partial^{(2)} \circ \alpha_{n-1}(w) = 0$, we also have $\alpha_{n-2} \circ \partial^{(1)}(w) = 0$; by Proposition 14.3, α_{n-2} is injective, so $\partial^{(1)}(w) = 0$, so $w \in Z_{n-1}(A \cap B)$.

So, it appears that we have made a map $\delta : Z_n(A \cup B) \to Z_{n-1}(A \cap B)$. However, this is just an illusion. The problem is that *it is not well-defined*: we had many choices for y, and we just chose one at random. In fact, there is generally no systematic way of picking y so as to make δ into a homomorphism. (We can make a function, but it will not have the homomorphism property.)

Nonetheless, all is not lost: Imagine we have two elements y and y' in $C_n(A) \times C_n(B)$ with $\beta_n(y) = \beta_n(y')$. Construct z' and w' similarly to the way we constructed z and w before. Since $\beta_n(y) = \beta_n(y') = x$, we have $\beta_n(y - y') = 0$, which means by Proposition 14.3 that $y - y' \in \mathrm{im}(\alpha_n)$, say $y - y' = \alpha_n(v)$. Then $\partial^{(1)}(v) = w - w'$. By definition, this means that $w - w' \in B_{n-1}(A \cap B)$. As a result, although δ didn't give us a well-defined map from $Z_n(A \cup B)$ to $Z_{n-1}(A \cap B)$, it did give us a well-defined map to $Z_{n-1}(A \cap B)$ *up to a boundary*—which is exactly the same as a map to $H_{n-1}(A \cap B)$. Thus we have a well-defined map $\delta : Z_n(A \cup B) \to H_{n-1}(A \cap B)$. One can now check that δ is in fact a homomorphism, as you will do in Problem 3.

In fact, δ induces a well-defined map $\partial_* : H_n(A \cup B) \to H_{n-1}(A \cap B)$. In order to verify that, we have to check that if x and x' in $Z_n(A \cup B)$ differ by a boundary, then $\delta(x) = \delta(x')$; equivalently, $\delta(x - x') = 0$. We leave this for you to do in Problem 4.

Now that we have constructed the homomorphisms, we have to show that the sequence is exact. We need to show that the sequence is exact at $H_n(A \cap B)$, $H_n(A) \times H_n(B)$, and at $H_n(A \cup B)$. We prove exactness at $H_n(A) \times H_n(B)$ below and leave the other two for exercises. **Warning:** Diagram chases can be enjoyable to work out on your own, but they are never fun to read. Try it on your own first!

We start with exactness at $H_n(A) \times H_n(B)$. First, $\operatorname{im}(\alpha_*) \leq \ker(\beta_*)$, because this is true at the level of chains: if $x \in Z_n(A \cap B)$, then we have $[\beta \circ \alpha(x)] = \beta_* \circ \alpha_*([x])$, and the left side is zero by Proposition 14.3. For the other direction, suppose that $y \in Z_n(A) \times Z_n(B)$ and that $\beta_*([y]) = 0$. Then $\beta_n(y) \in B_n(A \cup B)$, so that $\beta_n(y) = \partial^{(3)}(z)$ for some $z \in C_{n+1}(A \cup B)$. Because β_{n+1} is surjective, there is some $w \in C_{n+1}(A) \times C_{n+1}(B)$ so that $\beta_{n+1}(w) = z$. Now, $\beta_n(y - \partial^{(2)}(w)) = 0$, so $y - \partial^{(2)}(w) = \alpha_n(v)$ for some $v \in C_n(A \cap B)$. Now, $\alpha_{n-1} \circ \partial^{(1)}(v) = \partial^{(2)} \circ \alpha_n(v) = 0$, and since α_{n-1} is injective, $\partial^{(1)}(v) = 0$, so $v \in Z_1(A \cap B)$. Finally, $\alpha_*([v]) = [y - \partial^{(2)}(w)] = [y]$, so $[y] \in \operatorname{im} \alpha_*$, as desired.

The arguments for exactness at $H_n(A \cap B)$ and $H_n(A \cup B)$ are similar diagram chases. We leave them for Problem 5. ■

14.3 Homology of Orientable Surfaces

Okay, so now that we have the theorem, it's time to learn how to use it! We will use it to compute the homology of an orientable surface of genus g. In order to do this, we first compute the homology of a once-punctured surface of genus g (or, up to homotopy equivalence, a surface of genus g with a small open disk removed). This will turn out to be easier, based on the following observation: we saw earlier that a punctured torus is homotopy equivalent to a wedge sum of two circles. More generally, a once-punctured surface of genus g is homotopy equivalent to a wedge sum of $2g$ circles. We will take advantage of the following fact that we did not prove (and whose proof is beyond the scope of this book): *If two spaces are homotopy equivalent, then they have the same homology groups*. So, we prove the following by induction using the Mayer–Vietoris sequence.

Proposition 14.6 *Let Y_r denote a wedge sum of r circles. Then*

$$H_n(Y_r) = \begin{cases} \mathbb{Z} & n = 0, \\ \mathbb{Z}^r & n = 1, \\ 0 & n \geq 2. \end{cases}$$

Proof When $r = 1$, the statement is true, so suppose $r \geq 2$ and work by induction. In the Mayer–Vietoris sequence, let A denote a wedge of $r - 1$ of the circles and let B denote the remaining circle, so that $A \cap B$ is a point and $A \cup B = Y_r$. Because A

and B have dimension 1, all homology in dimension ≥ 2 is zero. Thus, plugging in the values we know (and recalling that $H_0(Y_r) = \mathbb{Z}$ since Y_r is connected), we have an exact sequence

$$0 \to \mathbb{Z}^{r-1} \times \mathbb{Z} \to H_1(Y_r) \to \mathbb{Z} \to \mathbb{Z} \times \mathbb{Z} \to \mathbb{Z} \to 0.$$

