Graduate Texts in Mathematics

John M. Lee

Introduction to Topological Manifolds

Springer

Graduate Texts in Mathematics 202

Springer
New York
Berlin
Heidelberg
Barcelona
Hong Kong
London
Milan
Paris
Singapore
Tokyo

Graduate Texts in Mathematics

(continued after index)

John M. Lee

Introduction to Topological Manifolds

With 138 Illustrations

Springer

John M. Lee
Department of Mathematics
University of Washington
Seattle, WA 98195-4350
USA
lee@math.washington.edu

Mathematics Subject Classification (2000): 54-01, 55-01, 57-01, 57N65

Library of Congress Cataloging-in-Publication Data
Lee, John M., 1950–
Introduction to topological manifolds / John M. Lee.
p. cm. — (Graduate texts in mathematics ; 202)
Includes bibliographical references and index.
ISBN 0-387-98759-2 (hard cover : alk. paper) — ISBN 0-387-95026-5 (soft cover : alk. paper)
1. Topological manifolds. I. Title. II. Series.
QA613.2 .L44 2000
514′.3—dc21 00-026156

ISBN 0-387-98759-2 Springer-Verlag New York Berlin Heidelberg SPIN 10711695(hardcover)
ISBN 0-387-95026-5 Springer-Verlag New York Berlin Heidelberg SPIN 10763406 (softcover)

To Pm,
sine qua non

Preface

This book is an introduction to manifolds at the beginning graduate level. It contains the essential topological ideas that are needed for the further study of manifolds, particularly in the context of differential geometry, algebraic topology, and related fields. Its guiding philosophy is to develop these ideas rigorously but economically, with minimal prerequisites and plenty of geometric intuition. Here at the University of Washington, for example, this text is used for the first third of a year-long course on the geometry and topology of manifolds; the remaining two-thirds focuses on smooth manifolds.

There are many superb texts on general and algebraic topology available. Why add another one to the catalog? The answer lies in my particular vision of graduate education—it is my (admittedly biased) belief that every serious student of mathematics needs to know manifolds intimately, in the same way that most students come to know the integers, the real numbers, Euclidean spaces, groups, rings, and fields. Manifolds play a role in nearly every major branch of mathematics (as I illustrate in Chapter 1), and specialists in many fields find themselves using concepts and terminology from topology and manifold theory on a daily basis. Manifolds are thus part of the basic vocabulary of mathematics, and need to be part of the basic graduate education. The first steps must be topological, and are embodied in this book; in most cases, they should be complemented by material on smooth manifolds, vector fields, differential forms, and the like. (After all, few of the really interesting applications of manifold theory are possible without using tools from calculus.)

Of course, it is not realistic to expect all graduate students to take full-year courses in general topology, algebraic topology, and differential geometry. Thus, although this book touches on a generous portion of the material that is typically included in much longer courses, the coverage is selective and relatively concise, so that most of the book can be covered in a single quarter or semester, leaving time in a year-long course for further study in whatever direction best suits the instructor and the students. At U.W. we follow it with a two-quarter sequence on smooth manifold theory; but it could equally well lead into a full-blown course on algebraic topology.

It is easy to describe what this book is not. It is not a course on general topology—many of the topics that are standard in such a course are ignored here, such as metrization theorems; infinite products and the Tychonoff theorem; countability and separation axioms and the relationships among them (other than second countability and the Hausdorff axiom, which are part of the definition of manifolds); and function spaces. Nor is it a course in algebraic topology—although I treat the fundamental group in detail, there is barely a mention of the higher homotopy groups, and the treatment of homology theory is extremely brief, meant mainly to give the flavor of the theory and to lay some groundwork for the later introduction of de Rham cohomology. It is certainly not a comprehensive course on topological manifolds, which would have to include such topics as PL structures and maps, transversality, intersection theory, cobordism, bundles, characteristic classes, and low-dimensional geometric topology. Finally, it is not intended as a reference book, because few of the results are presented in their most general or most complete form.

Perhaps the best way to summarize what this book is would be to say that it represents, to a good approximation, my conception of the ideal amount of topological knowledge that should be possessed by beginning graduate students who are planning to go on to study smooth manifolds and differential geometry. Experienced mathematicians will probably observe that my choices of material and approach have been influenced by the fact that I am a differential geometer and analyst by training and predilection, not a topologist. Thus I give special emphasis to topics that will be of importance later in the study of smooth manifolds, such as group actions, orientations, and degree theory. (A few topological ideas that are important for manifold theory, such as paracompactness and embedding theorems, are omitted because they are better treated in the context of smooth manifolds.) But despite my prejudices, I have tried to make the book useful as a precursor to algebraic topology courses as well, and it could easily serve as a prerequisite to a more extensive course in homology and homotopy theory.

Prerequisites. The prerequisite for studying this book is, briefly stated, a solid undergraduate degree in mathematics; but this probably deserves some elaboration. Traditionally, "algebraic topology" has been seen as a separate subject from "general topology," and most courses in the former begin with the assumption that the students have already completed a course in the latter. However, the sad fact is that for a variety of reasons, many undergraduate mathematics majors in the U.S. never take a course in general topology. For that reason I have written this book without assuming that the reader has had any exposure to topological spaces. On the other hand, I do assume several essential prerequisites beyond calculus and linear algebra: basic logic and set theory such as what one would encounter in any rigorous undergraduate analysis or algebra course; real analysis at the level of Rudin's *Principles of Mathematical Analysis* [Rud76], including, in particular, a thorough understanding of metric spaces and their continuous functions and compact subsets; and group theory at the level of Hungerford's *Abstract Algebra: An Introduction* [Hun90] or Herstein's *Topics in Algebra* [Her75]. Because it is vitally important that the reader be comfortable with this prerequisite material, I have collected in the Appendix a summary of the main points that are used throughout the book, together with a representative collection of exercises. These exercises, which should be relatively straightforward for anyone who has had the prerequisite courses, can be used by the student to refresh his or her knowledge, or can be assigned by the instructor at the beginning of the course to make sure that everyone starts with the same background.

Organization. The book is divided into thirteen chapters, which can be grouped into an introduction and five major substantive sections.

The introduction (Chapter 1) is meant to whet the student's appetite and create a "big picture" into which the many details can later fit.

The first major section, Chapters 2 through 4, is a brief and highly selective introduction to the ideas of general topology: topological spaces; their subspaces, products, and quotients; and connectedness and compactness. Of course, manifolds are the main examples and are emphasized throughout. These chapters emphasize the ways in which topological spaces differ from the more familiar Euclidean and metric spaces, and carefully develop the machinery that will be needed later, such as quotient maps, local path connectedness, and locally compact Hausdorff spaces.

The second major section, comprising Chapters 5 and 6, explores in detail the main examples that motivate the rest of the theory: simplicial complexes, 1-manifolds, and 2-manifolds. Chapter 5 introduces simplicial complexes in two ways—first concretely, as locally finite collections of simplices in Euclidean space that intersect nicely; and then abstractly, as collections of finite vertex sets. Both approaches are useful: The concrete definition helps students develop their geometric intuition, while the abstract point of view emphasizes the fact that all statements about simplicial complexes can be reduced to combinatorics. There are several reasons for introducing simplicial complexes at this stage: They furnish a rich source of examples; they give a very concrete way of thinking about orientations and the Euler characteristic; they provide the concept of triangulability needed for the classifications of 1-manifolds and 2-manifolds; and they set the stage for the treatment of homology later. Chapter 6 begins by proving a classification theorem for 1-manifolds using the triangulability theorem proved in the preceding chapter. The rest of the chapter is devoted to a detailed study of 2-manifolds. After exploring the basic examples of surfaces—the sphere, the torus, the projective plane, and their connected sums—I give a complete proof of the classification theorem for compact surfaces, essentially following the treatment in [Mas89].

The third major section, Chapters 7 through 10, is the core of the book. In it, I give a fairly complete and traditional treatment of the fundamental group. Chapter 7 introduces the definitions and proves the topological and homotopy invariance of the fundamental group. At the end of the chapter I insert a brief introduction to category theory. Categories are not used in a central way anywhere in the book, but it is natural to introduce them after having proved the topological invariance of the fundamental group, and it is useful for students to begin thinking in categorical terms early. Chapter 8 gives a detailed proof that the fundamental group of the circle is infinite cyclic. Because the techniques used here are the precursor and motivation for the entire theory of covering spaces, I introduce some of the terminology of the latter subject—evenly covered neighborhoods, local sections, lifting—in the special case of the circle, and the proofs here form a model for the proofs of more general theorems involving covering spaces to come in a later chapter. Chapter 9 is a brief digression into group theory. Although a basic acquaintance with group theory is an essential prerequisite, most undergraduate algebra courses do not treat free products, free groups, presentations of groups, or free abelian groups, so I develop these subjects from scratch. (The material on free abelian groups is included primarily for use in the treatment of homology in Chapter 13, but some of the results play a role also in classifying the coverings of the torus in Chapter 12.) The last chapter of this section gives the statement and proof of the Seifert–Van Kampen theorem, which expresses the fundamental group of a space in terms of the fundamental groups of its subsets, and describes several applications of the theorem including computation of the fundamental groups of graphs and of all the compact surfaces.

The fourth major section consists of two chapters on covering spaces. Chapter 11 defines covering spaces, gives a few examples, and develops the theory of the covering group. Much of the development goes rapidly here, because it is parallel to what was done earlier in the concrete case of the circle. The ostensible goal of Chapter 12 is to prove the classification theorem for coverings—that there is a one-to-one correspondence between isomorphism classes of coverings of X and conjugacy classes of subgroups of the fundamental group of X—but along the way two other ideas are developed that are of central importance in their own right. The first is the notion of the universal covering space, together with proofs that every manifold has a universal covering and that the universal covering space covers every other covering space. The second is the fact that the quotient of a manifold by a free, proper action of a discrete group yields a manifold. These ideas are applied to a number of important examples, including classifying coverings of the torus and the lens spaces, and proving that surfaces of higher genus are covered by the hyperbolic disk.

The fifth major section of the book consists of one chapter only, Chapter 13, on homology theory. In order to cover some of the most important applications of homology to manifolds in a reasonable time, I have chosen a "low-tech" approach to the subject. I focus mainly on singular homology because it is the most straightforward generalization of the fundamental group. After defining the homology groups, I prove a few essential properties, including homotopy invariance and the Mayer–Vietoris theorem, with a minimum of homological machinery. I could not resist including a (terribly brief) introduction to simplicial homology, just because it immediately yields the topological invariance of the Euler characteristic. The last section of the chapter is a brief introduction to cohomology, mainly with field coefficients, to serve as background for a treatment of de Rham theory in a later course. In keeping with the overall philosophy of focusing only on what is necessary for a basic understanding of manifolds, I do not even mention relative homology, homology with arbitrary coefficients, simplicial approximation, or the axioms for a homology theory.

Although this book grew out of notes designed for a one-quarter graduate course, there is clearly too much material here to cover adequately in ten weeks. It should be possible to cover all or most of it in a semester with well prepared students. The book could even be used for a full-year course, allowing the instructor to adopt a much more leisurely pace and to work out some of the problems as examples in class.

Each instructor will have his or her own ideas about what to leave out in order to fit the material into a short course. At the University of Washington, we typically do not cover the chapter on homology at all, and give short shrift to some of the simplicial theory and some of the more involved examples of covering maps. Others may wish to leave out some or all of the material on covering spaces, or the classification of surfaces. With students who have had a solid topology course, the first four chapters could be skipped or assigned as outside reading.

Exercises and Problems. As is the case with any new mathematical material, and perhaps even more than usual with material like this that is so different from the mathematics most students have seen as undergraduates, it is impossible to learn the subject without getting one's hands dirty and working out a large number of examples and problems. I have tried to give the reader ample opportunity to do so throughout the book. In every chapter, and especially in the early ones, there are "exercises" woven into the text. Do not ignore them; without their solutions, the text is incomplete.

The reader should take each exercise as a signal to stop reading, pull out a pencil and paper, and work out the answer before proceeding further. The exercises are usually relatively easy, and typically involve proving minor results or working out examples that are essential to the flow of the exposition. Some require techniques that the student probably already knows from prior courses; others ask the student to practice techniques or apply results that have recently been introduced in the text. A few are straightforward but rather long arguments that are more enlightening to work through on one's own than to read. In the later chapters, fewer things are singled out as exercises, but there are still plenty of omitted details in the text that the student should work out before going on; it is my hope that by the time the student reaches the last few chapters he or she will have developed the habit of stopping and working through most of the details that are not spelled out without having to be told.

At the end of each chapter is a selection of "problems." These are, with a few exceptions, harder and/or longer than the exercises, and give the student a chance to grapple with more significant issues. The results of a number of the problems are used later in the text. There are more problems than most students could do in a quarter or a semester, so the instructor will want to decide which ones are most germane and assign those as homework.

Acknowledgments. Those of my colleagues at the University of Washington with whom I have discussed this material—Tom Duchamp, Judith Arms, Steve Mitchell, Scott Osborne, and Ethan Devinatz—have provided invaluable help in sorting out what should go into this book and how it should be presented. Each has had a strong influence on the way the book has come out, for which I am deeply grateful. (On the other hand, it is likely that none of them would wholeheartedly endorse all my choices regarding which topics to treat and how to treat them, so they are not to be blamed for any awkwardnesses that remain.) I would like to thank Ethan Devinatz in particular for having had the courage to use the book as a course text when it was still in an inchoate state, and for having the grace and patience to wait while I prepared chapters at the last minute for his course.

Thanks are due also to Mary Sheetz, who did an excellent job producing some of the illustrations under the pressures of time and a finicky author.

My debt to the authors of several other textbooks will be obvious to anyone who knows those books: William Massey's *Algebraic Topology: An Introduction* [Mas89], Allan Sieradski's *An Introduction to Topology and Homotopy* [Sie92], Glen Bredon's *Topology and Geometry*, and James Munkres's *Topology: A First Course* [Mun75] and *Elements of Algebraic Topology* [Mun84] are foremost among them.

Finally, I would like to thank my wife, Pm, for her forbearance and unflagging support while I was spending far too much time with this book

and far too little with the family; without her help I unquestionably could not have done it.

John M. Lee Seattle

Contents

1

Introduction

A course on manifolds differs from most other introductory graduate mathematics courses in that the subject matter is often completely unfamiliar. Most beginning graduate students have had undergraduate courses in algebra and analysis, so that graduate courses in those areas are continuations of subjects they have already begun to study. But it is possible to get through an entire undergraduate mathematics education, at least in the United States, without ever hearing the word "manifold."

One reason for this anomaly is that even the definition of manifolds involves rather a large number of technical details—for example, in this book the formal definition will not come until the end of Chapter 2. Since it is disconcerting to embark on such an adventure without even knowing what it is about, we devote this introductory chapter to a nonrigorous definition of manifolds, an informal exploration of some examples, and a consideration of where and why they arise in various branches of mathematics.

What Are Manifolds?

Let us begin by describing informally how one should think about manifolds. The underlying idea is that manifolds are like curves and surfaces, except, perhaps, that they might be of higher dimension. Every manifold has a *dimension*, which is, roughly speaking, the number of independent numbers (or "parameters") needed to specify a point. The prototype of

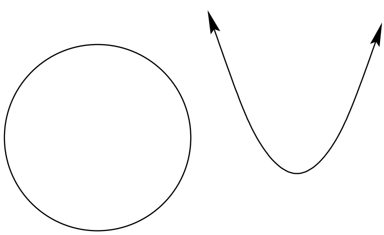


FIGURE 1.1. Plane curves.

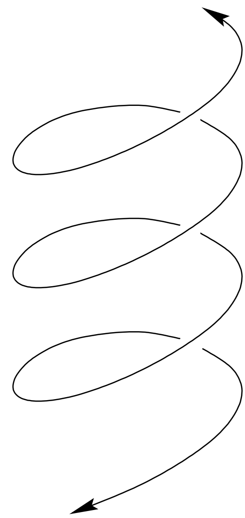


FIGURE 1.2. Space curve.

an n-dimensional manifold is n-dimensional Euclidean space $\mathbb{R}^n$, in which each point *is* an n-tuple of real numbers.

An n-dimensional manifold is an object modeled *locally* on $\mathbb{R}^n$; this means that it takes exactly n numbers to specify a point, at least if we do not stray too far from a given starting point. A physicist would say that an n-dimensional manifold is an object with n "degrees of freedom."

Manifolds of dimension 1 are commonly called *curves* (although they need not be "curved" in the ordinary sense of the word). The simplest example is the real line; other examples are provided by familiar plane curves such as circles, parabolas, or the graph of any continuous function of the form $y = f(x)$ (Figure 1.1). Still other familiar 1-dimensional manifolds are space curves, which are often described parametrically by equations such as $(x, y, z) = (f(t), g(t), h(t))$ for some continuous functions f, g, h (Figure 1.2).

In each of these examples, a point on the curve can be unambiguously specified by a single real number. For example, a point on the real line *is* a real number. We might specify a point on the circle by its angle, a point on a graph by its x coordinate, and a point on a parametrized curve by its parameter t. Note that although a parameter value determines a point, different parameter values may correspond to the same point, as in the case of angles on the circle. But in every case, as long as we stay close to some initial point, there is a one-to-one correspondence between nearby real numbers and nearby points on the curve.

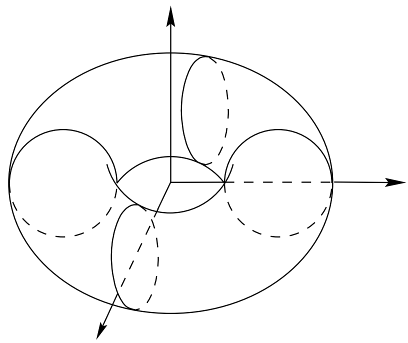


FIGURE 1.3. Doughnut surface.

Manifolds of dimension 2 are *surfaces.* The two most common examples are planes and spheres. (When mathematicians speak of a sphere, we invariably mean a spherical *surface*, which is 2-dimensional, not a solid ball, which is 3-dimensional.) Other familiar surfaces include cylinders, ellipsoids, paraboloids, and the doughnut-shaped surface in $\mathbb{R}^3$ obtained by revolving a circle around the z-axis (Figure 1.3). (This doughnut-shaped surface is often called a *torus*, but we will reserve that name for a slightly different but closely related object, to be introduced in the next chapter.)

In these cases two coordinates are needed to determine a point. For example, on the plane we typically use Cartesian or polar coordinates; on the sphere we might use latitude and longitude; while on the doughnut surface we might use two angles. As in the 1-dimensional case, the correspondence between points and pairs of numbers is in general only local.

The only higher-dimensional manifold that we can visualize is Euclidean 3-space. But it is not hard to construct subsets of higher-dimensional Euclidean spaces that might reasonably be called manifolds. First, any open subset of $\mathbb{R}^n$ is an n-manifold for obvious reasons. More interesting examples are obtained by using one or more equations to "cut out" lower-dimensional subsets. For example, the set of points (x_1, x_2, x_3, x_4) in $\mathbb{R}^4$ satisfying the equation

$$(x_1)^2 + (x_2)^2 + (x_3)^2 + (x_4)^2 = 1 \tag{1.1}$$

is called the (*unit*) 3-*sphere*. It is a 3-dimensional manifold because in a neighborhood of any given point it takes exactly three coordinates to specify a nearby point: Starting at, say, the "north pole" $(0,0,0,1)$, we can solve equation (1.1) for x_4, and then each nearby point is uniquely determined by choosing appropriate (small) (x_1, x_2, x_3) coordinates and setting $x_4 = (1 - (x_1)^2 - (x_2)^2 - (x_3)^2)^{1/2}$. Near other points, we may need to solve for different variables; but in each case three coordinates suffice.

The key feature of these examples is that an n-dimensional manifold "looks like" $\mathbb{R}^n$ locally. To make sense of the intuitive notion of "looks like," we will say that two subsets of Euclidean spaces $U \subset \mathbb{R}^k$, $V \subset \mathbb{R}^n$ are *topologically equivalent* or *homeomorphic* (Greek for "same form") if there exists a one-to-one correspondence $\varphi\colon U \to V$ such that both φ and its inverse are continuous maps. (Such a correspondence is called a *homeomorphism.*) A subset M of some Euclidean space $\mathbb{R}^k$ is *locally Euclidean of dimension n* if every point of M has a neighborhood in M that is topologically equivalent to a ball in $\mathbb{R}^n$.

Now we can give a preliminary definition of manifolds. An *n-dimensional manifold* (*n-manifold* for short) is a subset of some Euclidean space $\mathbb{R}^k$ that is locally Euclidean of dimension n. Later, after we have developed more machinery, we will give a considerably more general definition; but this one will get us started.

Why Study Manifolds?

What follows is an incomplete survey of some of the fields of mathematics in which manifolds play an important role.

Topology

Roughly speaking, topology is the branch of mathematics that is concerned with properties of sets that are unchanged by "continuous deformations." More accurately, a topological property is one that is preserved by homeomorphisms.

The subject in its modern form was invented a century ago by the French mathematician Henri Poincaré, as an outgrowth of his attempts to classify geometric objects that appear in analysis. In a seminal 1895 paper titled *Analysis Situs* (the old name for topology, Latin for "analysis of position") and a series of companion papers in 1899–1905, Poincaré laid out the main problems of topology and introduced an astonishing array of new ideas for solving them. As you read this book, you will see that his name is written all over the subject. In the intervening century, topology has taken on the role of providing the foundations for just about every branch of mathematics that has any use for a concept of "space." (An excellent historical account of the first six decades of the subject can be found in [Die89].)