Let us first think about the H_0 part of the sequence

$$\mathbb{Z} \to \mathbb{Z} \times \mathbb{Z} \to \mathbb{Z} \to 0.$$

The map $\mathbb{Z} \times \mathbb{Z} \to \mathbb{Z}$ is surjective, and its kernel is the subgroup of the form (n, n). But that is exactly the image of the map $\mathbb{Z} \to \mathbb{Z} \times \mathbb{Z}$, and this is injective. Thus the map $H_1(A \cup B) \to H_0(A \cap B)$ is the zero map (its image is the kernel of the map $\mathbb{Z} \to \mathbb{Z} \times \mathbb{Z}$, which is zero), so the map $H_1(A) \times H_1(B) \to H_1(A \cup B)$ must be surjective. It is also injective because $H_1(A \cap B) = 0$, so it is an isomorphism. Since we already know that $H_1(A) \cong \mathbb{Z}^{r-1}$ and $H_1(B) \cong \mathbb{Z}$, we have $H_1(A \cup B) = H_1(Y_r) \cong \mathbb{Z}^r$. ■

Now we have everything we need to compute the homology of surfaces of genus g.

Theorem 14.7 *Let X_g be a surface of genus g. Then*

$$H_n(X_g) \cong \begin{cases} \mathbb{Z} & n = 0, 2, \\ \mathbb{Z}^{2g} & n = 1, \\ 0 & n \geq 3. \end{cases}$$

Proof Let A be X_g with a point removed, and let B be a small neighborhood of the deleted point. Then $A \cap B$ is an annulus, which is homotopy equivalent to a circle. So, we know the homology groups of A, B, and $A \cap B$. Plugging in everything we know into the Mayer–Vietoris sequence, we have (starting with 0 for $H_2(A) \times H_2(B)$)

$$0 \to H_2(X_g) \to \mathbb{Z} \to \mathbb{Z}^{2g} \to H_1(X_g) \to 0,$$

where we end with a 0 because the map $H_1(X_g) \to H_0(A \cap B)$ is the zero map, as in Proposition 14.6. The map $H_2(X_g) \to H_1(A \cap B) \cong \mathbb{Z}$ is injective, and there are only two groups with injective maps to $\mathbb{Z}$: these are 0 and $\mathbb{Z}$ itself. So, which is it?

We actually already know the answer to this: in Section 13.8, we saw that the second homology group of a compact connected orientable surface is isomorphic to $\mathbb{Z}$. Thus $H_2(X_g) \cong \mathbb{Z}$.

Now, to compute $H_1(X_g)$, we note that the homomorphism $H_1(A) \times H_1(B) \to H_1(X_g)$ is surjective, and that $H_1(A) \cong \mathbb{Z}^{2g}$. Furthermore, we can use exactness to compute the rank of $H_1(X_g)$: in any exact sequence of finitely generated abelian groups starting and ending with 0's, the alternating sum of the ranks is equal to 0. (The argument is very similar to the one we used to show that the Euler characteristic

is the alternating sum of the ranks of the homology groups.) This shows that the rank of $H_1(X_g)$ is $2g$, and the only abelian group of rank $2g$ for which there is a surjection from $\mathbb{Z}^{2g}$ is $\mathbb{Z}^{2g}$ itself. This completes the proof. ■

14.4 The Jordan Curve Theorem

We now tackle one of the most infamous problems in mathematics: The Jordan Curve Theorem.

Theorem 14.8 (Jordan Curve Theorem) *Let $h : \mathbb{S}^1 \to \mathbb{S}^2$ be an injective continuous map, so that $h(\mathbb{S}^1)$ is a simple closed curve in $\mathbb{S}^2$. Then $\mathbb{S}^2 - h(\mathbb{S}^1)$ consists of two connected components.*

One can also consider $\mathbb{R}^2 - h(\mathbb{S}^1)$; the theorem holds in this case as well, with a similar proof. In the case of the plane, we can call the two regions "the inside" and "the outside." More precisely, one of the regions is bounded, and the other is unbounded.

The Jordan Curve Theorem seems obvious: in the case of a curve in the plane, it seems clear that there is an inside and an outside. In fact, we can probably even tell whether a point is on the inside or the outside. One popular way of doing this is using ray intersections. Pick a point x not on the curve. Draw a ray, in some direction, from x to ∞. If the ray intersects the curve an even number of times, it is on the outside, and if it intersects an odd number of times, it is on the inside.

This seems like a proof, but it is not. The problem is that the ray might intersect *infinitely many* times, and then we have learned nothing. In fact, it might be the case that *every* ray from x intersects the curve infinitely many times. Some complicated drawings of Jordan curves can be found in [RR11].

Fortunately, homology and the Mayer–Vietoris sequence provide us with a completely rigorous proof. We will work with the sphere version, although the planar version is similar. (See Problem 6.) We break the circle $\mathbb{S}^1$ up into two closed semicircles, which we call C^+ and C^-; their intersection consists of two points. Let $A = \mathbb{S}^2 - h(C^+)$ and $B = \mathbb{S}^2 - h(C^-)$. Thus $A \cup B$ is a sphere with two points deleted. It is also easy to show that A and B are both homeomorphic to $\mathbb{R}^2$. The mystery is in $A \cap B$, which is $\mathbb{S}^2 - h(\mathbb{S}^1)$.

Our goal is to compute $H_0(A \cap B)$, because H_0 tells us the number of connected components. As we saw in the previous chapter, H_0 is always of the form $\mathbb{Z}^r$ for some r, and in fact r is the number of connected components.

In order to compute $H_0(A \cap B)$, we throw A and B into the Mayer–Vietoris sequence and replace the terms we know. We already know the homology of $\mathbb{R}^2$ and of $\mathbb{S}^1 \times \mathbb{R}$ (being homotopy equivalent to $\mathbb{S}^1$), so we are only left with our mystery space. Starting with $H_2(A \cup B)$, the Mayer–Vietoris sequence becomes

$$0 \to H_1(A \cap B) \to 0 \to \mathbb{Z} \to H_0(A \cap B) \to \mathbb{Z} \times \mathbb{Z} \to \mathbb{Z} \to 0.$$

Because $H_1(A \cap B)$ is surrounded by zeros, it is zero, although this is not of major concern to us. We now compute $H_0(A \cap B)$. We use a similar trick to the one we used when showing that the alternating sum of the Betti numbers is the Euler characteristic: in an exact sequence bounded by zeros, the alternating sum of the ranks is 0. We know all the ranks in the exact subsequence

$$0 \to \mathbb{Z} \to H_0(A \cap B) \to \mathbb{Z} \times \mathbb{Z} \to \mathbb{Z} \to 0,$$

so we can compute the rank of $H_0(A \cap B)$, which is 2. Since H_0 is of the form $\mathbb{Z}^r$, we must have $r = 2$. Thus $A \cap B$ consists of two connected components, and we're done!

14.5 The Hurewicz Map

As we mentioned earlier, homology is an alternative invariant to the fundamental group and higher homotopy groups. But we might also wonder whether there is any connection between the homotopy groups and the homology groups. In particular, is there any connection between π_1 and H_1?

The answer is yes, at least when the space is path-connected. They can't always be the same—because H_1 is always abelian whereas π_1 doesn't have to be—but there's still a connection between the two. In order to state the result, we need the notion of *abelianization*.

Definition 14.9 Let G be a group. Its *commutator subgroup* is the subgroup $[G, G]$ generated by all elements of the form $[g, h] = ghg^{-1}h^{-1}$.

Proposition 14.10 *The commutator subgroup* $[G, G]$ *is a normal subgroup of* G.

Proof A typical element of $[G, G]$ has the form

$$a = [g_1, h_1][g_2, h_2] \cdots [g_k, h_k].$$

Let $g \in G$ be any element. Then

$$gag^{-1} = [gg_1g^{-1}, gh_1g^{-1}] \cdots [gg_kg^{-1}, gh_kg^{-1}],$$

which is an element of $[G, G]$. ■

Definition 14.11 The *abelianization* of a group G is the quotient group $G^{\text{ab}} = G/[G, G]$.

For any group G, G^{ab} is an abelian group. In fact, it is the *largest* abelian group that is a quotient of G. If we have a presentation for G, with some generators and relations, then we obtain the abelianization by adding the extra relations that force any pair of generators to commute, and no others.

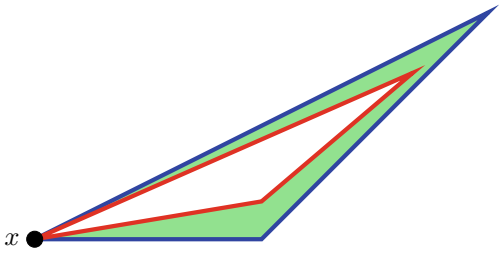


Figure 14.1 The red and blue cycles differ by a boundary, namely the boundary of the green 2-chain.

Example Let $G = \langle a_1, \ldots, a_n \mid \rangle$ be a free group on n generators. Then its abelianization is

$$G^{\text{ab}} = \langle a_1, \ldots, a_n \mid a_i a_j = a_j a_i \rangle \cong \mathbb{Z}^n$$

is the free *abelian* group on n generators.

Exercise 14.12 Show that, if $n \geq 2$, the abelianization of the symmetric group S_n is isomorphic to $\mathbb{Z}/2\mathbb{Z}$.

We can now state Hurewicz's Theorem in the case of π_1 and H_1. (Hurewicz more generally gives a relation between π_n and H_n in the case that $\pi_0, \pi_1, \ldots, \pi_{n-1}$ are all trivial.)

Theorem 14.13 (Hurewicz) *If X is path-connected and $x \in X$ is a basepoint, then $H_1(X) \cong \pi_1(X, x)^{\text{ab}}$.*

We will not give a complete proof of Hurewicz's Theorem (see [Hat02, Section 2.A] for a complete proof), but we can at least give an outline of how the argument might go. Suppose we have any loop γ in X based at x. Then we can break up γ into a bunch of 1-simplices whose sum is a 1-cycle in X. (This is more subtle than it may appear, because γ does not necessarily live in the 1-simplices of X. So, we may have to subdivide it first and perhaps homotope it slightly.) One can then check that two homotopic loops give rise to two 1-chains that only differ by a boundary, namely the boundary of the region between the two loops. See Figure 14.1 for a picture. As a result, we obtain a map, called the Hurewicz map, from $\pi_1(X, x)$ to $H_1(X)$. In fact, the Hurewicz map is a homomorphism.

Next, we must check that the Hurewicz map is surjective, and that its kernel is $[\pi_1(X, x), \pi_1(X, x)]$, the commutator subgroup of the fundamental group. This is believable: If we have a cycle in X, we can break it down as a bunch of loops, then we can add tails to each one of the loops so that they become based at x, as shown in Figure 14.2. Once we check the details, this shows that the Hurewicz map is surjective. The trickiest part is showing that the kernel is the commutator subgroup. Since $H_1(X)$ is abelian, we know that the kernel of the Hurewicz map is at least as large as the commutator subgroup. But one must show that it is no larger. And that is the part that we will skip.