Here is a simple but telling example of the kind of problem that topology was invented to solve. Consider two surfaces in space: a sphere and a cube. It should not be hard to convince yourself that the cube can be continuously deformed into the sphere without tearing or collapsing it. It is not much harder to come up with an explicit formula for a homeomorphism between them (we will do so in Chapter 2). Similarly, with a little more work, you

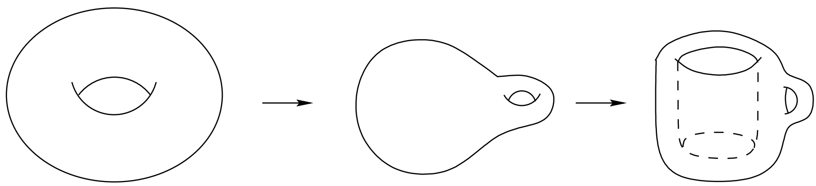


FIGURE 1.4. Deforming a doughnut into a coffee cup.

should be able to see how a doughnut surface can be continuously deformed into the surface of a one-handled coffee cup, by stretching out one-half of the doughnut to become the cup, and shrinking the other half to become the handle (Figure 1.4). Once you decide on an explicit set of equations to define a "coffee-cup surface" in $\mathbb{R}^3$, you could in principle come up with a set of formulas to describe a homeomorphism between it and the doughnut surface. On the other hand, a little reflection will probably convince you that there is *no* homeomorphism from the sphere to the doughnut surface: Any such map would have to tear open a "hole" in the sphere, and thus could not be continuous.

It is usually relatively straightforward (though not always easy!) to prove that two manifolds are topologically equivalent once you have convinced yourself intuitively that they are: Just write down an explicit homeomorphism between them. What is much harder is to prove that two manifolds are *not* homeomorphic—even when it seems "obvious" that they are not as in the case of the sphere and the doughnut—because you would need to show that no one, no matter how clever, could find such a map.

History abounds with examples of operations that mathematicians long believed to be impossible, only to be proved wrong. Here is an example from topology. Imagine a spherical surface colored white on the outside and gray on the inside, and imagine that it can move freely in space, including passing freely through itself. Under these conditions you could turn the sphere inside out by continuously deforming it, so that the gray side ends up facing out, but it seems obvious that in so doing you would have to introduce a crease somewhere. (There are precise mathematical definitions of the terms "continuously deforming" and "creases," but you do not need to know them to get the general idea.) The simplest way to proceed would be to push the northern hemisphere down and the southern hemisphere up, allowing them to pass through each other, until the two hemispheres had switched places (Figure 1.5); but this would introduce a crease along the equator. The topologist Stephen Smale stunned the mathematical community in 1958 [Sma58] when he proved it was possible to turn the sphere inside out without introducing any creases. Several ways to do this are beautifully illustrated in video recordings [Max77, LMM94, SFL98].

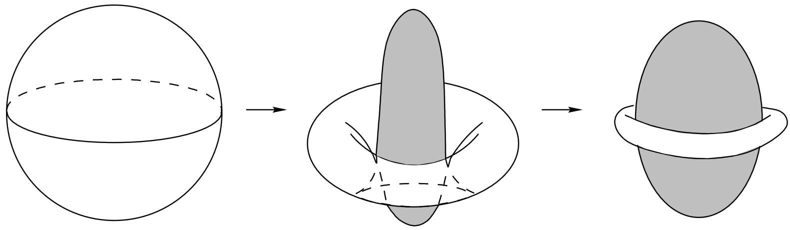


FIGURE 1.5. Turning a sphere inside out (with a crease).

The usual way to prove that two manifolds are *not* topologically equivalent is by finding *topological invariants*: properties (which could be numbers or other mathematical objects such as groups, matrices, polynomials, or vector spaces) that are preserved by homeomorphisms. If two manifolds have different invariants, they cannot be homeomorphic.

It is evident from the examples above that geometric properties such as circumference and area are not topological invariants, because they are not generally preserved by homeomorphisms. Intuitively, the property that distinguishes the sphere from the doughnut surface is the fact that the latter has a "hole," while the former does not. But it turns out that giving a precise definition of what is meant by a hole takes rather a lot of work. One invariant that is commonly used to count holes in a manifold is called the *fundamental group* of the manifold, which is a group (in the algebraic sense) attached to each manifold in such a way that homeomorphic manifolds have isomorphic groups. Then the "size" of the fundamental group is a measure of the number of holes possessed by the manifold. The study of the fundamental group will occupy a major portion of this book. It is the starting point for *algebraic topology*, which is the subject that studies topological properties of manifolds (or other geometric objects) by attaching algebraic structures such as groups and rings to them in a topologically invariant way.

One of the most important problems of topology is the problem of classifying manifolds. Ideally, one would like to produce a list of n-dimensional manifolds, and a theorem that says every n-dimensional manifold is homeomorphic to exactly one on the list, together with a list of computable topological invariants that could be used to decide where on the list any given manifold belongs. Precisely such a theorem is known for surfaces: It says that every compact surface is homeomorphic to a sphere, or to a doughnut surface with a finite number of holes, or to a connected sum of projective planes. (We will define these terms and prove the theorem in Chapter 6.)

For higher-dimensional manifolds, the situation is much more complicated. For example, Poincaré conjectured around 1900 that any compact 3-

manifold whose fundamental group is the trivial (one-element) group must be homeomorphic to the 3-sphere. For a long time, topologists thought of this as the simplest first step in a potential classification of 3-manifolds. But although analogous conjectures have been made for higher-dimensional manifolds and were proved in the intervening years (for 5-manifolds and higher by Stephen Smale in 1961 [Sma61], and for 4-manifolds by Michael Freedman in 1982 [Fre82]), the original Poincaré conjecture remains as of this writing a preeminent unsolved problem in topology. The best hope for a classification of 3-manifolds is the *geometrization conjecture* made in the 1970s by William Thurston (see [Sco83, Thu97] for an explanation), which says, roughly, that every compact 3-manifold can be cut into finitely many pieces each of which admits one of eight (mostly non-Euclidean) geometric structures. Since the manifolds with geometric structures are much better understood, a proof of this conjecture would go a long way toward providing a complete classification of 3-manifolds; in particular, it would imply that the Poincaré conjecture is true.

In dimensions 4 and higher, on the other hand, there is no hope for a complete classification: It was proved in 1958 by A. A. Markov that there is no algorithm for classifying manifolds of dimension greater than 3 (see [Sti93]). Nonetheless, there is much that can be said using sophisticated combinations of techniques from algebraic topology, differential geometry, partial differential equations, and algebraic geometry, and spectacular progress was made in the last half of the twentieth century in understanding the variety of manifolds that exist. The topology of 4-manifolds, in particular, is currently a highly active field of research.

Geometry

The principal objects of study in Euclidean plane geometry, as you encountered it in secondary school, are figures constructed from portions of lines, circles, and other curves—in other words, 1-manifolds. Similarly, solid geometry is concerned with figures made from portions of planes, spheres, and other 2-manifolds. The properties that are of interest are those that are invariant under rigid motions. These include simple properties such as lengths, angles, areas, and volumes, as well as more sophisticated properties derived from them such as curvature. The curvature of a curve or surface is a quantitative measure of how it bends and in what directions; for example, a positively curved surface is "bowl-shaped," while a negatively curved one is "saddle-shaped."

Geometric theorems involving curves and surfaces range from the trivial to the very deep. A typical theorem you have undoubtedly seen before is the *angle-sum theorem*: The sum of the interior angles of any Euclidean triangle is π radians. This seemingly trivial result has profound generalizations to the study of curved surfaces, where angles may add up to more or less than π depending on the curvature of the surface. The high point of surface theory is the *Gauss–Bonnet theorem*: For a closed, bounded surface in $\mathbb{R}^3$, this theorem expresses the relationship between the total curvature (i.e., the integral of curvature with respect to area) and the number of holes it has. If the surface is topologically equivalent to an n-holed doughnut surface, the theorem says that the total curvature is exactly equal to $4\pi - 4\pi n$. In the case $n = 1$ this implies that no matter how a one-holed doughnut surface is bent or stretched, the regions of positive and negative curvature will always precisely cancel each other out so that the total curvature is zero.

The introduction of manifolds has allowed the study of geometry to be carried into higher dimensions. The appropriate setting for studying geometric properties in arbitrary dimensions is that of *Riemannian manifolds*, which are manifolds on which there is a rule for measuring distances and angles, subject to certain natural restrictions to ensure that these quantities behave analogously to their Euclidean counterparts. The properties of interest are those that are invariant under *isometries*, or distance-preserving transformations. For example, one can study the relationship between the curvature of an n-dimensional Riemannian manifold (a local property) and its global topological type. A typical theorem is that a complete Riemannian n-manifold whose curvature is everywhere larger than some fixed positive number must be compact and have a finite fundamental group (not too many holes). The search for such relationships is one of the principal activities in Riemannian geometry, a thriving field of contemporary research. See Chapter 1 of [Lee97] for an informal introduction to the subject.

Complex Analysis

Complex analysis is the study of holomorphic (i.e., complex analytic) functions. Some such functions are naturally "multiple-valued." A typical example is the complex square root. Except for zero, every complex number has two distinct square roots. But unlike the case of positive real numbers, where we can always unambiguously choose the positive square root to denote by the symbol $\sqrt{x}$, it is not possible to define a global continuous square root function on the complex plane. To see why, write z in polar coordinates as $z = re^{i\theta}$. Then the two square roots of z can be written $\sqrt{r}\, e^{i\theta/2}$ and $\sqrt{r}\, e^{i(\theta/2+\pi)}$. As θ increases from 0 to 2π, the first square root goes from the positive real axis through the upper half-plane to the negative real axis, while the second goes from the negative real axis through the lower half-plane to the positive real axis. Thus whichever continuous square root function we start with on the positive real axis, we are forced to choose the other after having made one circuit around the origin.

Even though a "two-valued function" is properly considered as a relation and not really a function at all, we can define the *graph* of such a relation in an unambiguous way. To warm up with a simpler example, consider the two-valued square root "function" on the nonnegative real axis. Its graph

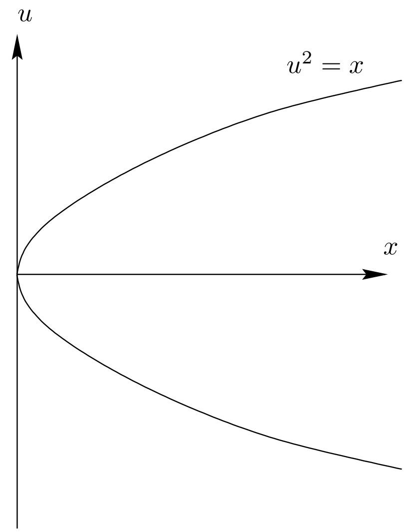


FIGURE 1.6. Graph of the two branches of the real square root.

is defined to be the set of pairs $(x, u) \in \mathbb{R} \times \mathbb{R}$ such that $u = \pm\sqrt{x}$, or equivalently $u^2 = x$. This is a parabola opening in the positive x direction (Figure 1.6), which we can think of as the two "branches" of the square root.

Similarly, the graph of the two-valued complex square root "function" is the set of pairs $(z, w) \in \mathbb{C} \times \mathbb{C}$ such that $w^2 = z$. Over each small disk $U \subset \mathbb{C}$ that does not contain 0, this graph has two branches or "sheets," corresponding to the two possible continuous choices of square root function on U (Figure 1.7). If you start on one sheet above the positive real axis and pass once around the origin in the counterclockwise direction, you end up on the other sheet. Going around once more brings you back to the first sheet.

It turns out that this graph in $\mathbb{C}^2$ is a 2-dimensional manifold, of a special type called a *Riemann surface*—this is essentially a 2-manifold on which there is some way to define holomorphic functions. Riemann surfaces are of great importance in complex analysis, since any holomorphic function gives rise to a Riemann surface by a procedure analogous to the one we sketched above. The surface we constructed turns out to be topologically equivalent to a plane, but more complicated functions can give rise to more complicated surfaces. For example, the two-valued "function" $f(z) = \pm\sqrt{z^3 - z}$ yields a Riemann surface that is homeomorphic to a plane with one "handle" attached.

One of the fundamental tasks of complex analysis is to understand the topological type (number of "holes" or "handles") of the Riemann surface

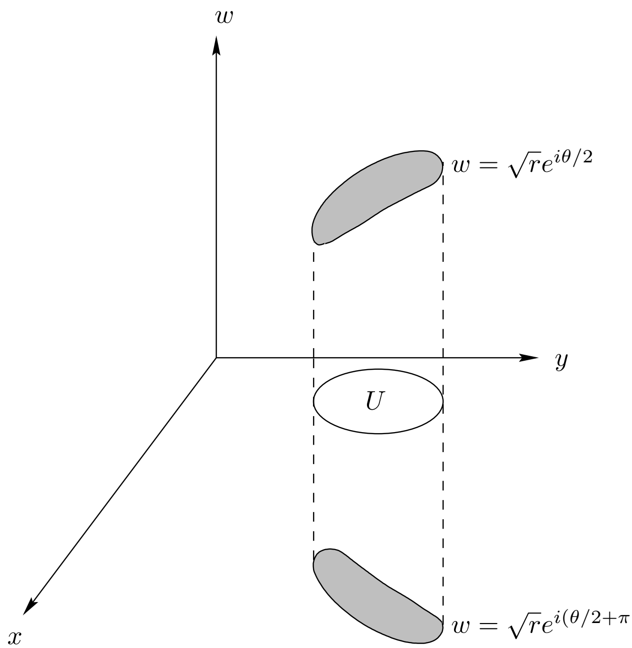


FIGURE 1.7. Two branches of the complex square root.

of a given function, and how it relates to the analytic properties of the function.

Algebra

One of the most important objects studied in abstract algebra is the *general linear group* $\mathrm{GL}(n, \mathbb{R})$, which is the group of $n \times n$ invertible real matrices. As a set, it can be identified with a subset of n^2-dimensional Euclidean space, simply by stringing all the matrix entries out in a row. Since a matrix is invertible if and only if its determinant is nonzero, $\mathrm{GL}(n, \mathbb{R})$ is an open subset of $\mathbb{R}^{n^2}$, and is therefore an n^2-dimensional manifold. Similarly, the *complex general linear group* $\mathrm{GL}(n, \mathbb{C})$ is the group of $n \times n$ invertible complex matrices; it is a $2n^2$-manifold, because we can identify $\mathbb{C}^{n^2}$ with $\mathbb{R}^{2n^2}$.

A *Lie group* is a group (in the algebraic sense) that is also a manifold, together with some technical conditions to ensure that the group structure and the manifold structure are compatible with each other. They play a central role in differential geometry, representation theory, and mathematical physics, among many other fields. The most important Lie groups are subgroups of the real and complex general linear groups. Some commonly

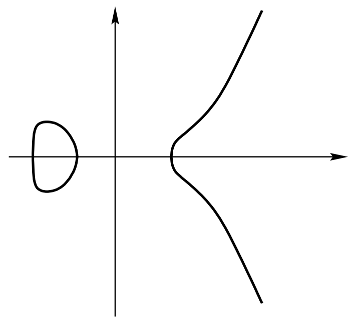


FIGURE 1.8. A plane curve with disconnected pieces.

encountered examples are the *special linear group* $\mathrm{SL}(n,\mathbb{R}) \subset \mathrm{GL}(n,\mathbb{R})$, consisting of matrices with determinant 1; the *orthogonal group* $\mathrm{O}(n) \subset \mathrm{GL}(n,\mathbb{R})$, consisting of matrices whose columns are orthonormal; the *special orthogonal group* $\mathrm{SO}(n) = \mathrm{O}(n) \cap \mathrm{SL}(n,\mathbb{R})$; and their complex analogues, the *complex special linear group* $\mathrm{SL}(n,\mathbb{C}) \subset \mathrm{GL}(n,\mathbb{C})$, the *unitary group* $\mathrm{U}(n) \subset \mathrm{GL}(n,\mathbb{C})$, and the *special unitary group* $\mathrm{SU}(n) = \mathrm{U}(n) \cap \mathrm{SL}(n,\mathbb{C})$.

It is important to understand the topological structure of a Lie group and how its topological structure relates to its algebraic structure. For example, it can be shown that $\mathrm{SO}(2)$ is topologically equivalent to a circle, $\mathrm{SU}(2)$ is topologically equivalent to the 3-sphere, and any connected abelian Lie group is topologically equivalent to a Cartesian product of circles and lines. Lie groups provide a rich source of examples of manifolds in all dimensions.

Algebraic Geometry

Algebraic geometers study the geometric properties of solution sets to systems of polynomial equations. Many of the basic questions of algebraic geometry can be posed very naturally in the elementary context of plane curves defined by polynomial equations. For example: How many intersection points can one expect between two plane curves defined by polynomials of degrees k and l? (Not more than kl, but sometimes fewer.) How many disconnected "pieces" does the solution set to a particular polynomial equation have (Figure 1.8)? Does a plane curve have any self crossings (Figure 1.9) or "cusps" (points where the tangent vector does not vary continuously—Figure 1.10)?

But the real power of algebraic geometry becomes evident only when one focuses on polynomials with coefficients in an *algebraically closed* field (one in which every polynomial decomposes into a product of linear factors), since polynomial equations always have the expected number of solutions (counted with multiplicity) in that case. The most deeply studied case is the complex field; in this context the solution set to a system of complex

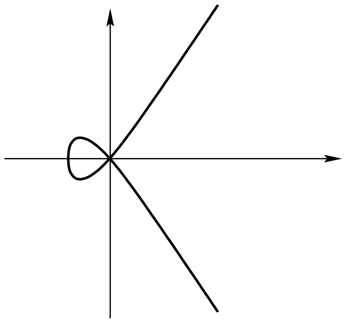


FIGURE 1.9. A self crossing.

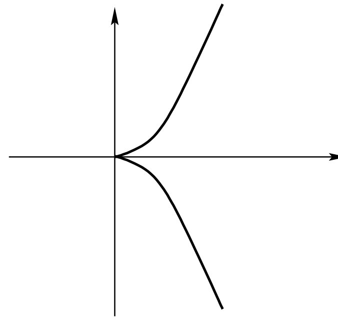


FIGURE 1.10. A cusp.

polynomials in n variables is a certain geometric object in $\mathbb{C}^n$ called an *algebraic variety*, which (except for a small subset where there might be self crossings, cusps, or more complicated kinds of behavior) is a manifold. The subject becomes even more interesting if one enlarges $\mathbb{C}^n$ by adding "ideal points at infinity" where parallel lines or asymptotic curves can be thought of as meeting; the resulting space is called *complex projective space*, and is an extremely important manifold in its own right.

The properties of interest are those that are invariant under projective transformations (the natural changes of coordinates on projective space). One can ask such questions as these: Is a given variety a manifold or does it have *singular points* (points where it fails to be a manifold)? If it is a manifold, what is its topological type? If it is not a manifold, what is the geometric structure of its singular set, and how does that set change when one varies the coefficients of the polynomials slightly? If two varieties are homeomorphic, are they equivalent under a projective transformation? How many times and in what way do two or more varieties intersect?

Algebraic geometry has contributed a prodigious supply of examples of manifolds. In particular, much of the recent progress in understanding 4-dimensional manifolds has been driven by the wealth of examples that arise as algebraic varieties.

Classical Mechanics

Classical mechanics is the study of systems that obey Newton's laws of motion. The positions of all the objects in the system at any given time can be described by a set of numbers, or coordinates; typically, these are not independent of each other but instead must satisfy some relations. The relations can usually be interpreted as defining a manifold in some Euclidean space.

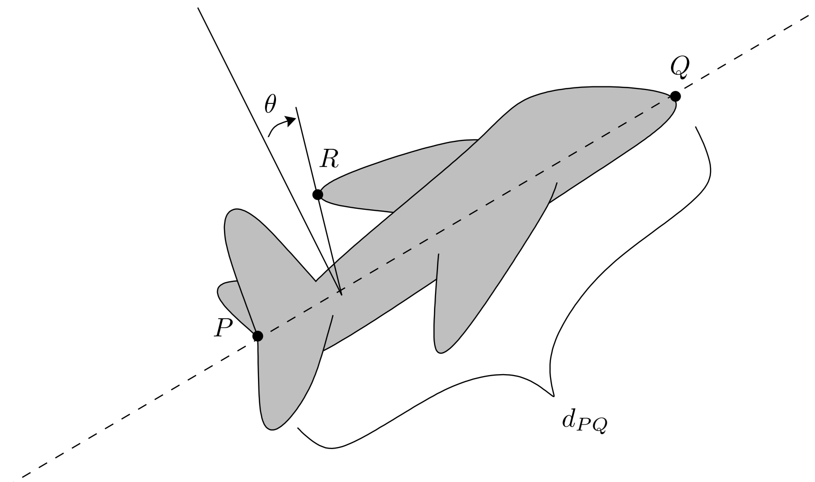


FIGURE 1.11. A rigid body in space.

For example, consider a rigid body moving through space under the influence of gravity. If we choose three noncollinear points P, Q, and R on the body (Figure 1.11), the position of the body is completely specified once we know the coordinates of these three points, which correspond to a point in $\mathbb{R}^9$. However, the positions of the three points cannot all be specified arbitrarily: Because the body is rigid, they are subject to the constraint that the distances between pairs of points are fixed. Thus, to determine the position of the body, we can arbitrarily specify the coordinates of P in space (three parameters), and then we can specify the position of Q by giving, say, its latitude and longitude on the sphere of radius d_{PQ}, the fixed distance between P and Q (two more parameters). Finally, having determined the position of the two points P and Q, the only remaining freedom is to rotate R around the line PQ; so we can specify the position of R by giving the angle θ that the plane PQR makes with some reference plane (one more parameter). Thus the set of possible positions of the body is a certain 6-dimensional manifold $M \subset \mathbb{R}^9$.

Newton's second law of motion expresses the acceleration of the object—that is, the second derivatives of the coordinates of P, Q, R—in terms of the force of gravity, which is a certain function of the object's position. This can be interpreted as a system of second-order ordinary differential equations for the position coordinates, whose solutions are all the possible paths the rigid body can take on the manifold M.

The study of classical mechanics can thus be interpreted as the study of ordinary differential equations on manifolds, also known as *smooth dynamical systems.* A wealth of interesting questions arise in this subject: How do solutions behave over the long term? Are there any equilibrium points or periodic trajectories? If so, are they *stable*, that is, do nearby trajectories stay nearby? A good understanding of manifolds is necessary to fully answer these questions.

General Relativity

Manifolds play a decisive role in Einstein's general theory of relativity, which describes the interactions among matter, energy, and gravitational forces. The central assertion of the theory is that spacetime (the collection of all points in space at all times in history) can be modeled by a 4-dimensional manifold that carries a certain kind of geometric structure called a *Lorentz metric*; and this metric satisfies a system of partial differential equations called the *Einstein field equations*. Gravitational effects are then interpreted as manifestations of the curvature of the Lorentz metric.

In order to describe the global structure of the universe, its history, and its possible futures, it is important to understand first of all what kinds of 4-manifolds exist and what kinds of Lorentz metrics they can carry. There are especially interesting relationships between the local geometry of spacetime (as reflected in the local distribution of matter and energy) and the global topological structure of the universe; these relationships are similar to those described above for Riemannian manifolds, but are more complicated because of the introduction of forces and motion into the picture. In particular, if we assume that on a cosmic scale the universe looks approximately the same at all points and in all directions (such a spacetime is said to be *homogeneous* and *isotropic*), then it turns out there is a critical value for the average density of matter and energy in the universe: Above this density, the universe closes up on itself spatially and will collapse to a point singularity in a finite time (the "big crunch"); below it, the universe extends infinitely far in all directions and will expand forever. Interestingly, physicists' best current estimates place the average density rather near the critical value, and they have so far been unable to determine whether it is above or below it, so they do not know whether the universe will go on existing forever or not.

Quantum Field Theory

The theory of elementary particle interactions, called quantum field theory, has become increasingly geometric in recent decades. In particular, the latest attempts to unify quantum theory and gravitation have led to ever more interesting and exotic geometric structures. The approach to quantum gravity that is currently considered most promising by many physicists is *string theory*, in which manifolds appear in several different starring roles.

First, in order to obtain a consistent theory, it seems to be necessary to assume that spacetime has more than four dimensions. We experience only four of them directly, because the dimensions beyond four are so tightly "curled up" that they are not visible on a macroscopic scale, much as a long but microscopically narrow cylinder would appear to be one-dimensional when viewed from far enough away. The topological properties of the manifold that appears as the "cross section" of the curled-up dimensions have such a profound effect on the observable dynamics of the resulting quantum field theory that it is possible to rule out most cross sections a priori. It currently appears that a consistent theory can be constructed only if the cross section is a certain kind of 6-dimensional manifold known as a *Calabi–Yau manifold*. These developments in physics have stimulated profound developments in the mathematical understanding of 6-manifolds in general and Calabi-Yau manifolds in particular.

Another role that manifolds play in string theory is in describing the history of an elementary particle. One of the central tenets of string theory is that particles should be represented not as points, but as tiny 1-dimensional objects ("strings") moving through spacetime. As a particle moves, it traces out a 2-dimensional manifold called its *world sheet*. Physical phenomena arise from the interactions among these different topological and geometric structures: the world sheet, the Calabi-Yau cross section, and the macroscopic four-dimensional spacetime that we see.

Manifolds are used in many more areas of mathematics than the ones listed here, but this brief survey should be enough to show you that manifolds have a rich assortment of applications. It is time to get to work.

2
Topological Spaces

In this chapter we begin our study in earnest. The first order of business is to build up enough machinery to give a proper definition of manifolds. The chief problem with the preliminary definition given in Chapter 1 is that it depends on having an "ambient Euclidean space" in which our n-manifold lives. This introduces a great deal of extraneous structure that is irrelevant to our purposes. Instead, we would like to view a manifold as a mathematical object in its own right, not as a subset of some larger space. The key concept that makes this possible is that of a "topological space," which is the main topic of this chapter.

We begin by defining topological spaces, motivated by the open set criterion for continuity in metric spaces. After the definition we introduce some of the important elementary notions associated with topological spaces such as convergence, continuity, homeomorphisms, closures, interiors, and exteriors, and then explore how to construct topologies from bases. At the end of the chapter we give the official definition of a manifold as a topological space with special properties.

Topologies

One of the most useful tools in analysis is the concept of a metric space. (See the Appendix for a brief review of metric space theory.) The most important examples, of course, are (subsets of) Euclidean spaces with the

Euclidean metric, but many others, such as function spaces, arise frequently in analysis.

Our goal in this book is to study manifolds and those of their properties that are preserved by homeomorphisms (continuous maps with continuous inverses). To accomplish this, we could choose to view our manifolds as metric spaces. However, a metric still contains extraneous information. It is obvious that a homeomorphism between metric spaces need not preserve distances (just think of the obvious homeomorphism between two spheres of different radii). So we will push the process of abstraction a step further, and come up with a kind of "space" without distances in which continuous functions still make sense.

The key motivation behind the definition of this new kind of space is the open set criterion for continuity (Lemma A.5 in the Appendix), which shows that continuous functions between metric spaces can be detected knowing only the open sets. Motivated by this observation, we make the following definition. A *topology* on a set X is a collection $\mathcal{T}$ of subsets of X, called *open sets*, satisfying the following properties:

(i) X and $\varnothing$ are elements of $\mathcal{T}$.

(ii) $\mathcal{T}$ is closed under finite intersections: If $U_1, \dots, U_n \in \mathcal{T}$, then their intersection $U_1 \cap \cdots \cap U_n$ is in $\mathcal{T}$.

(iii) $\mathcal{T}$ is closed under arbitrary unions: If $\{U_\alpha\}_{\alpha \in A}$ is any (finite or infinite) collection of elements of $\mathcal{T}$, then their union $\bigcup_{\alpha \in A} U_\alpha$ is in $\mathcal{T}$.

A pair $(X, \mathcal{T})$ consisting of a set X and a topology $\mathcal{T}$ on X is called a *topological space*. The elements of a topological space are usually called its *points*. Since we will rarely have occasion to discuss any other type of space in this book, we will sometimes follow the common practice of calling a topological space simply a *space*. As is common in mathematics in discussing a set endowed with a particular kind of structure, if the topology is understood from the context, we will typically omit it from the notation and simply say "X is a topological space" or "X is a space."

Aside from the simplicity of the open set criterion for continuity, the other reason for choosing open sets as the primary objects in the definition of a topological space is that they give us a qualitative way to detect "nearness" to a point without necessarily having a quantitative measure of nearness as we would in a metric space. If X is a topological space and $q \in X$, a *neighborhood* of q is just an open set containing q. More generally, a neighborhood of a subset $K \subset X$ is an open set containing K. (In some books, the word neighborhood is used in the more general sense of a set containing an open set containing q; but for us neighborhoods will always be open.) We think of something being true "near q" if it is true in some (or every, depending on the context) neighborhood of q.

The following exercises give some simple examples of topological spaces.

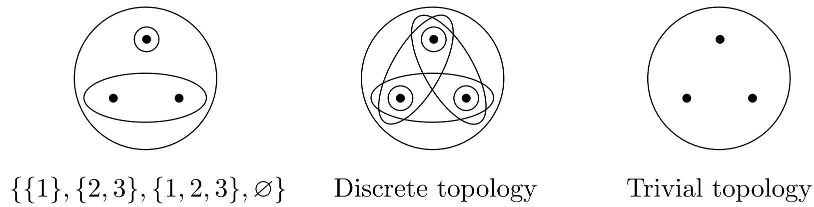


FIGURE 2.1. Topologies on $\{1,2,3\}$.

Exercise 2.1. Show that each of the following is a topological space. (See Figure 2.1.)

(a) Let X denote the set $\{1,2,3\}$, and declare the open sets to be $\{1\}$, $\{2,3\}$, $\{1,2,3\}$, and the empty set.

(b) Any set X whatsoever, with $\mathcal{T} = \{\text{all subsets of } X\}$. This is called the *discrete topology* on X, and $(X, \mathcal{T})$ is called a *discrete space.*

(c) Any set X, with $\mathcal{T} = \{\varnothing, X\}$. This is called the *trivial topology* on X.

(d) Any metric space (M, d), with $\mathcal{T}$ equal to the collection of all subsets of M that are open in the metric space sense. This topology is called the *metric topology* on M.

Metric spaces provide a rich source of examples of topological spaces. In fact, a large percentage of the topological spaces we will need to consider are actually subsets of Euclidean spaces $\mathbb{R}^n$, with the metric topology induced by the Euclidean metric (which we call the *Euclidean topology*). Unless we specify otherwise, subsets of $\mathbb{R}^n$ will always be considered as topological spaces with this topology. Thus our intuition regarding topological spaces will rely heavily on our understanding of subsets of Euclidean space.

Another important class of examples of topological spaces is obtained by taking open subsets of other spaces. If X is a topological space, and Y is any open subset of X, then we can define a topology on Y just by declaring the open sets of Y to be those open sets of X that are contained in Y. It is trivial to check that the three defining properties of a topology are satisfied. (In the next chapter, we will show how to put a topology on *any* subset of a topological space.)

Convergence and Continuity

The primary reason topological spaces were invented was that they provide the most general setting for studying the notions of convergence and continuity. For this reason, it is appropriate to introduce these concepts next. We begin with convergence.

The definition of convergence of a sequence of points in a metric space (see the Appendix) is really just a fancy way of saying that as we go far enough out in the sequence, the points of the sequence become "arbitrarily close" to q.

In topological spaces, we use neighborhoods to encode the notion of "arbitrarily close." Thus, if X is a topological space and $\{q_i\}$ is any sequence of points in X, we say that the sequence *converges to* $q \in X$, and q is the *limit* of the sequence, if for every neighborhood U of q there exists N such that $q_i \in U$ for all $i \geq N$. Symbolically, this is denoted by either $q_i \to q$ or $\lim_{i\to\infty} q_i = q$.

Exercise 2.2. Show that in a metric space, this topological definition of convergence is equivalent to the metric space definition.

For the types of topological spaces we will be chiefly interested in (mostly manifolds), convergent sequences behave very much the same way we are used to from our experience with Euclidean space. Nevertheless, it is good to be aware that for some of the stranger examples of topological spaces, convergence can sometimes have an unintuitive meaning, as the following exercises show.

Exercise 2.3.

(a) Let X be a discrete topological space. Show that the only convergent sequences in X are the ones that are "eventually constant," that is, sequences $\{q_i\}$ such that $q_i = q$ for all i greater than some N.

(b) Let Y be a trivial topological space (that is, a set with the trivial topology $\{\varnothing, Y\}$). Show that every sequence in Y converges to every point of Y.

At the end of this chapter we will describe a restricted class of topological spaces (Hausdorff spaces) for which the pathological behavior of (b) cannot occur.

Next we address the most important topological concept of all: continuous maps. If X and Y are topological spaces, a map $f\colon X \to Y$ is said to be *continuous* if for every open set $U \subset Y$, $f^{-1}(U)$ is open in X.

The open set criterion (Lemma A.5) for continuity in metric spaces says precisely that a map between metric spaces is continuous in this sense if and only if it is continuous in the usual ε-δ sense. Therefore, all the maps that you know to be continuous from metric space theory are also continuous as maps of topological spaces. Examples include polynomial functions from $\mathbb{R}$ to $\mathbb{R}$, linear transformations from $\mathbb{R}^n$ to $\mathbb{R}^k$, and, more generally, any map from a subset of $\mathbb{R}^n$ to $\mathbb{R}^k$ whose component functions are continuous in the ordinary sense, such as polynomial, exponential, rational, logarithmic, absolute value, and trigonometric functions (where they are defined), and functions built up from these by composition.

The next lemma gives some elementary but important properties of continuous maps. The ease with which properties like this can be proved is one of the virtues of defining continuity in terms of open sets.

Lemma 2.1. *Let X, Y, and Z be topological spaces.*

(*a*) *Any constant map $f\colon X \to Y$ is continuous.*

(*b*) *The identity map* $\mathrm{Id}\colon X \to X$ *is continuous.*

(*c*) *If $f\colon X \to Y$ is continuous, so is the restriction of f to any open subset of X.*

(*d*) *If $f\colon X \to Y$ and $g\colon Y \to Z$ are continuous, so is their composition $g \circ f\colon X \to Z$.*

Proof. We will prove (d) and leave the other parts as exercises. We have to show that if U is any open subset of Z, then $(g \circ f)^{-1}(U)$ is an open subset of X. By elementary set-theoretic considerations, $(g \circ f)^{-1}(U) = f^{-1}(g^{-1}(U))$. Applying the definition of continuity to g, $g^{-1}(U)$ is open; and then doing the same for f shows that $f^{-1}(g^{-1}(U))$ is open. □

Exercise 2.4. Prove parts (a)–(c) of Lemma 2.1.

In metric spaces it makes sense to talk about a map being "continuous at a point" ($f\colon M_1 \to M_2$ is continuous at $x \in M_1$ if for all $\varepsilon > 0$, there exists $\delta > 0$ such that for each $y \in M_1$, $d_1(y, x) < \delta$ implies $d_2(f(y), f(x)) < \varepsilon$), and a map is continuous if and only if it is continuous at every point. In topological spaces, continuity at a point is generally not a very useful concept. However, it is an important fact that continuity is a "local" property, in the sense that a map is continuous if and only if it is continuous in a neighborhood of every point. The precise statement is given in the following important lemma.

Lemma 2.2 (Local Criterion for Continuity). *A map $f\colon X \to Y$ between topological spaces is continuous if and only if each point of X has a neighborhood on which* (*the restriction of*) *f is continuous.*

Proof. If f is continuous, we may simply take each neighborhood to be X itself. Conversely, suppose f is continuous in a neighborhood of each point, and let $U \subset Y$ be any open set; we have to show that $f^{-1}(U)$ is open. Any point $x \in f^{-1}(U)$ has a neighborhood V_x on which f is continuous (Figure 2.2). Continuity of $f|_{V_x}$ implies, in particular, that $(f|_{V_x})^{-1}(U)$ is open in V_x, and therefore also open in X. Unwinding the definitions, we see that

$$(f|_{V_x})^{-1}(U) = \{x \in V_x : f(x) \in U\} = f^{-1}(U) \cap V_x,$$

which contains x and is contained in $f^{-1}(U)$. Since $f^{-1}(U)$ is the union of all such open sets as x ranges over $f^{-1}(U)$, it follows that $f^{-1}(U)$ is open, as desired. □

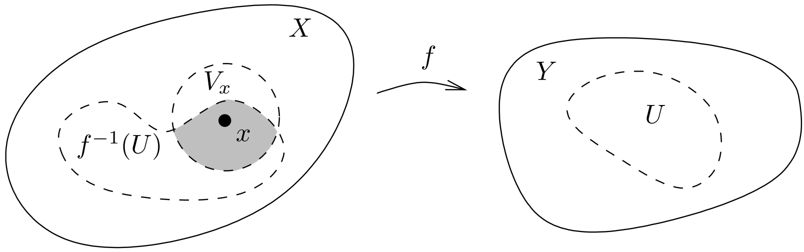


FIGURE 2.2. Local criterion for continuity.

If X and Y are topological spaces, a *homeomorphism* from X to Y is defined to be a continuous bijective map $\varphi\colon X \to Y$ with continuous inverse. If there exists a homeomorphism between X and Y, we say that X and Y are *homeomorphic* or *topologically equivalent*. Sometimes this is abbreviated $X \approx Y$.

Exercise 2.5. Show that "homeomorphic" is an equivalence relation.

The homeomorphism relation is the most fundamental relation in topology. In fact, as we mentioned in Chapter 1, "topological properties" are exactly those that are preserved by homeomorphisms.

Here are a few explicit examples of homeomorphisms that you should keep in mind.

Example 2.3. Any open ball in $\mathbb{R}^n$ is homeomorphic to any other open ball; the homeomorphism can easily be constructed as a composition of translations $x \mapsto x + x_0$ and dilations $x \mapsto cx$. Similarly, all spheres in $\mathbb{R}^n$ are homeomorphic to each other. These examples illustrate that "size" is not a topological property.

Example 2.4. Let $\mathbb{B}^n$ denote the open unit ball $B_1(0) \subset \mathbb{R}^n$, and define a map $F\colon \mathbb{B}^n \to \mathbb{R}^n$ by

$$y = F(x) = \frac{x}{1 - |x|^2}.$$

Note that $|F(x)| = |x|/(1 - |x|^2) \to \infty$ as $|x| \to 1$. It is straightforward to check that the inverse of F is given by

$$x = F^{-1}(y) = \frac{2y}{1 + \sqrt{1 + 4|y|^2}}.$$

Thus F is a homeomorphism, so $\mathbb{R}^n$ is homeomorphic to $\mathbb{B}^n$. This shows that "boundedness" is not a topological property.

Example 2.5. Another illustrative example is the homeomorphism between a cube and a sphere alluded to in Chapter 1. Let $\mathbb{S}^2 = \{x \in$

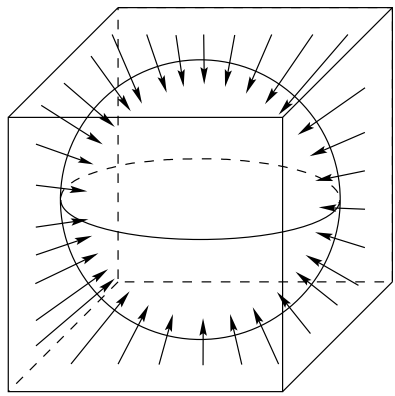


FIGURE 2.3. Deforming a cube into a sphere.

$\mathbb{R}^3 : |x| = 1\}$ denote the unit sphere in $\mathbb{R}^3$, and set $C = \{(x,y,z) : \max(|x|,|y|,|z|) = 1\}$, which is the cubical surface of side 2 centered at the origin. Let $\varphi\colon C \to \mathbb{S}^2$ be the map that projects each point on C radially inward to the sphere (Figure 2.3). More precisely, given a point $q \in C$, $\varphi(q)$ is the unit vector in the direction of q. Thus φ is given by the formula

$$\varphi(x,y,z) = \frac{(x,y,z)}{\sqrt{x^2+y^2+z^2}},$$

which is continuous by the usual arguments of elementary analysis. The next exercise shows that φ is a homeomorphism. This example demonstrates that "corners" are not topological properties.

Exercise 2.6. Show that the map $\varphi\colon C \to \mathbb{S}^2$ is a homeomorphism by showing that its inverse can be written

$$\varphi^{-1}(x,y,z) = \frac{(x,y,z)}{\max\left(|x|,|y|,|z|\right)}.$$

In the definition of a homeomorphism, it is important to note that although the assumption that φ is bijective guarantees that the inverse map φ^{-1} exists for set-theoretic reasons, continuity of φ^{-1} is not automatic. The next exercise gives an example of a continuous bijection whose inverse is not continuous.

Exercise 2.7. Let X denote the half-open interval $[0,1) \subset \mathbb{R}$, and let $\mathbb{S}^1$ denote the unit circle in $\mathbb{R}^2$ (both with the Euclidean metric topologies, of course). Define a map $a\colon X \to \mathbb{S}^1$ by $a(t) = (\cos 2\pi t, \sin 2\pi t)$ (Figure 2.4). Show that a is continuous and bijective but not a homeomorphism.

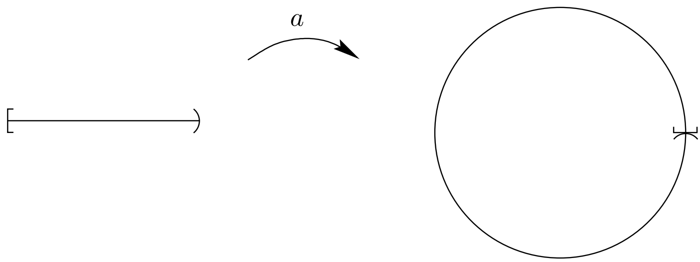


FIGURE 2.4. A map that is bijective but not a homeomorphism.

A map $f\colon X \to Y$ (continuous or not) is said to be an *open map* if for any open set $U \subset X$, the image set $f(U)$ is open in Y. A map can be open but not continuous, continuous but not open, both, or neither.

There is a generalization of homeomorphisms that is often useful. We say that a continuous map $f\colon X \to Y$ between topological spaces is a *local homeomorphism* if every point $x \in X$ has a neighborhood $U \subset X$ such that $f(U)$ is an open subset of Y and $f|_U\colon U \to f(U)$ is a homeomorphism.

Exercise 2.8.

(a) Show that every local homeomorphism is an open map.

(b) Show that every homeomorphism is a local homeomorphism.

(c) Show that a bijective continuous open map is a homeomorphism.

(d) Show that a bijective local homeomorphism is a homeomorphism.

Closed Sets

Because of the importance of neighborhoods in understanding a topological space and its continuous maps and convergent sequences, the definition of a topological space takes open sets as the primary objects. There is also a complementary notion that is nearly as important.

A subset F of a topological space X is said to be *closed* if its complement $X \setminus F$ is open. From the definition of topological spaces, several properties follow immediately:

(i) X and $\varnothing$ are closed.

(ii) Finite unions of closed sets are closed.

(iii) Arbitrary intersections of closed sets are closed.

A topology on a set X can be defined by describing the collection of closed sets, as long as they satisfy these three properties; the open sets are then just those sets whose complements are closed.

Here are some examples of closed subsets of familiar topological spaces.

Example 2.6 (Closed Sets).

(a) Any closed interval $[a, b] \subset \mathbb{R}$ is a closed set, as are the half-infinite closed intervals $[a, \infty)$ and $(-\infty, b]$.

(b) Any closed ball in a metric space is a closed set (Exercise A.11(b) in the Appendix).

(c) Every subset of a discrete space is closed.

It is important to be aware that just as in metric spaces, "closed" is not the same as "not open"—sets can be both open and closed, or neither open nor closed. For example, in any topological space X, the sets X and $\varnothing$ are both open and closed. On the other hand, the half-open interval $[0, 1)$ is neither open nor closed in $\mathbb{R}$.

Continuity can be detected by closed sets as well as open ones.