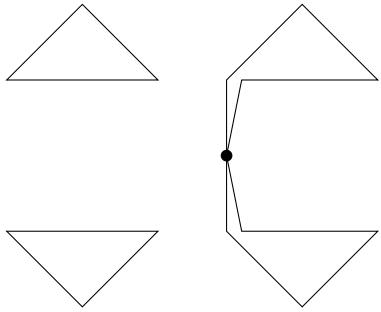


Figure 14.2 A cycle (left) and its corresponding loop (right).

14.6 Problems

(1) Suppose that $X = A \cup B$, where $A \cap B$ is contractible. Express the homology groups of X in terms of those of A and B.

(2) Let X be a topological space. Define its *suspension* SX to be $X \times [0, 1]/\sim$, where $(x, 0) \sim (y, 0)$ and $(x, 1) \sim (y, 1)$ for all $x, y \in S$. (That is, make a cylinder out of X, then collapse the top and bottom of this cylinder.)

   (a) Is the suspension $S\mathbb{S}^n$ of an n-sphere homeomorphic to a familiar space? If so, which one?
   (b) Compare the homology groups of X and SX.

(3) Show that $\delta : Z_n(A \cup B) \to H_{n-1}(A \cap B)$, constructed in the proof of the Mayer–Vietoris sequence, is a homomorphism.

(4) Show that if $w \in B_n(A \cup B)$, then $\delta(w) = 0$, where δ is as in the proof of the Mayer–Vietoris sequence.

(5) In the proof of the Mayer–Vietoris sequence, prove that the sequence is exact at $H_n(A \cap B)$ and $H_n(A \cup B)$.

(6) Modify our proof of the Jordan Curve Theorem for embeddings of $\mathbb{S}^1$ into $\mathbb{S}^2$, to the case of embeddings of $\mathbb{S}^1$ into $\mathbb{R}^2$.

(7) Let X be a connected sum of g tori and one projective plane. What is the homology of X?

Appendix A
Topological Notions

A.1 Compactness Results

We "defined" a compact set as one that is closed and bounded. While this is true in the case of subsets of $\mathbb{R}^n$ (and equivalent to the other definition), it is not the most general or most useful definition. A better one, as given in Problem 9 of Chapter 2, is as follows:

Definition A.1 A set X is *compact* if, whenever $\{U_\alpha\}_{\alpha\in A}$ is a collection of open subsets of X such that $\bigcup_{\alpha\in A} U_\alpha = X$, there exists a *finite* subset $B \subset A$ such that $\bigcup_{\beta\in B} U_\beta = X$.

That is, every open cover of X has a finite subcover.
With this definition, it is easy to prove important theorems about compactness.

Theorem A.2 *Let X be a compact set, and let $f : X \to Y$ be a continuous function. Then $f(X)$ is compact.*

That is, the image of a compact set under a continuous function is compact.

Proof Let $\{V_\alpha\}_{\alpha\in A}$ be any open cover of $f(X)$. We wish to show that there is a finite subcover. Let $U_\alpha = f^{-1}(V_\alpha) \subset X$. Because f is continuous, each U_α is open in X. As $\{V_\alpha\}$ is an open cover of $f(X)$, $\{U_\alpha\}$ is an open cover of X. Because X is compact, $\{U_\alpha\}$ must have a finite subcover, say $\{U_\beta\}_{\beta\in B}$, where B is a finite subset of A. But then $\{V_\beta\}_{\beta\in B}$ is an open cover of $f(X)$, so $\{V_\alpha\}_{\alpha\in A}$ has a finite subcover. Since the original choice of cover of $f(X)$ was arbitrary, it follows that *any* open cover of $f(X)$ has a finite subcover, so $f(X)$ is compact. ■

Corollary A.3 (Extreme Value Theorem) *If X is a nonempty compact set and $f : X \to \mathbb{R}$ is a continuous function, then f attains a maximum and a minimum value on X.*

C. Bray et al., *Algebraic Topology*,
https://doi.org/10.1007/978-3-030-70608-1

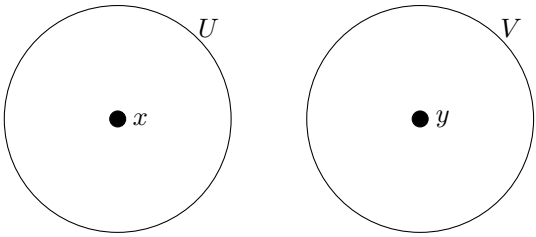


Figure A.1 This happens for any two points in a Hausdorff space.

Proof Under these hypotheses, $f(X)$ is a nonempty compact subset of $\mathbb{R}$, so let us show that a nonempty compact set of $\mathbb{R}$ must have a maximum (and, by symmetry, a minimum). Let $Y \subseteq \mathbb{R}$ be a nonempty compact set. Thus Y is closed and bounded. Let $s = \sup(Y)$, i.e. s is the smallest number t such that $t \geq y$ for all $y \in Y$. (It is a standard property of real numbers, called the *Least Upper Bound Property*, that such an s exists; see [Pug15, Theorem 2].) We claim that $s \in Y$, so that s is the maximum of Y. For each positive integer n, we can find some $y_n \in Y$ such that $y_n > s - \frac{1}{n}$. Thus $\lim_{n\to\infty} y_n \geq s$, and since each $y_n \leq s$, so have $\lim_{n\to\infty} y_n = s$. Since Y is closed and thus contains all its limit points, $s \in Y$, as claimed. ■

A.2 Technical Conditions for Abstract Surfaces

When we define an abstract surface, we want it to mimic all the properties of a surface in $\mathbb{R}^n$, except without the embedding. The most relevant point is the surface looks locally like $\mathbb{R}^2$. However, there are some more "global" conditions that hold automatically for any subset of $\mathbb{R}^n$, which we also expect abstract surfaces to have.