Lemma 2.7. *A map between topological spaces is continuous if and only if the inverse image of every closed set is closed.*

Exercise 2.9. Prove Lemma 2.7.

Given any set $A \subset X$, we define several related sets as follows. The *closure* of A in X, denoted by $\overline{A}$, is the set

$$\overline{A} = \bigcap \{B \subset X : B \supset A \text{ and } B \text{ is closed in } X\}.$$

The *interior* of A, written $\operatorname{Int} A$, is

$$\operatorname{Int} A = \bigcup \{C \subset X : C \subset A \text{ and } C \text{ is open in } X\}.$$

It is obvious from the properties of open and closed sets that $\overline{A}$ is closed and $\operatorname{Int} A$ is open. In words, $\overline{A}$ is "the smallest closed set containing A," and $\operatorname{Int} A$ is "the largest open set contained in A."

We also define the *exterior* of A, written $\operatorname{Ext} A$, as

$$\operatorname{Ext} A = X \smallsetminus \overline{A},$$

and the *boundary* of A, written ∂A, as

$$\partial A = X \smallsetminus (\operatorname{Int} A \cup \operatorname{Ext} A).$$

It follows immediately from the definitions that for any subset $A \subset X$, the whole space X is equal to the disjoint union of $\operatorname{Int} A$, $\operatorname{Ext} A$, and ∂A. The set A always contains all of its interior points and none of its exterior points, and may contain all, some, or none of its boundary points.

For many purposes, it is useful to have alternative characterizations of open and closed sets, and of the interior, exterior, closure, and boundary of a given set. The following lemma gives such characterizations. Some of these are probably familiar to you from your study of Euclidean and metric spaces. See Figure 2.5 for illustrations of some of these characterizations.

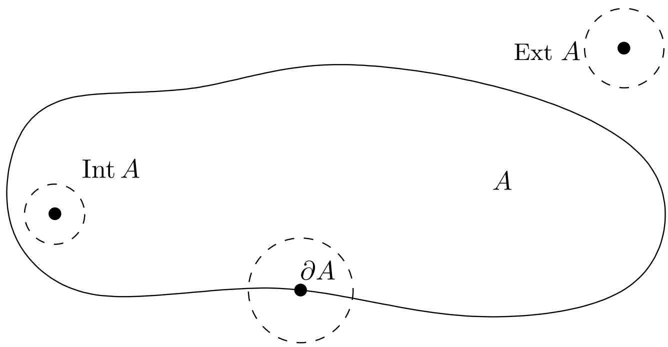


FIGURE 2.5. Interior, exterior, and boundary points.

Lemma 2.8. *Let X be a topological space and $A \subset X$ any subset.*

(*a*) *A point q is in the interior of A if and only if q has a neighborhood contained in A.*

(*b*) *A point q is in the exterior of A if and only if q has a neighborhood contained in $X \smallsetminus A$.*

(*c*) *A point q is in the boundary of A if and only if every neighborhood of q contains both a point of A and a point of $X \smallsetminus A$.*

(*d*) $\operatorname{Int} A$ *and* $\operatorname{Ext} A$ *are open in X, while ∂A is closed in X.*

(*e*) *A is open if and only if $A = \operatorname{Int} A$.*

(*f*) *A is closed if and only if it contains all its boundary points, which is true if and only if $A = \operatorname{Int} A \cup \partial A$.*

(*g*) $\overline{A} = A \cup \partial A = \operatorname{Int} A \cup \partial A$.

Exercise 2.10. Prove Lemma 2.8.

Given a topological space X and a set $A \subset X$, we say that a point $q \in X$ is a *limit point* of A if every neighborhood of q contains a point of A other than q (which might or might not itself be in A). Limit points are also sometimes called *accumulation points* or *cluster points*. For example, if $X = \mathbb{R}$ and $A = (0,1)$, then every point in $[0,1]$ is a limit point of A. If we let $B = \{1/n\}_{n=1}^{\infty} \subset \mathbb{R}$, then 0 is the only limit point of B.

Exercise 2.11. Show that a set A in a topological space is closed if and only if it contains all of its limit points.

A subset A of a topological space X is said to be *dense* in X if $\overline{A} = X$.

Exercise 2.12. Show that a subset $A \subset X$ is dense if and only if every nonempty open set in X contains a point of A.

Exercise 2.13. Show that $\mathbb{R}^n$ has a countable dense subset.

Analogous to open maps are closed maps: A map $f\colon X \to Y$ is said to be *closed* if it takes closed sets in X to closed sets in Y.

Exercise 2.14. Show that a bijective continuous map is a homeomorphism if and only if it is open if and only if it is closed.

Bases

To define a topology on a given set, it is often convenient to single out some "special" open sets and use them to define the rest of the open sets. For example, the metric topology is defined by first defining balls and then declaring a set to be open if it contains a ball around each of its points. This idea can be generalized easily to arbitrary topological spaces, as in the next definition.

Suppose X is any set. A *basis* in X is a collection $\mathcal{B}$ of subsets of X satisfying the following conditions:

(i) Every element of X is in some element of $\mathcal{B}$; in other words, $X = \bigcup_{B \in \mathcal{B}} B$.

(ii) If $B_1, B_2 \in \mathcal{B}$ and $x \in B_1 \cap B_2$, there exists an element $B_3 \in \mathcal{B}$ such that $x \in B_3 \subset B_1 \cap B_2$.

Proposition 2.9. *Let $\mathcal{B}$ be a basis in a set X, and let $\mathcal{T}$ be the collection of all unions of elements of $\mathcal{B}$. Then $\mathcal{T}$ is a topology on X.*

This topology $\mathcal{T}$ is called the *topology generated by* $\mathcal{B}$. Before we prove the proposition, it will be useful to have an alternative characterization of $\mathcal{T}$, analogous to the definition of open sets in a metric space in terms of balls. Given X and a collection $\mathcal{B}$ of subsets of X, we say that a subset $U \subset X$ satisfies the *basis criterion* with respect to $\mathcal{B}$ if for every $x \in U$, there exists $B \in \mathcal{B}$ such that $x \in B \subset U$.

Lemma 2.10. *Suppose $\mathcal{B}$ is a basis in X. Then the collection $\mathcal{T}$ defined in Proposition 2.9 is precisely the set of all subsets of X that satisfy the basis criterion with respect to $\mathcal{B}$.*

Proof. Let $U \subset X$, and suppose first that U satisfies the basis criterion. Let

$$V = \bigcup \{B \in \mathcal{B} : B \subset U\}.$$

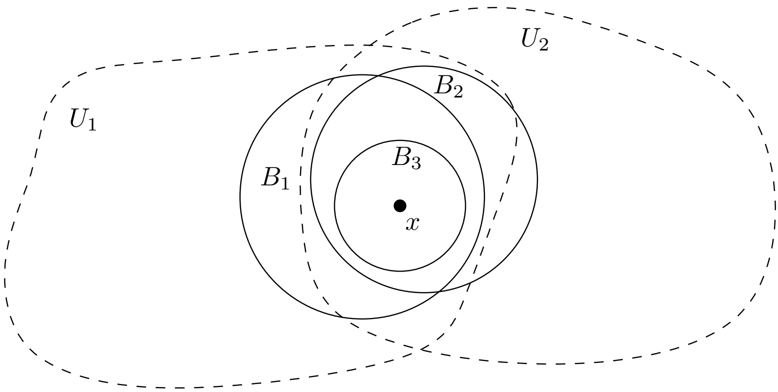


FIGURE 2.6. Proof that $U_1 \cap U_2$ satisfies the basis criterion.

Obviously, $V \in \mathcal{T}$, since V is by definition a union of basis sets. Thus if we can show that $U = V$, it will follow that $U \in \mathcal{T}$. Clearly, $V \subset U$, because V is a union of subsets of U. To show that $U \subset V$, let $x \in U$ be arbitrary; since U satisfies the basis criterion, there must exist a basis set $B \in \mathcal{B}$ such that $x \in B \subset U$. It follows immediately that $x \in V$, so we are done.

Conversely, suppose that $U \in \mathcal{T}$. This means that U is a union of elements of $\mathcal{B}$, i.e., for some subset $\mathcal{A} \subset \mathcal{B}$, $U = \bigcup_{B \in \mathcal{A}} B$. In other words, $x \in U$ if and only if $x \in B$ for some $B \in \mathcal{A}$. In particular, $x \in B \subset U$, so U satisfies the basis criterion. □

Proof of Proposition 2.9. We need to show that $\mathcal{T}$ satisfies the three defining conditions for a topology. First, the fact that $X = \bigcup_{B \in \mathcal{B}} B$ means that $X \in \mathcal{T}$. The empty set is also in $\mathcal{T}$ trivially. (It is the "union of no elements of $\mathcal{B}$"!) A union of elements of $\mathcal{T}$ is a union of unions of elements of $\mathcal{B}$, and therefore is itself a union of elements of $\mathcal{B}$; thus $\mathcal{T}$ is closed under arbitrary unions.

Finally, to show that $\mathcal{T}$ is closed under finite intersections, suppose first that $U_1, U_2 \in \mathcal{T}$. Then, for any $x \in U_1 \cap U_2$, the basis criterion says that there exist basis elements $B_1, B_2 \in \mathcal{B}$ such that $x \in B_1 \subset U_1$ and $x \in B_2 \subset U_2$ (Figure 2.6). But then part (ii) of the definition of basis guarantees that there exists $B_3 \in \mathcal{B}$ such that $x \in B_3 \subset B_1 \cap B_2 \subset U_1 \cap U_2$. Thus $U_1 \cap U_2$ satisfies the basis criterion, so it is again in $\mathcal{T}$. This shows that $\mathcal{T}$ is closed under pairwise intersections, and closure under finite intersections follows easily by induction. □

It often happens that we are given a particular topological space X together with a collection $\mathcal{B}$ of open subsets of X, and we would like to know whether $\mathcal{B}$ forms a basis for the topology of X. On the face of it, this would require showing first that $\mathcal{B}$ satisfies the two conditions in the definition of a basis, and then that each open subset of X is a union of elements of $\mathcal{B}$ (or alternatively satisfies the basis criterion). But as the following lemma shows, once we know that the sets in $\mathcal{B}$ are open subsets with respect to some topology, showing that they actually are a basis for that topology is much easier.

Lemma 2.11. *Suppose X is a topological space, and $\mathcal{B}$ is a collection of open subsets of X. If every open subset of X satisfies the basis criterion with respect to $\mathcal{B}$, then $\mathcal{B}$ is a basis for the topology of X.*

Proof. By Lemma 2.10, all we need to show is that $\mathcal{B}$ satisfies the two defining conditions for a basis.

For the first condition, since X itself is an open subset, the basis criterion tells us that any point $x \in X$ is in some element $B \in \mathcal{B}$; thus the union of all the elements of $\mathcal{B}$ is X.

For the second condition, suppose B_1 and B_2 are elements of $\mathcal{B}$ and $x \in B_1 \cap B_2$. The basis criterion applied to $B_1 \cap B_2$ (which is the intersection of two open subsets of X and therefore open) guarantees the existence of $B_3 \in \mathcal{B}$ such that $x \in B_3 \subset B_1 \cap B_2$. □

The next exercise describes bases for the topologies of Exercise 2.1. The preceding lemma makes the job of showing that they are indeed bases quite straightforward.

Exercise 2.15. In each of the following cases, prove that the given set $\mathcal{B}$ is a basis for the given topology.

(a) M is a metric space with the metric topology, and $\mathcal{B}$ is the collection of all open balls in M.

(b) X is a set with the discrete topology, and $\mathcal{B}$ is the collection of all one-point subsets of X.

(c) X is a set with the trivial topology, and $\mathcal{B} = \{X\}$.

Exercise 2.16. Show that each of the following collections $\mathcal{B}_i$ is a basis for the Euclidean topology on $\mathbb{R}^n$.

(a) $\mathcal{B}_1 = \{C_s(x) : x \in \mathbb{R}^n \text{ and } s > 0\}$, where $C_s(x)$ is the *open cube of side s around x*:

$$C_s(x) = \{y = (y_1, \ldots, y_n) : |x_i - y_i| < s/2,\ i = 1, \ldots, n\}.$$

(b) $\mathcal{B}_2 = \{B_r(x) :\ r \text{ is rational and } x \text{ has rational coordinates}\}$.

When we have a basis for a topology on Y, it is sufficient (and often much easier) to check continuity of maps into Y using only basis open sets, as the following lemma shows.

Lemma 2.12. *Let X and Y be topological spaces and let $\mathcal{B}$ be a basis for Y. A map $f\colon X \to Y$ is continuous if and only if for every basis open set $B \in \mathcal{B}$, $f^{-1}(B)$ is open in X.*

Proof. Clearly, if f is continuous, the inverse image of every basis open set is open. Conversely, suppose $f^{-1}(B)$ is open for every $B \in \mathcal{B}$. Let $U \subset Y$ be open, and let $x \in f^{-1}(U)$. By the basis criterion, there is a basis set B such that $f(x) \in B \subset U$. This implies that $x \in f^{-1}(B) \subset f^{-1}(U)$, which means that x has a neighborhood contained in $f^{-1}(U)$. Since this is true for every $x \in U$, U is the union of all these neighborhoods as x ranges over points of U, and therefore is open. □

Manifolds

We are almost ready to give the official definition of manifolds. We need just a few more preliminary definitions. The first one is easy, and captures very precisely the intuitive idea that a manifold should look "locally" like Euclidean space. Let X be a topological space. A topological space M is said to be *locally Euclidean of dimension* n if every point $q \in M$ has a neighborhood that is homeomorphic to an open subset of $\mathbb{R}^n$. Such a neighborhood is called a *Euclidean neighborhood* of q.

For some purposes, it is useful to be more specific about the kind of open subset we use to characterize locally Euclidean spaces. The next lemma shows that we could have replaced "open subset" by open ball or by $\mathbb{R}^n$ itself.

Lemma 2.13. *A topological space M is locally Euclidean of dimension n if and only if either of the following properties holds:*

(*a*) *Every point of M has a neighborhood homeomorphic to an open ball in $\mathbb{R}^n$.*

(*b*) *Every point of M has a neighborhood homeomorphic to $\mathbb{R}^n$.*

Proof. It is immediate that any space with property (a) or (b) is locally Euclidean of dimension n. Conversely, suppose M is locally Euclidean of dimension n. Because any open ball in $\mathbb{R}^n$ is homeomorphic to $\mathbb{R}^n$ itself (Example 2.4), properties (a) and (b) are equivalent, so we need only prove (a).

Given any point $q \in M$, let U be a neighborhood of q that admits a homeomorphism $\varphi\colon U \to V$, where V is an open subset of $\mathbb{R}^n$. The fact that V is open means that there is some open ball B around $\varphi(q)$ that is contained in V, and then $\varphi^{-1}(B)$ is a neighborhood of q homeomorphic to an open ball. □

If M is locally Euclidean of dimension n, a homeomorphism from an open subset $U \subset M$ to an open subset of $\mathbb{R}^n$ is called a *coordinate chart* (or just a *chart*) on U. We will call any open subset of M that is homeomorphic to a ball in $\mathbb{R}^n$ a *Euclidean ball* in M. (When M has dimension 2, we will sometimes use the term *Euclidean disk*.) The preceding lemma shows that every point in a locally Euclidean space has a Euclidean ball neighborhood.

The definition of locally Euclidean spaces makes sense even when $n = 0$. Since $\mathbb{R}^0$ is by convention a single point, Lemma 2.13(b) implies that a space is locally Euclidean of dimension 0 if and only if each point has a neighborhood homeomorphic to a one-point space, or in other words if and only if the space is discrete.

There are two other properties that we will include in the definition of manifolds, to rule out "pathological" spaces that might otherwise pass as manifolds.

Hausdorff Spaces

The first property we want to introduce ensures that there are "enough" open sets, so that neighborhoods behave more or less the way our intuition derived from Euclidean and metric spaces leads us to expect. For example, in a metric space, a one-point set $\{q\}$ is always closed, because around every point other than q there is a ball that does not include q. More generally, any two points in a metric space always have disjoint neighborhoods. However, these properties do not always hold in topological spaces. Consider the set $\{1, 2, 3\}$ with the topology $\{\varnothing, \{1\}, \{2, 3\}, \{1, 2, 3\}\}$ (Figure 2.1). In this case 2 and 3 do not have disjoint neighborhoods, since every open set that contains one also contains the other. Moreover, the set $\{2\}$ is not closed, because its complement is not open.

The problem with this example is that there are too few open sets, so neighborhoods do not have the same intuitive meaning they have in metric spaces. In our study of manifolds, we will want to rule out such "pathological" spaces, so we make the following definition. A topological space X is said to be a *Hausdorff space* if given any pair of distinct points $q_1, q_2 \in X$, there exist neighborhoods U_1 of q_1 and U_2 of q_2 with $U_1 \cap U_2 = \varnothing$. This property is often summarized by saying "open sets separate points."

Any metric space is Hausdorff. (If $d(q_1, q_2) = r$, then the open balls of radius $r/2$ around q_1 and q_2 are disjoint by the triangle inequality.) More generally, any open subset of a Hausdorff space is Hausdorff: If $V \subset X$ is open in the Hausdorff space X, and q_1, q_2 are distinct points in V, then in X there are open sets U_1, U_2 separating q_1 and q_2, and the sets $U_1 \cap V$ and $U_2 \cap V$ are open in V, disjoint, and contain q_1 and q_2, respectively.

Hausdorff spaces have many of the properties that we expect of metric spaces, such as those expressed in the following lemma.

Lemma 2.14. *Let X be a Hausdorff space.*

(*a*) *Every one-point set in X is closed.*

(*b*) *If a sequence $\{x_i\}$ in X converges to a limit $x \in X$, the limit is unique.*

Proof. For part (a), choose any $q_0 \in X$. Given $p \neq q_0$, the Hausdorff property says that there exist disjoint neighborhoods U_p of q_0 and V_p of p. This means that the complement of $\{q_0\}$ is equal to the union of the open sets V_p as p ranges over $X \smallsetminus \{q_0\}$, which is open, so $\{q_0\}$ is closed.

To prove that limits are unique, suppose that x and x' are two distinct limits of the sequence $\{x_i\}$. By the Hausdorff property, there exist disjoint neighborhoods U of x and U' of x'. By definition of convergence, there exist N, N' such that $i \geq N$ implies $x_i \in U$ and $i \geq N'$ implies $x_i \in U'$. But since U and U' are disjoint, this is a contradiction when $i \geq \max(N, N')$. □

Exercise 2.17. Show that the only Hausdorff topology on a finite set is the discrete topology.

The non-Hausdorff example above involving $\{1, 2, 3\}$ is obviously contrived, and has little relevance to our study of manifolds. But in Problem 3-8 at the end of the next chapter you will see a space that would be a manifold except for the fact that it fails to be Hausdorff.

Second Countability

Whereas the Hausdorff property ensures that there are enough open sets, the next property we will introduce ensures that there are not too many. We say that a topological space is *second countable* if it admits a countable basis.

As the "second" in the name second countable suggests, there is also another weaker notion of countability. It relies on the following definition: If X is a space and $q \in X$, a collection $\mathcal{B}_q$ of neighborhoods of q is called a *neighborhood basis* at q if every neighborhood of q contains some $B \in \mathcal{B}_q$. X is said to be *first countable* if there exists a countable neighborhood basis at each point. Second countability implies first countability: Given a countable basis for X, the collection of basis open sets containing q is easily seen to be a countable neighborhood basis at q.

One of the ways in which second countability is often used is in reducing the number of open sets one needs to "cover" a space. If X is any topological space, a collection $\mathcal{U}$ of subsets of X is said to *cover* X, or to be a cover of X, if every point in X is in one of the sets of $\mathcal{U}$. An *open cover* of X is a collection of open sets that covers X. Given any cover $\mathcal{U}$, a *subcover* of $\mathcal{U}$ is a subset of $\mathcal{U}$ that is still a cover.

Lemma 2.15. *If X is a second countable space, every open cover of X has a countable subcover.*

Proof. Let $\mathcal{B}$ be a countable basis for X, and let $\mathcal{U}$ be an arbitrary open cover of X. Let $\mathcal{B}'$ denote the subset of $\mathcal{B}$ consisting of those basis sets that are entirely contained in some element of $\mathcal{U}$. Because any subset of a countable set is countable, $\mathcal{B}'$ is a countable set.

Now, for each element $B \in \mathcal{B}'$, choose an element $U_B \in \mathcal{U}$ such that $B \subset U_B$ (this is possible by the way we defined $\mathcal{B}'$). The collection $\mathcal{U}' = \{U_B : B \in \mathcal{B}'\}$ is a countable subset of $\mathcal{U}$; the lemma will be proved if we can show that it still covers X.

If $x \in X$ is arbitrary, then $x \in U_0$ for some open set $U_0 \in \mathcal{U}$. By the basis criterion for U_0, there is some $B \in \mathcal{B}$ such that $x \in B \subset U_0$. This means, in particular, that $B \in \mathcal{B}'$, and therefore there is a set $U_B \in \mathcal{U}'$ such that $x \in B \subset U_B$. This shows that $\mathcal{U}'$ is a cover and completes the proof. □

Most "reasonable" spaces are second countable. For example, it follows from Exercise 2.16(b) above that $\mathbb{R}^n$ is second countable. Moreover, any open subset U of a second countable space X is second countable: Starting with a countable basis for X, just throw away all the elements of the basis that do not lie in U; then it is easy to check that the remaining basis sets form a countable basis for the topology of U.

In Problems 3-7 and 4-6 we will see examples of spaces that would be manifolds except for the failure of second countability.

Definition of Manifolds

We come now to the culmination of this chapter: the official definition of manifolds.

An *n-dimensional topological manifold* is a second countable Hausdorff space that is locally Euclidean of dimension n. Since the only kind of manifolds we will be considering in this book are topological manifolds, we will usually simply call them *n-dimensional manifolds*, or *n-manifolds*, or even just *manifolds* if the dimension is understood or irrelevant. (The term "topological manifold" is usually used only to emphasize that the kind of manifold under consideration is the kind we have defined here, which is a topological space with special properties, rather than other kinds of manifolds that can be defined, such as "smooth manifolds" or "complex manifolds." We will not treat any of these other kinds of manifolds in this book.)

A shorthand notation that is in common use is to write "let M^n be a manifold" to mean "let M be a manifold of dimension n." The superscript n is not part of the name of the manifold, and is usually dropped after the first time the manifold is introduced. One must be a bit careful to distinguish this notation from the n-fold Cartesian product $M^n = M \times \cdots \times M$, but it is usually clear from the context which is meant. We will not use this shorthand in this book, but you should be aware of it because you will encounter it in your reading.

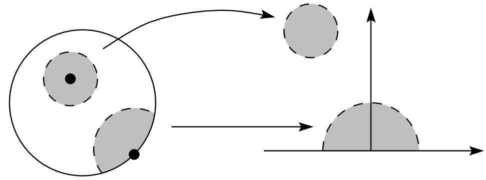


FIGURE 2.7. A manifold with boundary.

The most obvious example of an n-manifold is $\mathbb{R}^n$ itself. More generally, any open subset of $\mathbb{R}^n$—or in fact of any n-manifold—is again an n-manifold, as the next lemma shows.

Lemma 2.16. *Any open subset of an n-manifold is an n-manifold.*

Proof. Let M be an n-manifold, and let V be an open subset of M. Any $q \in V$ has a neighborhood (in M) that is homeomorphic to an open subset of $\mathbb{R}^n$; the intersection of that neighborhood with V is still open, still homeomorphic to an open subset of $\mathbb{R}^n$, and lies in V, so V is locally Euclidean. We remarked above that any open subset of a Hausdorff space is Hausdorff and any open subset of a second countable space is second countable. Therefore V is an n-manifold. □

In the next few chapters we will develop many more examples of manifolds.

Manifolds with Boundary

For some purposes it is useful also to have the following more general notion. An n-dimensional *manifold with boundary* is a second countable Hausdorff space in which every point has a neighborhood homeomorphic to an open subset of the n-dimensional *upper half space* $\mathbb{H}^n = \{(x_1, \dots, x_n) \in \mathbb{R}^n : x_n \geq 0\}$. Just as in the case of manifolds, we will call any homeomorphism from an open subset U of M to an open subset of $\mathbb{H}^n$ a *chart* on U.

Example 2.17 (Manifolds with Boundary). The upper half space $\mathbb{H}^n$ itself is obviously a manifold with boundary, as is any closed interval in $\mathbb{R}$, any closed disk in $\mathbb{R}^2$, or in fact a closed ball in any Euclidean space (see Figure 2.7). (This is not hard to see intuitively. You can probably construct appropriate charts yourself, or you can wait until Chapter 3 and use the ones suggested in Problem 3-4.)

The boundary of $\mathbb{H}^n$ in $\mathbb{R}^n$ is the set of points where $x_n = 0$. If M is a manifold with boundary, a point that is in the inverse image of $\partial\mathbb{H}^n$ under some chart is called a *boundary point* of M, and a point that is in the inverse image of $\operatorname{Int}\mathbb{H}^n$ is called an *interior point*. The *boundary* of M (the set of all its boundary points) is denoted by ∂M; similarly, its *interior* is denoted by $\operatorname{Int} M$.

Note that this use of the terms "boundary" and "interior" is distinct from their use earlier in this chapter in reference to subsets of topological spaces: If M is a manifold with boundary, its boundary as a subset of itself is always empty, even though its boundary as a manifold with boundary may not be. Usually the distinction will be clear from the context, but if necessary we can always distinguish the two meanings by referring to the *topological boundary* or the *manifold boundary* as appropriate.

There is a subtlety about this definition that might not be immediately evident. Although the interior and boundary of a manifold with boundary M are well-defined subsets, and it might seem intuitively rather obvious that they are disjoint sets, we have no way of proving at this stage that a point of M cannot be simultaneously both a boundary point and an interior point. After we have developed some more machinery, you will be asked to prove this fact (for the 1-dimensional case in Problem 4-14, the 2-dimensional case in Problem 8-6, and the general case in Problem 13-9). Nonetheless, we will go ahead and assume it when convenient.

Since any open ball in $\mathbb{R}^n$ is homeomorphic to an open subset of $\mathbb{H}^n$, an n-manifold is automatically an n-manifold with boundary (with empty boundary). But the converse is not true: For example, an endpoint of a closed interval has no Euclidean neighborhood. Assuming the (as yet unproved) fact that a boundary point cannot be an interior point, it follows that a manifold with boundary is a manifold if and only if its boundary is empty.

Problems

2-1. Let $(X_1, \mathcal{T}_1)$ and $(X_2, \mathcal{T}_2)$ be topological spaces and let $f\colon X_1 \to X_2$ be a bijective map. Show that f is a homeomorphism if and only if $f(\mathcal{T}_1) = \mathcal{T}_2$ in the sense that $U \in \mathcal{T}_1$ if and only if $f(U) \in \mathcal{T}_2$. [This shows, roughly speaking, that the topology is *precisely* the information preserved by homeomorphisms, and justifies the definition of topological spaces as the right setting for studying properties preserved by homeomorphisms.]

2-2. Suppose X is a set, and $\mathcal{B}$ is any collection of subsets of X whose union equals X. Let $\mathcal{T}$ be the collection of all unions of finite intersections of elements of $\mathcal{B}$. (Note that the empty set is the union of the empty collection of sets.)

(a) Show that $\mathcal{T}$ is a topology. (It is called the *topology generated by* $\mathcal{B}$, and $\mathcal{B}$ is called a *subbasis* for $\mathcal{T}$.)

(b) Show that $\mathcal{T}$ is the "smallest" topology for which all the sets in $\mathcal{B}$ are open; more precisely, show that $\mathcal{T}$ is the intersection of all topologies containing $\mathcal{B}$.

2-3. Let X be an infinite set. Consider the following collections of subsets of X:

$$\begin{aligned}
\mathcal{T}_1 &= \{U \subset X : X \smallsetminus U \text{ is finite or is all of } X\};\\
\mathcal{T}_2 &= \{U \subset X : X \smallsetminus U \text{ is infinite or is empty}\};\\
\mathcal{T}_3 &= \{U \subset X : X \smallsetminus U \text{ is countable or is all of } X\}.
\end{aligned}$$

For each collection, determine whether it is a topology.

2-4. Let $X = \{1, 2, 3\}$. Give a list of topologies on X such that any topology on X is homeomorphic to exactly one on your list.

2-5. Let $X = \mathbb{R}^2$ as a set, but with the topology determined by the following basis:

$$\mathcal{B} = \{\text{sets of the form } \{(c, y) : a < y < b\}, \text{ for fixed } a, b, c \in \mathbb{R}\}.$$

Determine which (if either) of the identity maps $X \to \mathbb{R}^2$, $\mathbb{R}^2 \to X$ is continuous.

2-6. Let X be a discrete space, Y be a space with the trivial topology, and Z be any topological space. Show that any maps $f\colon X \to Z$ and $g\colon Z \to Y$ are continuous. If Z is Hausdorff, show that the only continuous maps $h\colon Y \to Z$ are constant maps.

2-7. Give examples of maps between subsets of the plane (with the Euclidean topology) that are

(a) open but not closed or continuous;

(b) closed but not open or continuous;

(c) continuous but neither open nor closed;

(d) continuous and open but not closed;

(e) continuous and closed but not open;

(f) open and closed but not continuous.

2-8. Let $f\colon X \to Y$ be a continuous map between topological spaces, and let $\mathcal{B}$ be a basis for the topology of X. Let $f(\mathcal{B})$ denote the collection $\{f(B) : B \in \mathcal{B}\}$ of subsets of Y. If f is surjective and open, show that $f(\mathcal{B})$ is a basis for the topology of Y.

2-9. Suppose we are given an indexed collection of nonempty topological spaces $\{X_\alpha\}_{\alpha\in A}$. Declare a subset of the disjoint union $\coprod_{\alpha\in A} X_\alpha$ to be open if and only if its intersection with each X_α is open.

(a) Show that this is a topology on $\coprod_{\alpha\in A} X_\alpha$, called the *disjoint union topology.*

(b) Show that a subset of the disjoint union is closed if and only if its intersection with each X_α is closed.

(c) If each X_α is an n-manifold, show that the disjoint union $\coprod_{\alpha\in A} X_\alpha$ is an n-manifold if and only if the index set A is countable.

2-10. Suppose X is locally Euclidean of dimension n, and $f\colon X \to Y$ is a surjective local homeomorphism. Show that Y is also locally Euclidean of dimension n.

2-11. Show that a topological space is a 0-manifold if and only if it is a countable discrete space.

2-12. Let X be a totally ordered set (see the Appendix), and assume that X has at least two elements. For any $a \in X$, define sets $L(a), R(a) \subset X$ by

$$L(a) = \{c \in X : c < a\},$$
$$R(a) = \{c \in X : c > a\}.$$

Give X the topology generated by the subbasis $\{L(a), R(a) : a \in X\}$, called the *order topology.*

(a) Show that each set of the form (a, b) is open in X and each set of the form $[a, b]$ is closed (where (a, b) and $[a, b]$ are defined just as in $\mathbb{R}$).

(b) Show that X is Hausdorff.

(c) Show that for any $a, b \in X$, $\overline{(a,b)} \subset [a,b]$. Under what conditions does equality hold?

2-13. Let X be a second countable topological space. Show that every collection of disjoint open subsets of X is countable.

2-14. Show that locally Euclidean spaces and metric spaces are first countable.

2-15. (a) Show that every second countable space has a countable dense subset.

(b) Show that a metric space is second countable if and only if it has a countable dense subset.

2-16. Let X be a first countable space.

(a) For any set $A \subset X$ and any point $p \in X$, show that $p \in \overline{A}$ if and only if there is a sequence $\{p_n\}_{n=1}^{\infty}$ in A such that $p_n \to p$.

(b) Show that for any space Y, a map $f\colon X \to Y$ is continuous if and only if f takes convergent sequences in X to convergent sequences in Y.

2-17. Show that any manifold has a basis of Euclidean balls.

2-18. Suppose M is an n-dimensional manifold with boundary. Show that $\operatorname{Int} M$ is an n-manifold and ∂M is an $(n-1)$-manifold (without boundary).

3

New Spaces from Old

In this chapter we introduce three standard ways of constructing new topological spaces from given ones: subspaces, product spaces, and quotient spaces. We will explore how various topological properties are affected by these constructions, and we will show how each topology is characterized by which maps it makes continuous. At the end of the chapter we will explore in some detail the quotient spaces that arise from group actions. Throughout the chapter we will use these tools to build new examples of manifolds.

Subspaces

We have seen a number of examples of topological spaces that are subsets of $\mathbb{R}^n$, with the topology induced by the Euclidean metric. We have also seen that open subsets of a topological space inherit a topology from the containing space. It turns out that *arbitrary* subsets of topological spaces can also be viewed as topological spaces in their own right.

Let X be a space, and let $A \subset X$ be any subset. We define a topology $\mathcal{T}_A$ on A by

$$\mathcal{T}_A = \{U \subset A : U = A \cap V \text{ for some open set } V \subset X\}.$$

In other words, the open sets of $\mathcal{T}_A$ are the intersections with A of the open sets of X (Figure 3.1).

Exercise 3.1. Prove that $\mathcal{T}_A$ is a topology on A.

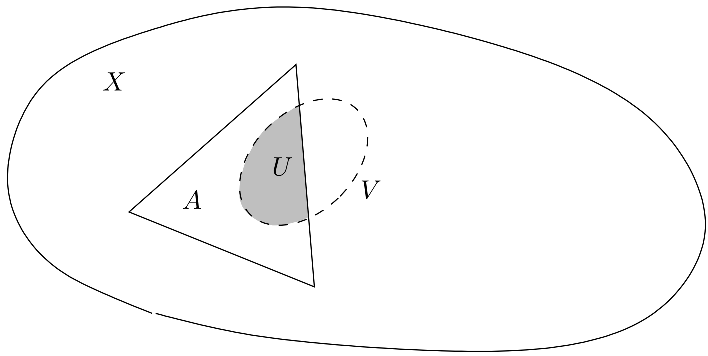


FIGURE 3.1. An open set in the subspace A.

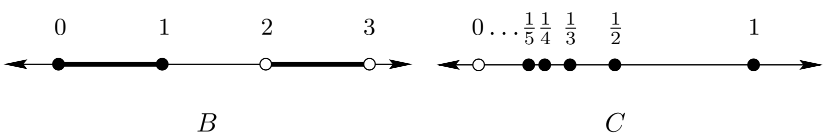


FIGURE 3.2. Subspaces of $\mathbb{R}$.

The topology $\mathcal{T}_A$ is called the *subspace topology* (or sometimes the *relative topology*) on A. A subset of a topological space X is called a *subspace* of X if it is endowed with the subspace topology. Henceforth, whenever we mention a subset of a topological space, we will always consider it as a topological space with the subspace topology unless otherwise specified.

Exercise 3.2. Let M be a metric space, and let $A \subset M$ be any subset. Show that the subspace topology on A is the same as the metric topology obtained by restricting the metric of M to points in A.

Example 3.1. Consider the subspaces $B = [0,1] \cup (2,3)$ and $C = \{1/n\}_{n=1}^{\infty}$ of $\mathbb{R}$ (Figure 3.2). Notice that the set $[0,1]$ is not open in $\mathbb{R}$. But it *is* an open subset of B, because $[0,1]$ is the intersection with B of the open interval $(-1,2)$. In C, the one-point sets $\{1/n\}$ are all open (why?), so the subspace topology on B is discrete.

These examples illustrate that openness and closedness are not properties of a set by itself, but rather of a set in relation to a particular topological space.

An injective continuous map that is a homeomorphism onto its image (in the subspace topology) is called a *topological embedding*. If $f\colon A \to X$ is such a map, we can think of the image set $f(A)$ as a homeomorphic copy of A embedded in X.

Example 3.2. Let $a\colon [0,1) \to \mathbb{R}^2$ be the map $a(s) = (\cos 2\pi s, \sin 2\pi s)$. In Exercise 2.7, you showed that a is not a homeomorphism onto its image in the subspace topology (which is the same as the metric topology by Exercise 3.2), so it is an example of a map that is continuous and injective but not an embedding. However, the restriction of a to any interval $[0, b)$ for $0 < b < 1$ is an embedding.

The first property we will prove about the subspace topology is so fundamental that, in a sense we will explain later, it completely characterizes the subspace topology among all the possible topologies on a subset. For any subset $A \subset X$, $\iota_A\colon A \hookrightarrow X$ denotes the inclusion map of A into X (see the Appendix).

Theorem 3.3 (Characteristic Property of Subspace Topologies). *Suppose $A \subset X$ is a subspace. For any topological space Y, a map $f\colon Y \to A$ is continuous if and only if the following composite map from Y to X is continuous:*

$$Y \xrightarrow{f} A \xhookrightarrow{\iota_A} X.$$

Proof. Directly from the definitions of continuity and the subspace topology,

$$\begin{aligned}
&f\colon Y \to A \text{ is continuous}\\
&\iff \text{for all } U \underset{\text{open}}{\subset} A,\ f^{-1}(U) \underset{\text{open}}{\subset} Y\\
&\iff \text{for all } V \underset{\text{open}}{\subset} X,\ f^{-1}(V \cap A) \underset{\text{open}}{\subset} Y\\
&\iff \text{for all } V \underset{\text{open}}{\subset} X,\ (\iota_A \circ f)^{-1}(V) \underset{\text{open}}{\subset} Y\\
&\iff \iota_A \circ f\colon Y \to X \text{ is continuous.}
\end{aligned}$$

□

Proposition 3.4 (Other Properties of the Subspace Topology). *Suppose A is a subspace of the topological space X.*

(*a*) *The inclusion map $\iota_A\colon A \hookrightarrow X$ is continuous, and in fact is a topological embedding.*

(*b*) *If $f\colon X \to Y$ is continuous, then its restriction to A is continuous.*

(*c*) *If $f\colon X \to Y$ is continuous, then $f\colon X \to f(X)$ is continuous.*

(*d*) *The closed subsets of A are precisely the intersections of A with closed subsets of X.*

(*e*) *If $B \subset A$ is a subspace of A, then B is a subspace of X; in other words, the subspace topologies that B inherits from A and from X agree.*

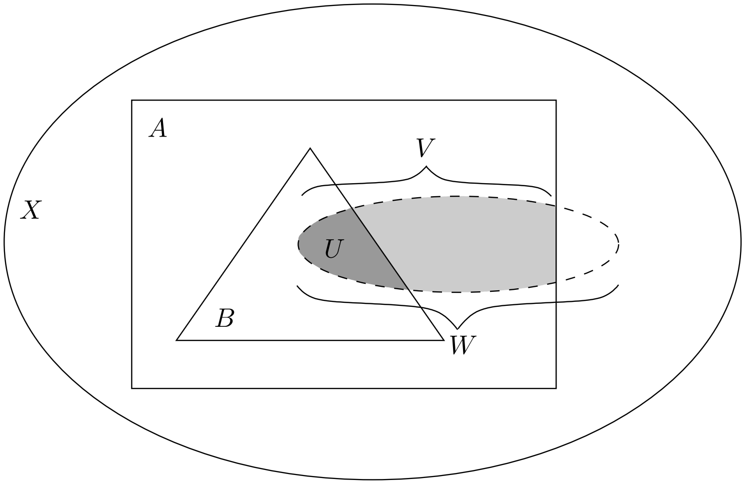


FIGURE 3.3. A subspace of a subspace.

(*f*) *If $B \subset A \subset X$, B is open in A, and A is open in X, then B is open in X.*

(*g*) *If $\mathcal{B}$ is a basis for the topology of X, then*

$$\mathcal{B}_A = \{B \cap A : B \in \mathcal{B}\}$$

is a basis for the topology of A.

(*h*) *Any subspace of a Hausdorff space is Hausdorff.*

(*i*) *Any subspace of a second countable space is second countable.*

Proof. Part (a) follows immediately from the characteristic property, just by taking Y to be equal to A and f to be the identity map. Then (b) follows from (a), since $f|_A = f \circ \iota_A$. Part (c) follows from the characteristic property because the hypothesis says that $\iota_{f(X)} \circ f \colon X \to Y$ is continuous, where $\iota_{f(X)}$ is the inclusion of $f(X)$ into Y.

For part (e), let $U \subset B$ be any subset. For the purposes of this proof, we say U *is open in B relative to A* if U is open in the subspace topology that B inherits from A; U *is open in B relative to X* is defined similarly. Then we argue as follows (Figure 3.3):

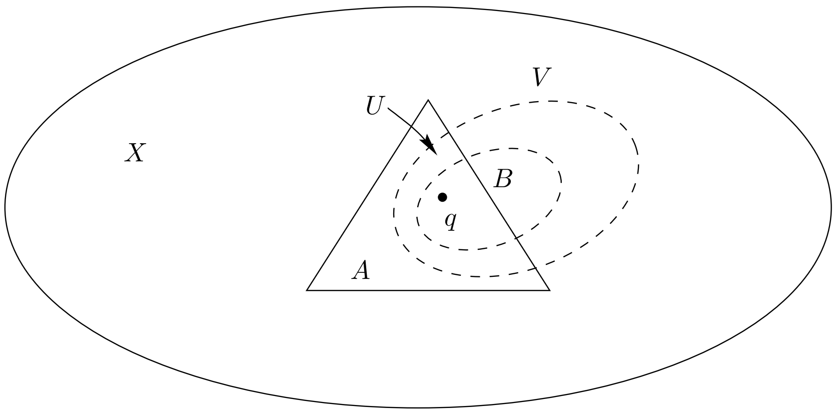


FIGURE 3.4. A basis open set for a subspace.

$$
\begin{aligned}
&U \text{ is open in } B \text{ relative to } A\\
&\quad\iff U = B \cap V \text{ for some } V \underset{\text{open}}{\subset} A\\
&\quad\iff U = B \cap V, \text{ where } V = A \cap W,\ W \underset{\text{open}}{\subset} X\\
&\quad\iff U = B \cap A \cap W \text{ for some } W \underset{\text{open}}{\subset} X\\
&\quad\iff U = B \cap W \text{ for some } W \underset{\text{open}}{\subset} X \text{ (since } B \subset A)\\
&\quad\iff U \text{ is open in } B \text{ relative to } X.
\end{aligned}
$$

To prove (g), we have to show that every open subset of A satisfies the basis criterion with respect to $\mathcal{B}_A$. Let U be an open subset of A, and let $q \in U$. Then by definition $U = A \cap V$ for some open subset $V \subset X$. By the basis criterion for V, there is an element $B \in \mathcal{B}$ such that $q \in B \subset V$ (Figure 3.4). It then follows that $q \in B \cap A \subset U$ with $B \cap A \in \mathcal{B}_A$.

Parts (d), (f), (h), and (i) are left as an exercise. $\square$

Exercise 3.3. Complete the proof of Proposition 3.4.

We can now produce many examples of manifolds as subspaces of Euclidean spaces. In particular, since the Hausdorff property and second countability are hereditary by parts (h) and (i) of the preceding proposition, to show that a subspace of $\mathbb{R}^n$ is a manifold we need only verify the locally Euclidean condition.

We begin with a very general construction.