Definition A.4 An abstract topological space X is said to be *Hausdorff* if, for any two distinct points $x, y \in X$, there are open sets $U, V \subset X$ with $x \in U$ and $y \in V$, and $U \cap V = \varnothing$.

That is, we can find entirely separate open sets around U and V. See Figure A.1.

Subsets of $\mathbb{R}^n$ are automatically Hausdorff, since the Hausdorff property is inherited from the Hausdorffness of $\mathbb{R}^n$. So, it may seem hard to imagine what a non-Hausdorff space might look like. One famous example is called the *line with a doubled origin*. To construct it, take two lines, such as the lines $y = 0$ and $y = 1$. So, points on the first line are of the form $(x, 0)$, and points on the second line are of the form $(x, 1)$. Now, glue together the points $(x, 0)$ and $(x, 1)$, but do *not* glue $(0, 0)$ and $(0, 1)$. The resulting figure is a single line, except that it has two points at 0 rather than one.

In order to make this space into a genuine topological space, we use the quotient topology: let X be the union of the two lines, and let $\sim$ be the equivalence relation such that $(x, 0) \sim (x, 1)$ if $x \neq 0$. Then the line with the doubled origin is $X/\sim$, and this endows it with a topology. More concretely, if $x \neq 0$ is a point on the line with

the doubled origin, then a small neighborhood around x is just a neighborhood on the line, since there are small neighborhoods that do not see the two origins. On the other hand, a small neighborhood around one of the origins is just a neighborhood on that line, which misses the other origin.

So, this space is not Hausdorff because if U is a neighborhood of one of the origins and V is a neighborhood of the other, then $U \cap V \neq \varnothing$.

Another technical condition we expect out of our abstract surfaces is known as *second countability*.

Definition A.5 A space X is said to be *second countable* if we can find *countably many* open sets $U_1, U_2, \ldots \subseteq X$ so that every open set is a union of some (possibly infinite) collection of U_i's.

Example $\mathbb{R}$ is second countable, which follows from the *denseness of the rationals*. That is, consider all the open intervals in $\mathbb{R}$ that have *rational* endpoints. Every open set in $\mathbb{R}$ is a union of open intervals with rational endpoints. (Why?) Thus $\mathbb{R}$ is second countable.

More generally, $\mathbb{R}^n$, and indeed any subset of $\mathbb{R}^n$, is second countable.

What would a non-second countable space look like? The standard example of one is called the *long line*. We can think of a normal line as consisting of a half-open interval $[n, n+1)$ for each integer n, and then connecting the (missing) right endpoint of one interval to the left endpoint of the next interval. Or, up to homeomorphism, we can let n run only over the nonnegative integers if we remove the point at 0.

But we can go further, if we know about ordinals. For every ordinal $\alpha < \omega$, where ω is the first infinite ordinal (so that α is just a nonnegative integer), we take an interval $I_\alpha = [\alpha, \alpha + 1)$, and then we glue the ends of consecutive intervals together as before. That's the normal line (or ray).

To modify this to get the long ray or long line, we do the same thing, but now we let α run over the ordinals less than ω_1, the first *uncountable* ordinal. That gets us a long ray. To make the long line, just glue together two long rays at their endpoints.

Exercise A.6 Show that the long line is not second countable.

The line with the doubled origin and the long line are just two of the many peculiar topological spaces out there. For a compendium of curious and interesting topological spaces, see the book [SS95].

Appendix B
A Brief Look at Singular Homology

There are many ways of constructing homology. In Chapter 13, we introduced simplicial homology, based on splitting a topological space X up into simplices. This has the advantage that we can calculate homology in an algorithmic manner, but it has the disadvantage that many crucial theorems are difficult to prove, and indeed we skipped most of the proofs. For instance, if we were to triangulate X in some different way, why should the homology groups with respect to the two triangulations coincide? It is possible to prove this using simplicial homology, but doing so requires a considerable amount of work.

It would be appealing to have a version of homology that "clearly" does not depend on a choice of triangulation, or anything else beyond the space X itself. One of the ways of doing that is with singular homology.

Singular homology is constructed in a somewhat similar way to simplicial homology, in that we have a chain complex with chain groups, and the corresponding cycles, boundaries, and homology groups. The difference is in the definition of the chain groups. To define it, we start by introducing the standard n-simplex.

Definition B.1 The *standard n-simplex* Δ^n is defined to be

$$\Delta^n = \{(x_0, x_1, \ldots, x_n) \in \mathbb{R}^{n+1} : x_0 + \cdots + x_n = 1, x_i \geq 0 \text{ for all } i\}.$$

You can easily verify that Δ^0 is a point, Δ^1 is an edge, Δ^2 is a triangle, and so forth.

We now use the standard n-simplex to construct singular n-simplices in X.

Definition B.2 Let X be a topological space. A *singular n-simplex* in X is a continuous function $T : \Delta^n \to X$.

The singular n-simplices replace the n-simplices in the simplicial version of homology. Note that there are many, many singular n-simplices in X: uncountably many for typical spaces X. This can cause difficulties when trying to do calculations with them, but it's not a problem when we're proving general theorems about them.

The next step is to construct the chain groups.

C. Bray et al., *Algebraic Topology*,
https://doi.org/10.1007/978-3-030-70608-1

Definition B.3 The n^{th} *chain group* of X is the free abelian group generated by the singular n-simplices of X. We denote this group by $C_n(X)$.

What this means is that an element of $C_n(X)$ can be written as $c_1T_1 + c_2T_2 + \cdots + c_rT_r$ for some r, where the c_i's are integers and the T_i's are singular n-simplices. Note that this is necessarily a *finite* sum, even though there are infinitely many singular n-simplices. Alternatively, we can write an element of $C_n(X)$ as $\sum c_T T$, where the sum runs over the singular n-simplices T, and we require that $c_T = 0$ for all but finitely many singular n-simplices T.