Example 3.5. If $U \subset \mathbb{R}^n$ is an open set and $f\colon U \to \mathbb{R}^k$ is any continuous map, the *graph* of f (Figure 3.5) is the set

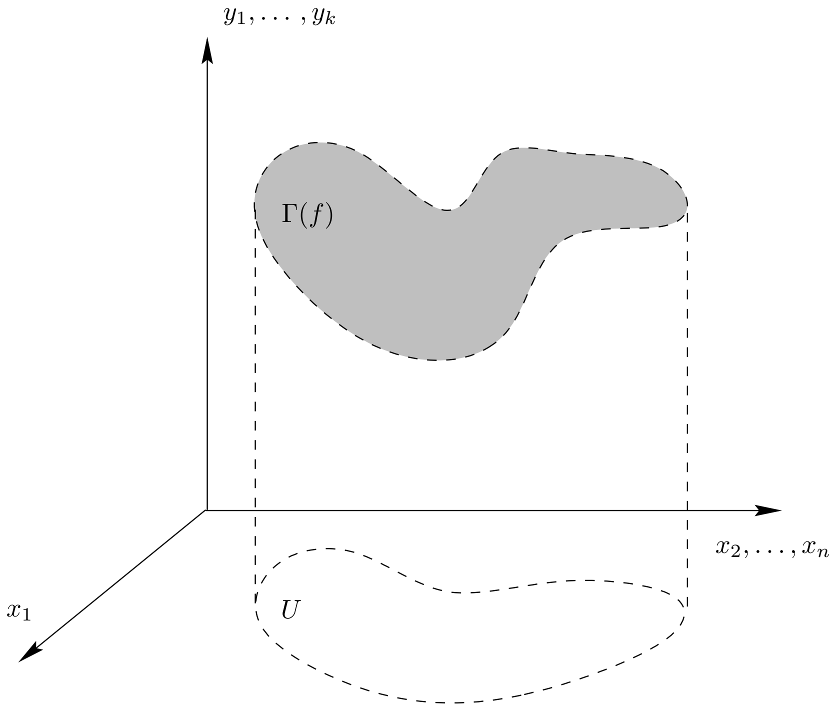


FIGURE 3.5. The graph of a continuous function.

$$\Gamma(f) = \{(x, y) = (x_1, \dots, x_n, y_1, \dots, y_k) \in \mathbb{R}^{n+k} : x \in U \text{ and } y = f(x)\}.$$

To verify that $\Gamma(f)$ is locally Euclidean, we construct an explicit homeomorphism between U and $\Gamma(f)$. Let $\Phi_f : U \to \Gamma(f)$ be the map

$$\Phi_f(x) = (x, f(x)).$$

It is continuous because f is, and it is easily seen that its inverse is the restriction to $\Gamma(f)$ of the map $\pi(x, y) = x$ (the projection onto the first n coordinates), which is continuous by Proposition 3.4(b). Thus Φ_f is a topological embedding. In particular, $\Gamma(f)$ is (globally) homeomorphic to the open set $U \subset \mathbb{R}^n$, so it is an n-manifold.

Example 3.6. Our next examples are arguably the most important manifolds of all, so it is worth taking some time to understand them well. The (*unit*) *n-sphere* is the set

$$\mathbb{S}^n = \{x \in \mathbb{R}^{n+1} : |x| = 1\}.$$

In low dimensions, spheres are easy to visualize: $\mathbb{S}^0$ is the two-point discrete space $\{\pm 1\} \subset \mathbb{R}$; $\mathbb{S}^1$ is the unit circle in the plane; and $\mathbb{S}^2$ is the familiar spherical surface of radius 1 in $\mathbb{R}^3$. In the case of the circle, it is often more convenient to identify the plane $\mathbb{R}^2$ with the set $\mathbb{C}$ of complex numbers by the correspondence $(x, y) \leftrightarrow x + iy$, and think of the circle as the set of complex numbers with unit modulus:

$$\mathbb{S}^1 = \{z \in \mathbb{C} : |z| = 1\}.$$

To see that $\mathbb{S}^n$ is a manifold, we need to show that each point has a Euclidean neighborhood. The most straightforward way is to show that each point has a neighborhood in which $\mathbb{S}^n$ is the graph of a continuous function. For each $i = 1, \dots, n+1$, let U_i^+ denote the subset of $\mathbb{R}^{n+1}$ where $x_i > 0$, and U_i^- the set where $x_i < 0$. If x is any point in $\mathbb{S}^n$, some coordinate x_i must be nonzero there, so the sets $U_1^\pm, \dots, U_{n+1}^\pm$ cover a neighborhood of $\mathbb{S}^n$. On $U_i^\pm$, we can solve the equation $|x| = 1$ for x_i and find that $x \in S^n \cap U_i^\pm$ if and only if

$$x_i = \pm\sqrt{1 - (x_1)^2 - \cdots - (x_{i-1})^2 - (x_{i+1})^2 - \cdots - (x_{n+1})^2}.$$

In other words, the portion of $\mathbb{S}^n$ in $U_i^\pm$ is the graph of a continuous function, and is therefore locally Euclidean. This proves that $\mathbb{S}^n$ is an n-manifold.

Here is another useful way to show that $\mathbb{S}^n$ is a manifold. Consider both $\mathbb{R}^n$ and $\mathbb{S}^n$ as subsets of $\mathbb{R}^{n+1}$ (identifying $\mathbb{R}^n$ with the set of points whose x_{n+1} coordinate is zero), and let $N = (0, \dots, 0, 1)$ denote the "north pole." Define *stereographic projection* $\sigma : \mathbb{S}^n \smallsetminus \{N\} \to \mathbb{R}^n$ to be the map given by

$$\sigma(x_1, \dots, x_{n+1}) = \frac{(x_1, \dots, x_n)}{1 - x_{n+1}}.$$

Geometrically, it sends a point $x \in \mathbb{S}^n \smallsetminus \{N\}$ to the point $u \in \mathbb{R}^n$ where the line from N to x intersects $\mathbb{R}^n$ (Figure 3.6), as you can easily check. It is a homeomorphism, because it has an inverse given by

$$\sigma^{-1}(u_1, \dots, u_n) = \frac{(2u_1, \dots, 2u_n, |u|^2 - 1)}{|u|^2 + 1}.$$

Thus $\mathbb{S}^n \smallsetminus \{N\}$ is homeomorphic to $\mathbb{R}^n$. In particular, this provides a Euclidean neighborhood of every point of $\mathbb{S}^n$ except N, and the analogous projection from the south pole works in a neighborhood of N.

Example 3.7. Finally, consider the *doughnut surface* D, which is the surface of revolution in $\mathbb{R}^3$ defined by revolving the circle $(y-2)^2 + z^2 = 1$ (called the *generating circle*) around the z-axis (Figure 3.7). It is characterized by the equation $(r-2)^2 + z^2 = 1$, where $r = \sqrt{x^2 + y^2}$. This surface can be parametrized by two angles θ (measured around the z axis from the

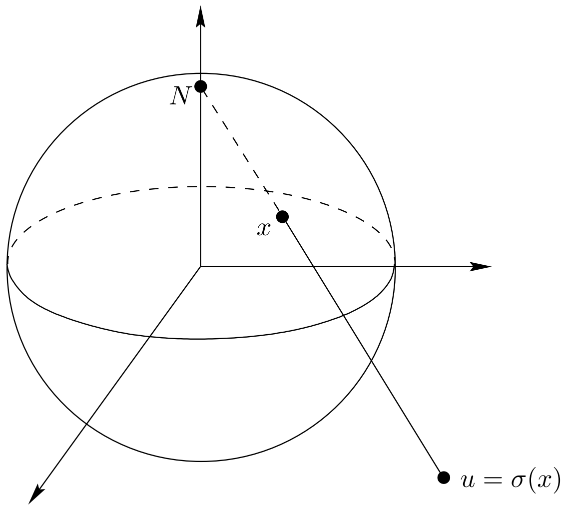


FIGURE 3.6. Stereographic projection.

xz-plane) and φ (measured around the generating circle from the horizontally outward direction). It is more convenient for calculations to make the substitutions $\varphi = 2\pi u$ and $\theta = 2\pi v$, and define a map $F\colon \mathbb{R}^2 \to D$ by

$$F(u,v) = ((2+\cos 2\pi u)\cos 2\pi v, (2+\cos 2\pi u)\sin 2\pi v, \sin 2\pi u). \tag{3.1}$$

This maps the plane onto D. It is not one-to-one, since $F(u+k, v+l) = F(u,v)$ for any pair of integers (k,l). However, F is injective if it is restricted to a small enough neighborhood of any point (u_0, v_0), and a straightforward calculation shows that a local inverse in a neighborhood of (u_0, v_0) can be constructed from the formulas

$$u = \frac{1}{2\pi}\tan^{-1}\frac{z}{r-2} + k; \qquad v = \frac{1}{2\pi}\tan^{-1}\frac{y}{x} + l;$$
$$u = \frac{1}{2\pi}\cot^{-1}\frac{r-2}{z} + k; \qquad v = \frac{1}{2\pi}\cot^{-1}\frac{x}{y} + l$$

for suitable choices of k, l. Thus D is a 2-manifold.

The next lemma is similar to the local criterion for continuity of Lemma 2.2, in that it asserts the global continuity of a map that is known to be continuous on certain subsets. In this case, however, the subsets must be closed, and there can be only finitely many of them. This lemma will turn out to be extremely useful in our investigations of surfaces.

Lemma 3.8 (Gluing Lemma). *Let X be a topological space, and suppose $X = A_1 \cup \cdots \cup A_k$, where each A_i is closed in X. For each i, let*

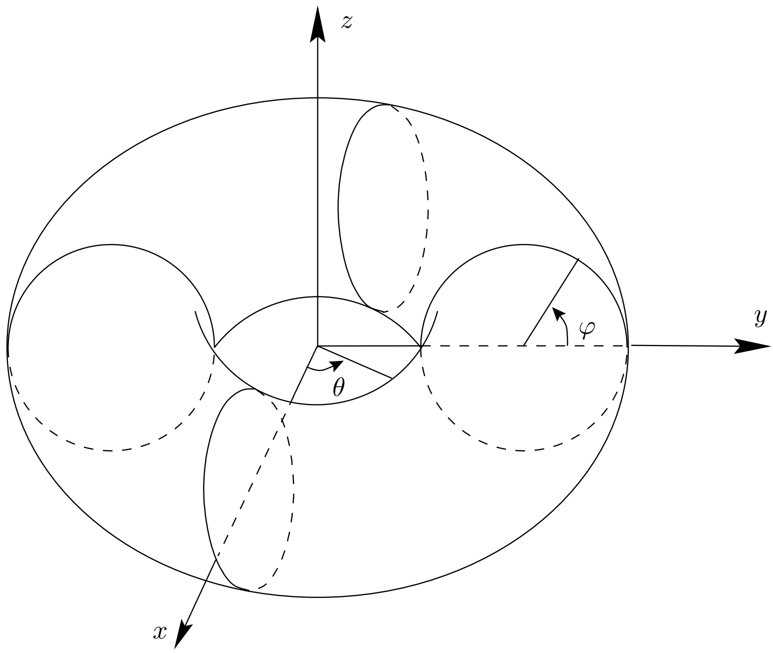


FIGURE 3.7. A doughnut surface of revolution.

$f_i\colon A_i \to Y$ be a continuous map such that $f_i|_{A_i\cap A_j} = f_j|_{A_i\cap A_j}$. There exists a unique continuous map $f\colon X \to Y$ such that $f|_{A_i} = f_i$ for each i.

Exercise 3.4. Prove Lemma 3.8.

In choosing a topology for a subset $A \subset X$, there are two competing priorities: We would like the inclusion map $A \hookrightarrow X$ to be continuous (from which it follows by composition that the restriction to A of any continuous map $f\colon X \to Y$ is continuous); and we would also like continuous maps into X whose images happen to lie in A also to be continuous as maps into A. For the first requirement, A needs to have enough open sets, while for the second it should not have too many. The subspace topology is chosen as the optimal compromise between these requirements.

As we will see several times in this chapter, natural topologies like the subspace topology can usually be characterized in terms of which maps are continuous with respect to them. This is why the "characteristic property" of the subspace topology (Theorem 3.3) is so named. The next lemma makes this precise.

Theorem 3.9 (Uniqueness of the Subspace Topology). *Suppose A is a subset of a topological space X. The subspace topology on A is the unique topology for which the characteristic property holds.*

Proof. Suppose we are given an arbitrary topology on A that is known to satisfy the characteristic property. For this proof, let A_g denote A with the given topology, and let A_s denote A with the subspace topology. To show that the given topology is equal to the subspace topology, it suffices to show that the identity map of A is a homeomorphism between A_g and A_s, by Problem 2-1.

First we note that the inclusion map from A_g into X is continuous, as follows: Since the identity map of any space is continuous, the characteristic property applied to the composition

$$A_g \xrightarrow{\text{Id}} A_g \hookrightarrow X$$

implies that this composite map is continuous; but this composition is just the inclusion $A_g \hookrightarrow X$ itself. Of course, the inclusion map $A_s \hookrightarrow X$ is also continuous, because it is the inclusion map of a subspace (Proposition 3.4(a)).

Now consider the two composite maps

$$A_s \xrightarrow{\text{Id}_{sg}} A_g \overset{\iota_g}{\hookrightarrow} X,$$
$$A_g \xrightarrow{\text{Id}_{gs}} A_s \overset{\iota_s}{\hookrightarrow} X.$$

Here both Id_{gs} and Id_{sg} represent the identity map of A, and ι_s and ι_g represent inclusion of A into X; we decorate them with subscripts only for the purpose of discussing their continuity.

Note that $\iota_g \circ \text{Id}_{sg} = \iota_s$, and $\iota_s \circ \text{Id}_{gs} = \iota_g$, both of which we have just shown to be continuous. Thus, applying the characteristic property to each of the compositions above, we conclude that both Id_{sg} and its inverse Id_{gs} are continuous. Therefore, Id_{sg} is a homeomorphism. □

Product Spaces

Suppose $X_1, \dots, X_n$ are topological spaces. We define a basis in their Cartesian product $X_1 \times \cdots \times X_n$ by

$$\mathcal{B} = \{U_1 \times \cdots \times U_n : U_i \text{ is open in } X_i,\ i = 1, \dots, n\}.$$

Exercise 3.5. Prove that $\mathcal{B}$ is a basis.

The topology generated by $\mathcal{B}$ is called the *product topology*, and the space $X_1 \times \cdots \times X_n$ with the product topology is called a *product space*. The basis open sets of the form $U_1 \times \cdots \times U_n$ are called *product open sets*.

For example, in the plane $\mathbb{R}^2 = \mathbb{R} \times \mathbb{R}$, the product topology is generated by sets of the form $I \times J$, where I and J are open sets in $\mathbb{R}$. A typical such set is an open rectangle.

Exercise 3.6. Show that the product topology on $\mathbb{R}^n = \mathbb{R} \times \cdots \times \mathbb{R}$ is the same as the metric topology induced by the Euclidean distance function.

The product topology can also be defined in the more general setting of infinite products, with a slightly more complicated definition. We will not need to use infinite products in this book, but the general definition of the product topology is given in Problem 7-12. For more information, consult [Sie92] or [Mun75].

The characteristic property relates continuity of a map into a product space to continuity of its component functions. In the special case of a map from $\mathbb{R}^n$ to $\mathbb{R}^m$, this reduces to a familiar result from advanced calculus.

Theorem 3.10 (Characteristic Property of Product Topologies). *Let $X_1 \times \cdots \times X_n$ be a product space. For any topological space B, a map $f\colon B \to X_1 \times \cdots \times X_n$ is continuous if and only if each of its component functions $f_i = \pi_i \circ f$ is continuous:*

$$\begin{tikzcd} & X_1 \times \cdots \times X_n \arrow[d, "\pi_i"] \\ B \arrow[ur, "f"] \arrow[r, "f_i"'] & X_i. \end{tikzcd}$$

Proof. Since it suffices to check continuity on basis open sets,

$$\begin{aligned} &f \text{ is continuous} \\ &\iff f^{-1}(U_1 \times \cdots \times U_n) \underset{\text{open}}{\subset} B \text{ for all } U_i \underset{\text{open}}{\subset} X_i \\ &\iff f_1^{-1}(U_1) \cap \cdots \cap f_n^{-1}(U_n) \underset{\text{open}}{\subset} B \text{ for all } U_i \underset{\text{open}}{\subset} X_i. \end{aligned}$$

If each f_i is continuous, the set in the last line above is the intersection of finitely many open sets and is therefore open in B, which shows that f is continuous. Conversely, if f is continuous, choose j between 1 and n and take $U_i = X_i$ for all i except $i = j$. Then $f_i^{-1}(U_i) = B$ when $i \neq j$, so the argument above shows that $f_j^{-1}(U_j)$ is open in B whenever U_j is open in X_j, or in other words, f_j is continuous. □

It follows from the characteristic property that if $f_1, f_2\colon X \to \mathbb{C}$ are complex-valued continuous functions, then their sum $(f_1+f_2)(x) = f_1(x)+f_2(x)$ is continuous, because it is the composition of the maps $f\colon X \to \mathbb{C}^2$ given by $f(x) = (f_1(x), f_2(x))$ and $s\colon \mathbb{C}^2 \to \mathbb{C}$ given by $s(w,z) = w + z$. A similar remark applies to the product $(f_1 f_2)(x) = f_1(x) f_2(x)$.

Just as in the case of the subspace topology, the product topology is uniquely determined by its characteristic property.

Theorem 3.11 (Uniqueness of the Product Topology). *Let X_1, $\ldots$, X_n be topological spaces. The product topology on $X_1 \times \cdots \times X_n$ is the unique topology that satisfies the characteristic property.*

Proof. Suppose that $X_1 \times \cdots \times X_n$ is endowed with some topology that satisfies the characteristic property. First we note that the projection maps π_i are continuous (in either topology) by the characteristic property applied to the identity map of $X_1 \times \cdots \times X_n$. Now inserting $X_1 \times \cdots \times X_n$ with the product topology in place of B shows that the identity map from the product topology to the given topology is continuous, and reversing roles shows that its inverse is also continuous. Thus the two topologies are equal. □

Proposition 3.12 (Other Properties of the Product Topology). *Let $X_1, \dots, X_n$ be topological spaces.*

(*a*) *The projection maps $\pi_i\colon X_1 \times \cdots \times X_n \to X_i$ are all continuous.*

(*b*) *The product topology is "associative" in the sense that the three product topologies $X_1 \times X_2 \times X_3$, $(X_1 \times X_2) \times X_3$, and $X_1 \times (X_2 \times X_3)$ on the set $X_1 \times X_2 \times X_3$ are all equal.*

(*c*) *For any i and any points $x_j \in X_j$, $j \neq i$, the map $f_i\colon X_i \to X_1 \times \cdots \times X_n$ given by*

$$f_i(x) = (x_1, \dots, x_{i-1}, x, x_{i+1}, \dots, x_n)$$

is a topological embedding of X_i into the product space.

(*d*) *If for each i, $\mathcal{B}_i$ is a basis for the topology of X_i, then the set*

$$\{B_1 \times \cdots \times B_n : B_i \in \mathcal{B}_i\}$$

is a basis for the product topology on $X_1 \times \cdots \times X_n$.

(*e*) *If A_i is a subspace of X_i for $i = 1, \dots, n$, the product topology and the subspace topology on $A_1 \times \cdots \times A_n \subset X_1 \times \cdots \times X_n$ are equal.*

(*f*) *If each X_i is Hausdorff, so is $X_1 \times \cdots \times X_n$.*

(*g*) *If each X_i is second countable, so is $X_1 \times \cdots \times X_n$.*

Exercise 3.7. Prove Proposition 3.12.

If $f_i\colon X_i \to Y_i$ are maps (continuous or not) for $i = 1, \dots, k$, their *product map* is

$$f_1 \times \cdots \times f_k\colon X_1 \times \cdots \times X_k \to Y_1 \times \cdots \times Y_k$$

given by

$$f_1 \times \cdots \times f_k(x_1, \dots, x_k) = (f_1(x_1), \dots, f_k(x_k)).$$

Proposition 3.13. *A product of continuous maps is continuous, and a product of homeomorphisms is a homeomorphism.*

Proof. Because it suffices to check that the inverse images of basis open sets are open, the first claim follows from the fact that $(f_1 \times \cdots \times f_k)^{-1}(U_1 \times \cdots \times U_k)$ is just the product of the open sets $f_1(U_1), \ldots, f_k(U_k)$. The second claim follows from the first, because the inverse of a bijective product map is itself a product map. □

Product spaces provide us with another rich source of examples of manifolds. The key is the following proposition.

Proposition 3.14. *If $M_1, \ldots, M_k$ are manifolds of dimensions $n_1, \ldots, n_k$, respectively, the product space $M_1 \times \cdots \times M_k$ is a manifold of dimension $n_1 + \cdots + n_k$.*

Proof. Proposition 3.12 shows that the product is Hausdorff and second countable, so only the locally Euclidean property needs to be checked. Given any point $q = (q_1, \ldots, q_k) \in M_1 \times \cdots \times M_k$, for each i there exists a neighborhood U_i of q_i and a homeomorphism φ_i from U_i to an open subset of $\mathbb{R}^{n_i}$. By the preceding lemma, the product map $\varphi_1 \times \cdots \times \varphi_k$ is a homeomorphism from a neighborhood of q to an open set in $\mathbb{R}^{n_1 + \cdots + n_k}$. □

A particularly important example is the product manifold $\mathbb{T}^n = \mathbb{S}^1 \times \cdots \times \mathbb{S}^1$, which is an n-dimensional manifold called the *n-torus*. In particular, the 2-torus is usually just called the *torus*. Because $\mathbb{S}^1$ is a subspace of $\mathbb{R}^2$, $\mathbb{T}^2$ can be considered as a subspace of $\mathbb{R}^4$ by Proposition 3.12(e): It is just the set of points $(x_1, x_2, x_3, x_4) \in \mathbb{R}^4$ such that $(x_1)^2 + (x_2)^2 = 1$ and $(x_3)^2 + (x_4)^2 = 1$. As the next lemma shows, $\mathbb{T}^2$ is homeomorphic to a familiar surface.

Lemma 3.15. *The torus $\mathbb{T}^2$ is homeomorphic to the doughnut surface D of Example 3.7.*

Proof. The key idea is that both surfaces are parametrized by two angles. For D, the angles are $\varphi = 2\pi u$ and $\theta = 2\pi v$ as in (3.1); for $\mathbb{T}^2$, they are the angles in the two circles. Although one must be careful using angle functions because they cannot be defined continuously on a whole circle, with some care we can eliminate the angles altogether and come up with formulas that are manifestly continuous.

With this in mind, we write $x_1 = \cos\theta$, $x_2 = \sin\theta$, $x_3 = \cos\varphi$, $x_4 = \sin\varphi$. Substituting into (3.1) suggests defining a map $G\colon \mathbb{T}^2 \to D$ by

$$G(x_1, x_2, x_3, x_4) = ((2 + x_3)x_1,\ (2 + x_3)x_2,\ x_4).$$

This is obviously continuous, and a little algebra shows that G maps $\mathbb{T}^2$ into D. To see that it is a homeomorphism, just check that its inverse is given by

$$G^{-1}(x, y, z) = (x/r,\ y/r,\ r - 2,\ z),$$

where $r = \sqrt{x^2 + y^2}$ as in Example 3.7. □

Quotient Spaces

Our third technique for constructing new topological spaces from old ones is somewhat more involved than the preceding two. It is a way to identify some points in a given topological space with each other, to obtain a new, smaller space.

Let X be a topological space, Y be any set, and $\pi\colon X \to Y$ be a surjective map. Define a topology on Y by declaring a subset $U \subset Y$ to be open if and only if $\pi^{-1}(U)$ is open in X. This is called the *quotient topology* induced by the map π.

Exercise 3.8. Show that the quotient topology is indeed a topology.

More generally, if X and Y are topological spaces, a map $\pi\colon X \to Y$ is called a *quotient map* if it is surjective and continuous and Y has the quotient topology induced by π. If π is known to be surjective, this is the same as saying that U is open in Y if and only if $\pi^{-1}(U)$ is open in X.

An easy example to keep in mind is the map $\pi\colon \mathbb{R}^{n+k} \to \mathbb{R}^n$ given by projection onto the first n coordinates. It is straightforward to check directly from the definition that it is a quotient map.

The most common source of quotient maps is the following construction. Let $\sim$ be an equivalence relation on a topological space X (see the Appendix). For each $q \in X$ let $[q]$ denote the equivalence class of q, and let $X/{\sim}$ denote the set of equivalence classes: This is a *partition* of X, which is a decomposition of X into a collection of disjoint subsets whose union is X. Let $\pi\colon X \to X/{\sim}$ be the natural projection sending each element of X to its equivalence class. Then $X/{\sim}$ together with the quotient topology determined by π is called the *quotient space* (or sometimes *identification space*) of X by the given equivalence relation. The quotient map π is called the *projection*.

Alternatively, a quotient space can be defined by explicitly giving a partition of X. Whether a given quotient space is defined in terms of an equivalence relation or a partition is a matter of convenience.

If $\pi\colon X \to Y$ is a quotient map, a subset $U \subset X$ is said to be *saturated* (with respect to π) if $U = \pi^{-1}(V)$ for some subset $V \subset Y$. (In fact, you can check that U is saturated if and only if $U = \pi^{-1}(\pi(U))$.) If Y is a quotient space determined by an equivalence relation, the saturated sets are those that are unions of equivalence classes. More generally, for any quotient map $\pi\colon X \to Y$, a subset $\pi^{-1}(y) \subset X$ for $y \in Y$ is called a *fiber* of π; a saturated set is one that is a union of fibers.

Although quotient maps do not always take open sets to open sets, there is a useful alternative characterization of quotient maps in terms of saturated open sets.

Lemma 3.16. *A continuous surjective map $\pi\colon X \to Y$ is a quotient map if and only if it takes saturated open sets to open sets, or saturated closed sets to closed sets.*

Exercise 3.9. Prove Lemma 3.16.

Lemma 3.17. *Suppose $f\colon X \to Y$ is a quotient map. The restriction of f to any saturated open or closed set is a quotient map.*

Exercise 3.10. Prove Lemma 3.17.

It is not always a trivial matter to check that a continuous surjective map is a quotient map—it may well not be, as the following example shows.

Example 3.18. Consider the map $a\colon [0,1) \to \mathbb{S}^1$ defined (in complex notation) by $a(s) = e^{2\pi i s}$. It is surjective and continuous, but not a quotient map, because $[0, \frac{1}{2})$ is a saturated open subset of $[0,1)$ whose image is not open in $\mathbb{S}^1$.

The following lemma gives two very useful sufficient conditions for a surjective continuous map to be a quotient map.

Lemma 3.19. *If $\pi\colon X \to Y$ is a surjective continuous map that is also an open or closed map, then it is a quotient map.*

Proof. If π is open, it takes saturated open sets to open sets (because it takes *all* open sets to open sets). If π is closed, it takes saturated closed sets to closed sets. In either case, it is a quotient map by Lemma 3.16. □

One simple property of quotient maps is that they behave well with respect to composition.

Lemma 3.20 (Composition Property of Quotient Maps). *Suppose $\pi_1\colon X \to Y$ and $\pi_2\colon Y \to Z$ are quotient maps. Then their composition $\pi_2 \circ \pi_1\colon X \to Z$ is also a quotient map.*

Proof. Just note that $U \subset Z$ is open if and only if $\pi_2^{-1}(U)$ is open in Y, which is true if and only if $\pi_1^{-1}(\pi_2^{-1}(U)) = (\pi_2 \circ \pi_1)^{-1}(U)$ is open in X. □

As it happens, quotient spaces do not generally behave well with respect to products, subspaces, bases, or topological properties such as locally Euclidean, Hausdorff, or second countable. In particular, quotient spaces of manifolds are generally not manifolds. In fact, it is not difficult to construct

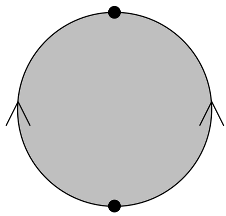


FIGURE 3.8. A quotient of $\overline{\mathbb{B}^2}$.

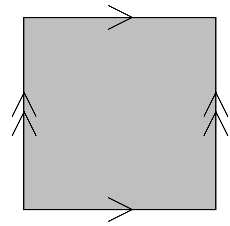


FIGURE 3.9. A quotient of $I \times I$.

a quotient space of a manifold that satisfies all the definitions of a manifold except that it is not Hausdorff (see Problem 3-8). Thus if we wish to prove that a given quotient space is a manifold, we have to prove at least the locally Euclidean and Hausdorff properties directly.

The following lemma shows that in many cases this is sufficient. In particular, it shows that a quotient of a manifold is again a manifold, provided that it is locally Euclidean and Hausdorff.

Lemma 3.21. *Suppose P is a second countable space and $\pi\colon P \to M$ is a quotient map. If M is locally Euclidean, it is second countable.*

Proof. Let $\mathcal{U}$ be a covering of M by Euclidean balls. The collection $\{\pi^{-1}(U) : U \in \mathcal{U}\}$ is an open cover of P, which has a countable subcover by Lemma 2.15. If we let $\mathcal{U}' \subset \mathcal{U}$ denote a countable subset of $\mathcal{U}$ such that $\{\pi^{-1}(U) : U \in \mathcal{U}'\}$ covers P, then $\mathcal{U}'$ is a countable cover of M by Euclidean balls. Each such ball has a countable basis, and the union of all these bases is a countable basis for M. □

Because quotient spaces are probably less familiar to you than subspaces or products, we will introduce a number of examples before going any farther.

Example 3.22. The map $\alpha\colon [0,1] \to \mathbb{S}^1$ given by $\alpha(s) = e^{2\pi i s}$ is a closed map and therefore a quotient map.

Example 3.23. Let $\overline{\mathbb{B}^2}$ denote the closed unit disk in $\mathbb{R}^2$. Let $\sim$ be the equivalence relation on $\overline{\mathbb{B}^2}$ generated by $(x,y) \sim (-x,y)$ for all $(x,y) \in \partial\mathbb{B}^2$ (Figure 3.8). (You can think of this space as being obtained from $\overline{\mathbb{B}^2}$ by "pasting" the left half of the boundary to the right half.) We will see in Chapter 6 that $\overline{\mathbb{B}^2}/\sim$ is homeomorphic to $\mathbb{S}^2$.

Example 3.24. Let $I = [0,1]$ denote the closed unit interval in the real line; we will generally just call this the *unit interval.* Define an equivalence relation on the square $I \times I$ by setting $(x,0) \sim (x,1)$ for all $x \in I$, and $(0,y) \sim (1,y)$ for all $y \in I$ (Figure 3.9). This can be visualized as the space obtained by gluing the top boundary segment of the square to the

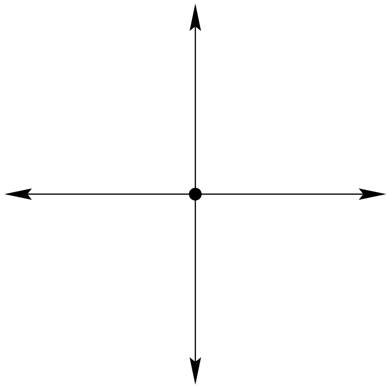


FIGURE 3.10. Wedge of two lines.

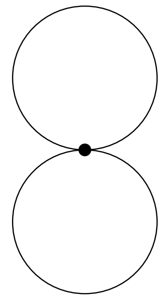


FIGURE 3.11. Wedge of two circles.

bottom to form a cylinder, and then gluing the left-hand boundary circle of the resulting cylinder to the right-hand one. Later we will see that the resulting quotient space is homeomorphic to the torus.

Example 3.25. Let $X_1, \dots, X_k$ be topological spaces, and let $q_i \in X_i$. The *wedge* of $X_1, \dots, X_k$ (also called their *one-point union*), written $X_1 \vee \cdots \vee X_k$, is the quotient space obtained from $X_1 \amalg \cdots \amalg X_k$ by identifying $q_1 \sim \cdots \sim q_k$. In other words, we glue the spaces together by identifying all their distinguished points together. For example, the wedge $\mathbb{R} \vee \mathbb{R}$ is homeomorphic to the union of the x-axis and the y-axis in the plane (Figure 3.10), and the wedge $\mathbb{S}^1 \vee \mathbb{S}^1$ is homeomorphic to the *figure eight space* $\mathcal{E}$ consisting of the union of the two circles of radius 1 centered at $(0,1)$ and $(0,-1)$ in the plane (Figure 3.11). A wedge of finitely many copies of $\mathbb{S}^1$ is sometimes called a *bouquet of circles*.

Example 3.26. Define an equivalence relation on $\mathbb{R}$ by declaring $x \sim y$ if x and y differ by an integer. We will see below that the resulting quotient space is homeomorphic to the circle.

Example 3.27. Consider the map $q\colon \mathbb{R}^{n+1} \smallsetminus \{0\} \to \mathbb{S}^n$ defined by $q(x) = x/|x|$. Observe that q is continuous and surjective, and the fibers of q are rays in $\mathbb{R}^{n+1} \smallsetminus \{0\}$. Thus the saturated sets are the unions of rays, and it is easy to check that q takes saturated open sets to open sets and is therefore a quotient map.

Example 3.28. Define $\mathbb{P}^n$, the *real projective space* of dimension n, to be the set of 1-dimensional linear subspaces (lines through the origin) in $\mathbb{R}^{n+1}$. There is a natural map $\pi\colon \mathbb{R}^{n+1} \smallsetminus \{0\} \to \mathbb{P}^n$ defined by sending a point x to its span. We topologize $\mathbb{P}^n$ by giving it the quotient topology with respect to this map.

This space can also be viewed in another way. If we define an equivalence relation on $\mathbb{R}^{n+1} \smallsetminus \{0\}$ by declaring two points x, y to be equivalent if $x = \lambda y$ for some nonzero real number λ, then there is an obvious identification between $\mathbb{P}^n$ and the set of equivalence classes. Under this identification, the map π defined above is just the map sending a point to its equivalence class.

The Characteristic Property of the Quotient Topology

The characteristic property of the quotient topology is even more important than those of the subspace or product topologies.

Theorem 3.29 (Characteristic Property of Quotient Topologies). *Let $\pi\colon X \to Y$ be a quotient map. For any topological space B, a map $f\colon Y \to B$ is continuous if and only if the composite map $f \circ \pi$ is continuous:*

$$\begin{tikzcd} X \arrow[d, "\pi"'] \arrow[dr, "f \circ \pi"] & \\ Y \arrow[r, "f"'] & B. \end{tikzcd}$$

Proof. If f is continuous, $f \circ \pi$ is continuous by composition. Conversely, if $f \circ \pi$ is continuous and $U \subset B$ is open, then $\pi^{-1}(f^{-1}(U)) = (f \circ \pi)^{-1}(U)$ is open in X, which implies $f^{-1}(U)$ is open in Y by definition of the quotient topology. Thus f is continuous. □

The characteristic property has the following extremely important corollary:

Corollary 3.30 (Passing to the Quotient). *Suppose $\pi\colon X \to Y$ is a quotient map, B is a topological space, and $f\colon X \to B$ is any continuous map that is constant on the fibers of π (i.e., if $\pi(p) = \pi(q)$, then $f(p) = f(q)$). Then there exists a unique continuous map $\widetilde{f}\colon Y \to B$ such that $f = \widetilde{f} \circ \pi$:*

$$\begin{tikzcd} X \arrow[d, "\pi"'] \arrow[dr, "f"] & \\ Y \arrow[r, "\widetilde{f}"'] & B. \end{tikzcd}$$

Proof. The existence and uniqueness of $\widetilde{f}$ follow from elementary set theory: Given $q \in Y$, there is some $p \in X$ such that $\pi(p) = q$, and we can set $\widetilde{f}(q) = f(p)$ for any such p. The hypothesis on f guarantees that $\widetilde{f}$ is unique and well-defined. Continuity of $\widetilde{f}$ is then immediate from the characteristic property. $\square$

In the situation of the preceding corollary, we say that the map f *passes to the quotient* or *descends to the quotient.*

In the case of quotient spaces, there are two slightly different ways of phrasing the uniqueness associated with the characteristic property. The first one says that the characteristic property uniquely characterizes quotient maps.

Theorem 3.31 (Characterization of Quotient Maps). *Let X and Y be topological spaces, and let $\pi\colon X \to Y$ be any surjective map. Then π is a quotient map if and only if the characteristic property holds.*

Proof. If π is a quotient map, the characteristic property holds by Theorem 3.29. Conversely, suppose π has the characteristic property. Applying the characteristic property to the diagram

$$\begin{tikzcd} X \arrow[d, "\pi"'] \arrow[dr, "\pi"] & \\ Y \arrow[r, "\mathrm{Id}"'] & Y \end{tikzcd}$$

shows that π is continuous because the identity is. To show that π is a quotient map, we will show that Y with the given topology is homeomorphic to Y with the quotient topology. As before, let Y_g and Y_q denote Y with the given and quotient topologies, respectively, and let Id_{gq}, Id_{qg}, π_q, and π_g have the obvious meanings. Then the characteristic property applied to the two diagrams

$$\begin{tikzcd} X \arrow[d, "\pi_g"'] \arrow[dr, "\pi_q"] & \\ Y_g \arrow[r, "\mathrm{Id}_{gq}"'] & Y_q \end{tikzcd} \qquad \begin{tikzcd} X \arrow[d, "\pi_q"'] \arrow[dr, "\pi_g"] & \\ Y_q \arrow[r, "\mathrm{Id}_{qg}"'] & Y_g \end{tikzcd}$$

shows that Id_{qg} and Id_{gq} are both continuous, from which the result follows. $\square$

The second uniqueness result says that quotient spaces are uniquely determined up to homeomorphism by the identifications made by their quotient maps.

Corollary 3.32 (Uniqueness of Quotient Spaces). *Suppose $\pi_1\colon X \to Y_1$ and $\pi_2\colon X \to Y_2$ are quotient maps that make the same identifications (i.e., $\pi_1(p) = \pi_1(q)$ if and only if $\pi_2(p) = \pi_2(q)$). Then there is a unique homeomorphism $\varphi\colon Y_1 \to Y_2$ such that $\varphi \circ \pi_1 = \pi_2$.*

Proof. By Corollary 3.30, both π_1 and π_2 pass uniquely to the quotient as in the following diagrams:

$$
\begin{tikzcd}
X \arrow[d, "\pi_1"'] \arrow[dr, "\pi_2"] & \\
Y_1 \arrow[r, "\widetilde{\pi_2}"'] & Y_2
\end{tikzcd}
\qquad
\begin{tikzcd}
X \arrow[d, "\pi_2"'] \arrow[dr, "\pi_1"] & \\
Y_2 \arrow[r, "\widetilde{\pi_1}"'] & Y_1.
\end{tikzcd}
$$

Since both diagrams above commute, it follows that $\widetilde{\pi_1} \circ (\widetilde{\pi_2} \circ \pi_1) = \widetilde{\pi_1} \circ \pi_2 = \pi_1$. Consider another diagram:

$$
\begin{tikzcd}
X \arrow[d, "\pi_1"'] \arrow[dr, "\pi_1"] & \\
Y_1 \arrow[r] & Y_1.
\end{tikzcd}
$$

If the bottom arrow is interpreted as either $\widetilde{\pi_1} \circ \widetilde{\pi_2}$ or the identity map of Y_1, this diagram will commute; by the uniqueness part of Corollary 3.30, these maps must be equal. Similarly, $\widetilde{\pi_2} \circ \widetilde{\pi_1}$ is the identity on Y_2. Thus $\varphi = \widetilde{\pi_2}$ is the required homeomorphism, and it is the unique such map by the uniqueness statement of Corollary 3.30. □

Group Actions

Our next construction is a far-reaching generalization of Examples 3.26 and 3.28. A *topological group* is a group G endowed with a topology such that the maps $\mu\colon G \times G \to G$ and $\iota\colon G \to G$ given by

$$\mu(g_1, g_2) = g_1 g_2, \qquad \iota(g) = g^{-1}$$

are continuous, where the product and inverse are those of the group structure of G. A *discrete group* is a topological group that has the discrete topology.

Example 3.33. Each of the following is a topological group.

- The real line $\mathbb{R}$ with its additive group structure and the Euclidean topology.
- The set $\mathbb{R}^* = \mathbb{R} \smallsetminus \{0\}$ of nonzero real numbers under multiplication, with the subspace topology.
- The set $\mathbb{C}^* = \mathbb{C} \smallsetminus \{0\}$ of nonzero complex numbers under complex multiplication, with the subspace topology.

- The *general linear group* $\mathrm{GL}(n,\mathbb{R})$, which is the set of $n\times n$ invertible real matrices under matrix multiplication, with the subspace topology inherited from $\mathbb{R}^{n^2}$.

- Any group with the discrete topology.

Exercise 3.11. Verify that each of the above examples is a topological group.

Lemma 3.34. *Any subgroup of a topological group is a topological group with the subspace topology. Any finite product of topological groups is a topological group with the direct product group structure and the product topology.*

Exercise 3.12. Prove Lemma 3.34.

In view of Lemma 3.34, each of the following is a topological group:

- Euclidean space $\mathbb{R}^n$ as a group under vector addition.

- The circle $\mathbb{S}^1 \subset \mathbb{C}^*$ under complex multiplication, with the subspace topology.

- The n-torus $\mathbb{T}^n = \mathbb{S}^1 \times \cdots \times \mathbb{S}^1$ with the direct product group structure.

- The orthogonal group $\mathrm{O}(n)$, which is the subgroup of $\mathrm{GL}(n,\mathbb{R})$ consisting of matrices A such that AA^t is the identity.

If G is a topological group and $g \in G$, the *left translation* map $L_g : G \to G$ defined by $L_g(g') = gg'$ is continuous, because it is the restriction of the multiplication map to $\{g\} \times G$. Because $L_g \circ L_{g^{-1}} = \mathrm{Id}_G$, left translation by any element of g is a homeomorphism of G. Similarly, *right translation* $R_g(g') = g'g$ is also a homeomorphism.

Suppose G is a group and X is a topological space. A *left action* of G on X is a map $G \times X \to X$, written $(g, x) \mapsto g \cdot x$, with the following properties:

(i) For any $x \in X$ and any $g_1, g_2 \in G$, $g_1 \cdot (g_2 \cdot x) = (g_1g_2) \cdot x$.

(ii) For all $x \in X$, $1 \cdot x = x$.

Similarly, a *right action* is a map $X \times G \to X$, written $(x, g) \mapsto x \cdot g$, with the same properties except that composition works in reverse: $(x \cdot g_1) \cdot g_2 = x \cdot (g_1g_2)$. Any right action determines a left action in a canonical way, and vice versa, by the correspondence

$$g \cdot x = x \cdot g^{-1}.$$

Thus for many purposes, the choice of left or right action is a matter of taste. We usually choose to focus on left actions because the composition law mimics composition of functions, and unless we specify otherwise, groups will always be understood to act on the left. However, we will see some situations in which an action appears naturally as a right action.

If G is a topological group, an action of G on a space X is said to be *continuous* if the map $G \times X \to X$ is continuous. This means, in particular, that for each $g \in G$ the map $x \mapsto g \cdot x$ is continuous from X to itself, because it is the restriction of the action to the subspace $\{g\} \times X \subset G \times X$. In fact, each such map is a homeomorphism, because the definition of a group action guarantees that it has a continuous inverse $x \mapsto g^{-1} \cdot x$. When G is discrete, it is easy to check that the action is continuous if and only if $x \mapsto g \cdot x$ is continuous for each $g \in G$.

For any $x \in X$, the set $G \cdot x = \{g \cdot x : g \in G\}$ is called the *orbit* of x. The action is said to be *transitive* if for every pair of points $x, y \in X$, there is a group element g such that $g \cdot x = y$ or equivalently if the only orbit is the entire space X. It is said to be *free* if the only element of G that has any fixed points is the identity, i.e., $g \cdot x = x$ for some x implies $g = 1$.

Example 3.35 (Continuous Group Actions).

(a) The general linear group $\mathrm{GL}(n, \mathbb{R})$ acts continuously on the left on $\mathbb{R}^n$ by matrix multiplication, each vector in $\mathbb{R}^n$ considered as a column matrix. Because any nonzero vector in $\mathbb{R}^n$ can be taken to any other by a linear transformation, there are only two orbits: $\mathbb{R}^n \smallsetminus \{0\}$ and $\{0\}$.