As in the case of simplicial homology, the next step is to define the boundary map $\partial_n : C_n(X) \to C_{n-1}(X)$. For this, we need to define $n+1$ continuous maps from Δ_{n-1} to Δ_n. For $i = 0, \ldots, n$, define $\iota_i : \Delta_{n-1} \to \Delta_n$ by setting

$$\iota_i(x_0, \ldots, x_{n-1}) = (x_0, \ldots, x_{i-1}, 0, x_i, \ldots, x_{n-1}),$$

i.e. by inserting a 0 in the i^{th} coordinate. Intuitively, this means considering Δ_{n-1} as one of the $(n-1)$-dimensional faces of Δ_n.

Definition B.4 Let T be a singular n-simplex. We define $\partial_n(T)$ to be

$$\partial_n(T) = \sum_{i=0}^{n} (-1)^i T \circ \iota_i.$$

We then extend ∂_n linearly to a homomorphism from $C_n(X)$ to $C_{n-1}(X)$.

The boundary map has the same key property as does the one for simplicial homology, namely that $\partial_{n+1} \circ \partial_n = 0$ for all n. Thus, just as before, we may define the singular n-cycles as $Z_n(X) = \ker(\partial_n)$, and the singular n-boundaries as $B_n(X) = \text{im}(\partial_{n+1})$. Finally, the n^{th} singular homology group is $H_n(X) = Z_n(X)/B_n(X)$.

Note that the definition of singular homology doesn't depend on any auxiliary structure on X, in the way that simplicial homology at least appears to depend on the choice of triangulation. We can now easily say that homology is a homeomorphism invariant.

Proposition B.5 *If X and Y are homeomorphic, then $H_n(X) \cong H_n(Y)$ for all n.*

Proof Suppose that $f : X \to Y$ is a homeomorphism. We construct an isomorphism $f_* : C_n(X) \to C_n(Y)$ that descends to an isomorphism on homology. We define f_* on a singular n-simplex to be $f_*(T) = f \circ T$, which is a continuous function from Δ_n to Y and thus a singular n-simplex of Y. We then extend linearly to a homomorphism $f_* : C_n(X) \to C_n(Y)$. It is straightforward to check that f_* is in fact an isomorphism, and that it commutes with ∂_n in the sense that $\partial_n \circ f_* = f_* \circ \partial_n$. As we saw when constructing the induced homomorphism on simplicial homology, this is what we need for f_* to descend to a homomorphism (and in fact an isomorphism) $H_n(X) \to H_n(Y)$. ■

Simplicial and singular homology give the same answers for reasonable spaces, although proving that is beyond what we can do here; see [Mun84, Section 34] for a proof. What we can do is to show that there is a homomorphism from simplicial homology to singular homology. Again, we start by doing this at the level of chains. Let us write $C_n(X)^{\text{simp}}$ for simplicial chains (and similarly for cycles, boundaries, and homology), and $C_n(X)^{\text{sing}}$ for simplicial chains (etc.). Suppose we have a triangulation of X, and the n-simplices of this triangulation are $T_1, \ldots, T_r$. Then we can consider each T_i as being a singular simplex by choosing a way of mapping Δ_n to T_i, in such a way that the boundary of Δ_n gets mapped to the boundary of T_i. Let us call this function from ordinary simplices to singular simplices g. Then g extends to a homomorphism from $C_n(X)^{\text{simp}}$ to $C_n(X)^{\text{sing}}$ by setting

$$g\left(\sum_{i=1}^{r} c_i T_i\right) = \sum_{i=1}^{r} c_i g(T_i).$$

Since the boundary homomorphism commutes with g, g descends to a homomorphism on homology $g^\sharp : H_n(X)^{\text{simp}} \to H_n(X)^{\text{sing}}$. For nice spaces X, $g^\sharp$ is actually an isomorphism.

All the results we proved (or attempted to prove) using simplicial homology are also true with singular homology, and the proofs we tried to give turn into correct proofs. Recall in particular that we were only able to prove that induced homomorphisms exist if $f : X \to Y$ is a simplicial map. In singular homology, we are easily freed from such restrictions, since the image of a singular simplex in X is certainly a singular simplex in Y. Thus all these proofs go through as is, except that they are completely correct proofs. The Mayer–Vietoris sequence, too, holds for singular homology, with essentially no change to the proof.

Appendix C
Hints for Selected Problems

Chapter 1, Problem 9: If X and Y are two sets, then we usually prove that $X \subseteq Y$ by picking an element in X and showing that it's in Y. For the equality conditions, note that if $a, b \in A$ with $a \in U$, $b \notin U$, and $f(a) = f(b)$, then $a, b \in f^{-1}(f(U))$. Generalize from there.

Chapter 2, Problem 1: All the surjective functions exist. Imagine how you would construct a torus out of a clay sphere (or vice versa) in practice; you may wish to start by squashing it to something flat.

Chapter 2, Problem 7: For this problem, you must find a relation that is symmetric and transitive but not reflexive. Try to find the most trivial example you can.

Chapter 2, Problem 9: One approach is to start by showing that closed and bounded intervals in $\mathbb{R}$ satisfy the covering definition. Then show that if X and Y satisfy the covering definition, then so does $X \times Y$. Finally, show that if X satisfies the covering definition and $A \subseteq X$ is a closed set, then A does as well. All closed and bounded subsets in $\mathbb{R}^n$ are closed subsets of rectangular boxes.

Chapter 4, Problem 7: Show that in any graph on the surface of S, there is one vertex of degree $< \lfloor N \rfloor$. Thus if we can color the rest of the graph using $\lfloor N \rfloor$ colors such that no two adjacent vertices have the same color, then we can color that vertex differently from all its neighbors.

Chapter 4, Problem 8: If you're trying to find a homeomorphism, then just describe what it looks like. If you're trying to prove that no such homeomorphism exists, then you need to separate the two spaces by means of a homeomorphism invariant.

Chapter 6, Problem 4: Count the number of pairs (i, j) with $1 \leq i < j \leq n$ such that $\sigma(i) > \sigma(j)$.

Chapter 7, Problem 4c: Represent automorphisms of groups like $(\mathbb{Z}/n\mathbb{Z})^k$ or $\mathbb{Z}^k$ as matrices. What properties do these matrices have to have? When can you find non-commuting matrices of these types?

Chapter 8, Problem 6: Consider Figure C.1.

Chapter 9, Problem 3: This is easiest to picture if you work with an ID space.

C. Bray et al., *Algebraic Topology*,
https://doi.org/10.1007/978-3-030-70608-1

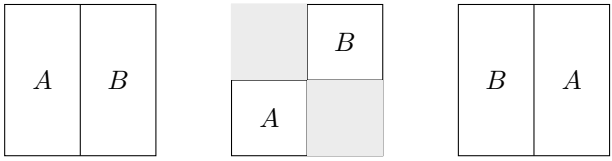


Figure C.1 An outline for a possible solution to Chapter 8, Problem 6.

Chapter 9, Problem 4: Show that a surface with genus g with $n > 0$ punctures is homotopy equivalent to a bunch of circles glued at one point. How many circles?

Chapter 9, Problem 6: The natural approach based on straight-line homotopies doesn't work. Instead, show that the concatenation $\gamma_0 * \bar{\gamma}_1$ is a loop in $\mathbb{S}^2$. Why must it be homotopic to the constant loop? The trickiest part is dealing with the case in which this loop is space-filling: it passes through all points on the sphere. (This can happen even if γ_0 and γ_1 are continuous.) In this case, show that you can homotope the path to one that isn't space-filling.

Chapter 9, Problem 8: There is a very nice continuous map $\mathbb{S}^2 \to \mathbb{RP}^2$.

Chapter 11, Problem 2b: If $x + \varepsilon e^x = 0$, then $-\varepsilon e^x = x$, so you need to show that $g(x) = -\varepsilon e^x$ has a fixed point.

Chapter 11, Problem 7: First, suppose that the jewels are continuous rather than discrete. How do you describe the number of jewels each person gets from a certain set of cuts by means of a point on $\mathbb{S}^n$? Finally, determine why the continuous case implies the discrete case.

Chapter 12, Problem 2: Find a simple homeomorphism invariant that distinguishes the earring from the wedge.

Chapter 12, Problem 7: There can be many different presentations for the same group.

Chapter 13, Problem 4: This is tedious to do with the material from Chapter 13. But going through the combinatorics here will make you appreciate the Mayer–Vietoris sequence even more when you get to it!

References

[AH77] Appel, K., Haken, W.: Every planar map is four colorable. I. Discharging. Ill. J. Math. **21**(3), 429–490 (1977). http://projecteuclid.org/euclid.ijm/1256049011

[AHK77] Appel, K., Haken, W., Koch, J.: Every planar map is four colorable. II. Reducibility. Ill. J. Math. **21**(3), 491–567 (1977). http://projecteuclid.org/euclid.ijm/1256049012

[AZ14] Aigner, M., Ziegler, G.M.: Proofs from the Book, 5th edn. Springer, Berlin (2014). Including illustrations by Karl H. Hofmann. https://doi.org/10.1007/978-3-662-44205-0

[Bae09] Baez, J.: This Week's Finds in Mathematical Physics (week 286) (2009). http://math.ucr.edu/home//baez/week286.html

[Bak09] Baker, K.: A (reverse) rational circle? Sketches of Topology (2009). https://sketchesoftopology.wordpress.com/2009/12/10/a-rational-circle/

[Bak10] Baker, K.: Bing's house. Sketches of Topology (2010). https://sketchesoftopology.wordpress.com/2010/03/25/bings-house/

[BEO02] Besche, H.U., Eick, B., O'Brien, E.: A millennium project: constructing small groups. Int. J. Algebra Comput. **12**(5), 623–644 (2002). https://doi.org/10.1142/S0218196702001115

[Bra21] Brahana, H.R.: Systems of circuits on two-dimensional manifolds. Ann. Math. (2) **23**(2), 144–168 (1921). https://doi.org/10.2307/1968030

[BT82] Bott, R., Tu, L.: Differential Forms in Algebraic Topology. Graduate Texts in Mathematics, vol. 82. Springer, New York (1982)

[Che04] Cheng, E.: Mathematics, morally (2004). http://cheng.staff.shef.ac.uk/morality/morality.pdf

[Cox94] Coxeter, H.S.M.: Projective Geometry. Springer, New York (1994). Revised reprint of the 2nd edn. (1974)

[DS84] Drobot, V., Sawka, J.: The teaching of mathematics: why the product topology? Am. Math. Mon. **91**(2), 137–138 (1984). https://doi.org/10.2307/2322114

[dS92] de Smit, B.: The fundamental group of the Hawaiian earring is not free. Int. J. Algebra Comput. **2**(1), 33–37 (1992). https://doi.org/10.1142/S0218196792000049

[EML45] Eilenberg, S., Lane, S.M.: General theory of natural equivalences. Trans. Am. Math. Soc. **58**, 231–294 (1945). https://doi.org/10.2307/1990284

[Fre82] Freedman, M.H.: The topology of four-dimensional manifolds. J. Differ. Geom. **17**(3), 357–453 (1982). http://projecteuclid.org/euclid.jdg/1214437136

[FT63] Feit, W., Thompson, J.G.: Solvability of groups of odd order. Pac. J. Math. **13**, 775–1029 (1963). http://projecteuclid.org/euclid.pjm/1103053943

[Gal79] Gale, D.: The game of Hex and the Brouwer fixed-point theorem. Am. Math. Mon. **86**(10), 818–827 (1979). https://doi.org/10.2307/2320146

C. Bray et al., *Algebraic Topology*,
https://doi.org/10.1007/978-3-030-70608-1

[Gal87] Gale, D.: The teaching of mathematics: the classification of 1-manifolds: a take-home exam. Am. Math. Mon. **94**(2), 170–175 (1987). https://doi.org/10.2307/2322421
[Gou97] Gouvêa, F.Q.: p-adic Numbers. Universitext, 2nd edn. Springer, Berlin (1997). An introduction. https://doi.org/10.1007/978-3-642-59058-0
[Gre02] Greene, J.E.: A new short proof of Kneser's conjecture. Am. Math. Mon. **109**(10), 918–920 (2002). https://doi.org/10.2307/3072460
[Hat02] Hatcher, A.: Algebraic Topology. Cambridge University Press, Cambridge (2002)
[HS09] Herrlich, F., Schmithüsen, G.: Dessins d'enfants and origami curves. Handbook of Teichmüller Theory. Vol. II. IRMA Lectures in Mathematics and Theoretical Physics, vol. 13, pp. 767–809. Eur. Math. Soc. Zürich (2009). https://doi.org/10.4171/055-1/19
[Man16] Manolescu, C.: Pin(2)-equivariant Seiberg–Witten Floer homology and the triangulation conjecture. J. Am. Math. Soc. **29**(1), 147–176 (2016). https://doi.org/10.1090/jams829
[Mar60] Markov, A.A.: Insolubility of the problem of homeomorphy. In: Proceedings of the International Congress of Mathematicians, 1958, pp. 300–306. Cambridge University Press, New York (1960)
[Mas91] Massey, W.S.: A Basic Course in Algebraic Topology. Graduate Texts in Mathematics, vol. 127. Springer, New York (1991)
[Mat03] Matoušek, J.: Using the Borsuk–Ulam Theorem. Universitext. Springer, Berlin (2003). Lectures on Topological Methods in Combinatorics and Geometry, Written in cooperation with Anders Björner and Günter M, Ziegler
[Mir95] Miranda, R.: Algebraic Curves and Riemann Surfaces. Graduate Studies in Mathematics, vol. 5. American Mathematical Society, Providence (1995). https://doi.org/10.1090/gsm/005
[Möb61] Möbius, A.F.: Zur theorie der polyëder und der elementarverwandtschaft. Oeuvres Complètes **2**, 519–559 (1861)
[Mor12] Morishita, M.: Knots and Primes. Universitext. Springer, London (2012). An introduction to arithmetic topology. https://doi.org/10.1007/978-1-4471-2158-9
[Mun84] Munkres, J.R.: Elements of Algebraic Topology. Addison–Wesley Publishing Company, Menlo Parkd (1984)
[Nas51] Nash, J.: Non-cooperative games. Ann. Math. (2) **54**, 286–295 (1951). https://doi.org/10.2307/1969529
[Per02] Perelman, G.: The entropy formula for the Ricci flow and its geometric applications. ArXiv Mathematics e-prints (2002). arXiv:math/0211159
[Per03a] Perelman, G.: Finite extinction time for the solutions to the Ricci flow on certain three-manifolds. ArXiv Mathematics e-prints (2003). arXiv:math/0307245
[Per03b] Perelman, G.: Ricci flow with surgery on three-manifolds. ArXiv Mathematics e-prints (2003). arXiv:math/0303109
[Pug15] Pugh, C.C.: Real Mathematical Analysis. Undergraduate Texts in Mathematics, 2nd edn. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-17771-7
[Ric63] Richards, I.: On the classification of noncompact surfaces. Trans. Am. Math. Soc. **106**, 259–269 (1963). https://doi.org/10.2307/1993768
[Rot95] Rotman, J.: An Introduction to the Theory of Groups. Graduate Texts in Mathematics, vol. 148, 4th edn. Springer, New York (1995). https://doi.org/10.1007/978-1-4612-4176-8
[RR11] Ross, F., Ross, W.: The Jordan curve theorem is non-trivial. J. Math. Arts **5**(4), 213–219 (2011). https://doi.org/10.1080/17513472.2011.634320
[Ser03] Serre, J.-P.: Trees. Springer Monographs in Mathematics. Springer, Berlin (2003). Translated from the French original by John Stillwell, Corrected 2nd printing of the 1980 English translation
[SS95] Steen, L.A., Seebach Jr., J.A.: Counterexamples in Topology. Dover Publications, Inc., Mineola (1995). Reprint of the 2nd edn. (1978)
[Su99] Su, F.E.: Rental harmony: Sperner's lemma in fair division. Am. Math. Mon. **106**(10), 930–942 (1999). https://doi.org/10.2307/2589747

[Tho92] Thomassen, C.: The Jordan–Schönflies theorem and the classification of surfaces. Am. Math. Mon. **99**(2), 116–130 (1992). https://doi.org/10.2307/2324180

[Wil05] Wild, M.: The groups of order sixteen made easy. Am. Math. Mon. **112**(1), 20–31 (2005). https://doi.org/10.2307/30037381

[Wil09] Wilson, R.A.: The Finite Simple Groups. Graduate Texts in Mathematics, vol. 251. Springer, London (2009). https://doi.org/10.1007/978-1-84800-988-2

Index

C. Bray et al., *Algebraic Topology*,
https://doi.org/10.1007/978-3-030-70608-1