(b) The orthogonal group $\mathrm{O}(n)$ acts continuously on $\mathbb{R}^n$ by matrix multiplication as well; this is just the restriction of the action in part (a) to $\mathrm{O}(n) \times \mathbb{R}^n \subset \mathrm{GL}(n, \mathbb{R}) \times \mathbb{R}^n$. Since orthogonal transformations preserve lengths of vectors, and any vector can be taken to any other of the same length by an orthogonal transformation, the orbits are $\{0\}$ and the spheres centered at 0.

(c) The restriction of the action of $\mathrm{O}(n)$ to the unit sphere in $\mathbb{R}^n$ yields a transitive action on $\mathbb{S}^{n-1}$.

(d) The group $\mathbb{R}^*$ acts on $\mathbb{R}^n \smallsetminus \{0\}$ by scalar multiplication. The action is free, and the orbits are the lines through the origin (with the origin removed).

(e) Any topological group G acts freely and transitively on itself on the left by left translation: $g \cdot g' = L_g(g') = gg'$. Similarly, G acts on itself on the right by right translation.

Given an action of G on a space X, we define an equivalence relation on X by setting $x_1 \sim x_2$ if there is an element $g \in G$ such that $g \cdot x_1 = x_2$.

The equivalence classes are precisely the orbits of the group action. The resulting quotient space is denoted by X/G, and is called the *orbit space* of the action. If the action is transitive, the orbit space is a single point, so only nontransitive actions yield interesting examples.

Exercise 3.13. Verify that the real projective space $\mathbb{P}^n$ of Example 3.28 is the orbit space of the action of $\mathbb{R}^*$ on $\mathbb{R}^{n+1} \setminus \{0\}$ by scalar multiplication.

A particularly important special case arises when we consider a subgroup Γ of a topological group G (with the subspace topology). Group multiplication on the left or right defines a left or right action of Γ on G; it is just the restriction of the action of G on itself to $\Gamma \times G$ or $G \times \Gamma$. This action is continuous and free, but in general not transitive. An orbit of the right action of Γ on G is a set of the form $\{g\gamma : \gamma \in \Gamma\}$, which is precisely the left coset $g\Gamma$. Thus the orbit space of the right action of Γ on G is the set G/Γ of left cosets with the quotient topology. This quotient space is called the (*left*) *coset space* of G by Γ. (It is unfortunate but unavoidable that the right action produces a left coset space and vice versa. If G is abelian, the situation is simpler, because then the left action and right action of Γ are equal to each other.)

Example 3.36. As an application, let us consider the coset space $\mathbb{R}/\mathbb{Z}$. Because $\mathbb{Z}$ (with the discrete topology) is a subgroup of the topological group $\mathbb{R}$, there is a natural free continuous action of $\mathbb{Z}$ on $\mathbb{R}$ by translation: $n \cdot x = n + x$. (Because $\mathbb{R}$ is abelian, we might as well consider it as a left action.) The orbits are exactly the equivalence classes of the relation defined in Example 3.26 above, $x \sim y$ if and only if $x - y \in \mathbb{Z}$. Thus the quotient space of that example is the same as the coset space $\mathbb{R}/\mathbb{Z}$.

Consider also the map $\varepsilon\colon \mathbb{R} \to \mathbb{S}^1$ defined by

$$\varepsilon(s) = e^{2\pi i s}.$$

It is straightforward to check that this is a local homeomorphism and thus an open map, so it is a quotient map. Because it makes the same identifications as the quotient map $\mathbb{R} \to \mathbb{R}/\mathbb{Z}$, the uniqueness of quotient spaces tells us that $\mathbb{R}/\mathbb{Z}$ is homeomorphic to $\mathbb{S}^1$. (We will be returning to this map ε, which we call the *exponential quotient map*, extensively in this book.)

More generally, the discrete subgroup $\mathbb{Z}^n$ acts freely on $\mathbb{R}^n$ by translation. By similar reasoning, the quotient space $\mathbb{R}^n/\mathbb{Z}^n$ is homeomorphic to the n-torus $\mathbb{T}^n = \mathbb{S}^1 \times \cdots \times \mathbb{S}^1$.

We will see more examples of this technique in the next few chapters.

Problems

3-1. Show that a finite product of open maps is open, and a finite product of closed maps is closed.

3-2. By considering the space $X = [0,1] \subset \mathbb{R}$, and the sets $A_0 = \{0\}$, $A_i = [1/(i+1), 1/i]$ for $i = 1, 2, \ldots$, show that the gluing lemma (Lemma 3.8) is false if $\{A_1, \ldots, A_k\}$ is replaced by an infinite sequence of closed sets.

3-3. Formulate a "characteristic property" for the disjoint union topology (Problem 2-9) and prove that the disjoint union topology is uniquely characterized by it.

3-4. Use stereographic projection to show that any closed ball in $\mathbb{R}^n$ is an n-dimensional manifold with boundary.

3-5. Let X be a topological space. The *diagonal* of $X \times X$ is the subset $\Delta = \{(x,x) : x \in X\} \subset X \times X$. Show that X is Hausdorff if and only if Δ is closed in $X \times X$.

3-6. If $X_1, \ldots, X_k$ are topological spaces, show that the projections $\pi_i\colon X_1 \times \cdots \times X_k \to X_i$ are quotient maps.

3-7. Let $M = \mathbb{R}_d \times \mathbb{R}$, where $\mathbb{R}_d$ is the set $\mathbb{R}$ with the discrete topology.

  (a) Show that M is homeomorphic to the space X of Problem 2-5.

  (b) Show that M is locally Euclidean (of what dimension?) and Hausdorff, but not second countable.

3-8. Let X be the subset $\mathbb{R} \times \{0\} \cup \mathbb{R} \times \{1\}$ of $\mathbb{R}^2$. Define an equivalence relation on X by declaring $(x,0) \sim (x,1)$ if $x \neq 0$. Show that the quotient space $X/{\sim}$ is locally Euclidean and second countable, but not Hausdorff. (This space is called the *line with two origins.*)

3-9. Lemma 3.17 showed that the restriction of a quotient map to a saturated open set is still a quotient map. Show that the "saturated" hypothesis is necessary, by giving an example of a quotient map $f\colon X \to Y$ and an open subset $U \subset X$ such that $f|_U$ is surjective but not a quotient map.

3-10. Show that real projective space $\mathbb{P}^n$ is an n-manifold. [Hint: Consider the subsets $U_i \subset \mathbb{R}^{n+1}$ where $x_i = 1$.]

3-11. Let $\mathbb{CP}^n$ denote the set of all 1-dimensional complex subspaces of $\mathbb{C}^{n+1}$, called *n-dimensional complex projective space.* Topologize $\mathbb{CP}^n$ as the quotient $\mathbb{C}^{n+1} \smallsetminus \{0\}/\mathbb{C}^*$, where $\mathbb{C}^*$ is the group of nonzero complex numbers acting by scalar multiplication. Show that $\mathbb{CP}^n$ is a $2n$-manifold. [Hint: Mimic what you did in Problem 3-10.]

3-12. Let G be a topological group and let $H \subset G$ be a subgroup. Show that $\overline{H}$ is also a subgroup.

3-13. If G is a group that is also a topological space, show that G is a topological group if and only if the map $G \times G \to G$ given by $(x, y) \mapsto xy^{-1}$ is continuous.

3-14. Let G be a topological group and $\Gamma \subset G$ be a subgroup.

(a) For any $g \in G$, show that left translation $L_g \colon G \to G$ passes to the quotient G/Γ and defines a homeomorphism of G/Γ with itself.

(b) A topological space X is said to be *homogeneous* if for any $x, y \in X$, there is a homeomorphism $\varphi \colon X \to X$ taking x to y. Show that every coset space is homogeneous.

3-15. Let G be a topological group acting continuously on a topological space X.

(a) Show that the quotient map $\pi \colon X \to X/G$ is open.

(b) Show that X/G is Hausdorff if and only if the orbit relation

$$\{(x_1, x_2) \in X \times X : x_2 = g \cdot x_1 \text{ for some } g \in G\}$$

is closed in $X \times X$.

3-16. If Γ is a normal subgroup of the topological group G, show that the coset space G/Γ is a topological group. [Hint: It might be helpful to use Problems 3-1 and 3-15(a).]

4

Connectedness and Compactness

In this chapter we treat two topological properties that will be of central importance in our study of manifolds: connectedness and compactness.

The definition of connectedness is formulated so that connected spaces will behave similarly to intervals in the real line, so, for example, a continuous real-valued function on a connected space satisfies the intermediate value theorem. Similarly, compactness is defined so that compact spaces will have many of the same properties enjoyed by closed and bounded subsets of Euclidean spaces. In particular, continuous real-valued functions on compact sets always achieve their maxima and minima.

Connectedness

One of the most important elementary facts about continuous functions is the *intermediate value theorem*: If f is a continuous real-valued function defined on a closed bounded interval $[a, b]$, then f takes on every value between $f(a)$ and $f(b)$. The key idea here is the "connectedness" of intervals. In this section we generalize this concept to topological spaces.

Definitions and Basic Properties

If X is a topological space, a *separation* of X is a pair of nonempty, disjoint, open subsets $U, V \subset X$ such that $X = U \cup V$. We say that X is *disconnected* if there exists a separation of X, and *connected* otherwise.

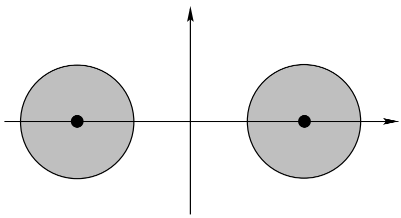


FIGURE 4.1. Union of two disks.

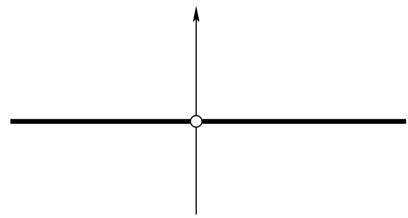


FIGURE 4.2. The x-axis minus 0.

By this definition, connectedness is a property of a *space*, not a property of subsets like openness or closedness. We can also talk about connected *subsets* of a topological space, by which we always mean connected in the subspace topology. In this context we can also consider a separation of A to be a pair of open subsets $U, V \subset X$ whose intersections with A are nonempty and disjoint, and whose union contains A: This is equivalent to the original definition because the open subsets of A are exactly the open subsets of X intersected with A.

Example 4.1. Each of the following subspaces of the plane is disconnected.

(a) X is the union of the two disjoint closed disks $\overline{B}_1(2,0)$ and $\overline{B}_1(-2,0)$ (Figure 4.1). Each of the disks is open in X, so the pair of disks is a separation of X.

(b) Y is the x-axis minus the origin (Figure 4.2). The two sets $\{(x,0) : x > 0\}$ and $\{(x,0) : x < 0\}$ separate Y.

(c) Z is the set of points with rational coordinates. A separation is given by, say, $\{(x,y) : x < \pi\}$ and $\{(x,y) : x > \pi\}$.

On the other hand, it is intuitively clear that the open and closed unit disks, the circle, the whole plane, and the x-axis are all connected, at least in the everyday sense of the word. Proving it, however, is not so easy, because we would have to show that it is impossible to find a separation. We will soon come up with an easy technique for proving connectedness that will work in most practical cases, including that of manifolds.

Here is a useful alternative characterization of connectedness.

Proposition 4.2. *A space X is connected if and only if the only subsets of X that are both open and closed in X are $\varnothing$ and X itself.*

Proof. Suppose first that X is connected, and assume that $U \subset X$ is open and closed. Then $V = X \smallsetminus U$ is also open and closed. If both U and V were nonempty, then $\{U, V\}$ would be a separation of X; therefore, either V is empty, which means that $U = X$, or U is empty.

Conversely, if X is disconnected, we can write $X = U \cup V$ where U and V are nonempty, open, and disjoint. This implies that U is open, closed, not empty, and not equal to X. □

The most important feature of connectedness is that continuous images of connected sets are connected.

Theorem 4.3 (Main Theorem on Connectedness). *Let X, Y be topological spaces and let $f: X \to Y$ be a continuous map. If X is connected, then $f(X)$ is connected.*

Proof. Suppose $f(X)$ is not connected. Then there exist open sets $U, V \subset Y$ whose intersections with $f(X)$ are nonempty and disjoint and whose union contains $f(X)$. It follows immediately that $\{f^{-1}(U), f^{-1}(V)\}$ is a separation of X, so X is not connected. □

Proposition 4.4 (Properties of Connected Sets).

(*a*) *Suppose X is any space and U, V are disjoint open subsets of X. If A is a connected subset of X contained in $U \cup V$, then either $A \subset U$ or $A \subset V$.*

(*b*) *Suppose X is any space and $A \subset X$ is connected. Then $\overline{A}$ is connected.*

(*c*) *Let X be a space, and let $\{B_\alpha\}_{\alpha \in A}$ be any collection of connected subspaces with a point in common. Then $\bigcup_{\alpha \in A} B_\alpha$ is connected.*

(*d*) *Any product of finitely many connected spaces is connected.*

(*e*) *Any quotient space of a connected space is connected.*

Proof. For part (a), if A contained points in both U and V, then $\{A \cap U, A \cap V\}$ would be a separation of A.

To prove part (b), suppose U and V are disjoint open subsets of X that separate $\overline{A}$. By (a), A is contained in one of the sets, say $A \subset U$. Each point of V has a neighborhood (namely V) disjoint from A, so every point of V is exterior to A. Therefore, $\overline{A} \subset U$, which means that $\overline{A} \cap V = \varnothing$, a contradiction.

For part (c), let q be a point contained in each B_α, and suppose $\{U, V\}$ is a separation of $\bigcup_{\alpha \in A} B_\alpha$. Suppose without loss of generality that q lies in U. By part (a), each B_α must be entirely contained in U, and thus so is their union.

For part (d), since $X_1 \times \cdots \times X_k = (X_1 \times \cdots \times X_{k-1}) \times X_k$, by induction it suffices to consider a product of two spaces. Thus let X and Y be connected, and suppose $\{U, V\}$ is a separation of $X \times Y$. Choose any point $(x_0, y_0) \in U$. The set $\{x_0\} \times Y$ is connected because it is homeomorphic to Y; since it contains the point $(x_0, y_0) \in U$, it must be entirely contained in U by part (a). For each $y \in Y$, the set $X \times \{y\}$ is also connected and has a point

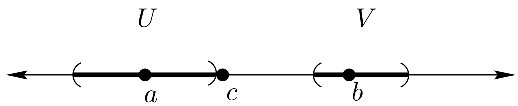


FIGURE 4.3. Proof that an interval is connected.

$(x_0, y) \in U$, so it must also be contained in U. Since the sets $X \times \{y\}$ exhaust $X \times Y$, the result follows.

Finally, (e) follows from Theorem 4.3 and the fact that quotient maps are surjective. $\square$

Although this proposition gives us a number of ways of building new connected spaces out of given ones, so far we have no examples of spaces to start with that are known to be connected (except a one-point space, which does not carry us very far). The one example of a space that can be shown to be connected by "brute force" is the one that enters into the proof of the intermediate value theorem: an interval in the real line (see the Appendix).

Proposition 4.5. *A nonempty subset of* $\mathbb{R}$ *is connected if and only if it is an interval.*

Proof. First assume that $J \subset \mathbb{R}$ is an interval. If it is not connected, there are open subsets $U, V \subset \mathbb{R}$ that separate J. Choose $a \in U \cap J$, $b \in V \cap J$, and assume (interchanging U and V if necessary) that $a < b$ (Figure 4.3). Then $[a, b] \subset J$ because J is an interval. Since U and V are both open in $\mathbb{R}$, there exists $\varepsilon > 0$ such that $[a, a + \varepsilon) \subset U \cap J$ and $(b - \varepsilon, b] \subset V \cap J$.

Let $c = \sup(U \cap [a, b])$. By our choice of ε, $a + \varepsilon \leq c \leq b - \varepsilon$. In particular, c is between a and b, so $c \in J \subset U \cup V$. But if c were in U, it would have a neighborhood $(c - \delta, c + \delta) \subset U$, which would contradict the definition of c. Similarly, $c \in V$ leads to a contradiction. Therefore, J is connected.

Conversely, assume that J is not an interval. This means that there exist $a < c < b$ with $a, b \in J$ but $c \notin J$. Then the sets $(-\infty, c)$ and (c, ∞) separate J, so J is not connected. $\square$

An immediate consequence of this proposition is the following generalized intermediate value theorem.

Theorem 4.6 (Intermediate Value Theorem). *Suppose* X *is a connected topological space, and* f *is a continuous real-valued function on* X*. If* $p, q \in X$*, then* f *takes on all values between* $f(p)$ *and* $f(q)$*.*

Proof. The image set $f(X)$ is connected, so it must be an interval. $\square$

Path Connectedness

Now we can give a simple but powerful sufficient condition for connectedness, based on the following definitions. Let X be a topological space and $p, q \in X$. A *path* in X from p to q is a continuous map $f\colon [0,1] \to X$ such that $f(0) = p$ and $f(1) = q$. We say that X is *path connected* if for every $p, q \in X$, there is a path in X from p to q.

Theorem 4.7. *Path connectedness implies connectedness.*

Proof. Suppose that X is path connected but not connected, and let $\{U, V\}$ be a separation of X. We can choose $p \in U$ and $q \in V$ (since neither set is empty), and find a path f from p to q in X. Then $f^{-1}(U)$ and $f^{-1}(V)$ are disjoint open subsets of $[0,1]$ that cover $[0,1]$; moreover, $0 \in f^{-1}(U)$ and $1 \in f^{-1}(V)$, so neither set is empty. This implies that $[0,1]$ is disconnected, which is a contradiction. □

Example 4.8. The following spaces are all easily shown to be path connected, and therefore they are connected.

(a) $\mathbb{R}^n$.

(b) Any subset $B \subset \mathbb{R}^n$ that is *convex*, which means that for any $x, x' \in B$, the line segment from x to x' lies entirely in B.

(c) $\mathbb{R}^n \smallsetminus \{0\}$ for $n \geq 2$.

Example 4.9. The following spaces are also connected.

(a) $\mathbb{S}^n$ for $n \geq 1$, because it is a quotient space of $\mathbb{R}^{n+1} \smallsetminus \{0\}$ by Example 3.27.

(b) The n-torus $\mathbb{T}^n$, because it is a product of connected spaces.

On the other hand, path connectedness is stronger in general than connectedness. Here is an example of a space that is connected but not path connected.

Example 4.10. Define subsets of the plane by

$$\begin{aligned} A &= \{(x,y) : x = 0 \text{ and } y \in [-1,1]\}; \\ B &= \{(x,y) : y = \sin(1/x) \text{ and } x \in (0,1]\}. \end{aligned}$$

Let $X = A \cup B$ (Figure 4.4). X is called the *topologist's sine curve*. In Problem 4-5 you will show that it is connected but not path connected.

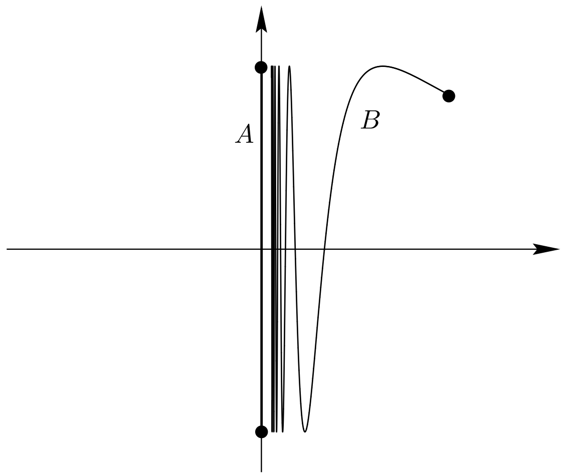


FIGURE 4.4. The topologists's sine curve.

Components and Path Components

Look back at Example 4.1. Our first example of a disconnected set, the union of two disjoint closed disks, could be separated in only one way, because any other separation would induce a separation of one of the closed disks, which is path connected. The same reasoning applies to the second example, the x-axis minus the origin. The set of rational points in the plane, however, admits infinitely many possible separations. Identifying the possible separations of a space amounts to finding "maximal" connected subsets, a concept we now explore more fully.

Let X be a topological space. Define a relation on X, called the *connectivity relation*, by saying that $p \sim q$ if there exists a connected subset of X containing both p and q.

Lemma 4.11. *The connectivity relation is an equivalence relation.*

Proof. It is reflexive because $\{q\}$ is a connected subset containing q, and symmetric because $p \sim q$ and $q \sim p$ both mean that there is a connected subset containing p and q. To prove transitivity, suppose $p \sim q$ and $q \sim r$, which means that there are connected subsets A containing $\{p, q\}$ and B containing $\{q, r\}$. Since A and B have the point q in common, $A \cup B$ is connected by Proposition 4.4(c). Thus $A \cup B$ is a connected set containing $\{p, r\}$, so $p \sim r$. $\square$

The equivalence classes in X under the connectivity relation are called the *components* of X.

Lemma 4.12. *The components of X are exactly the maximal connected subsets of X, that is, connected sets that are not contained in any larger connected set.*

Proof. Given $q \in X$, let A be the component of X containing q, and let B be the union of all connected sets containing q. Then B itself is connected by Proposition 4.4(c), and is thus a maximal connected subset. If $p \in B$, then p, q lie in the connected subset B, so $p \sim q$ and thus $p \in A$. Conversely, if $p \in A$, then $p \sim q$, so p lies in some connected subset containing q. Since B is the union of all such subsets, $p \in B$. □

Example 4.13. Consider the disconnected subsets of Example 4.1.

(a) The components of X (the union of two disjoint closed disks) are the two disks themselves.

(b) The components of Y (the x-axis minus the origin) are the positive x-axis and the negative x-axis.

(c) In the set Z of points with rational coordinates, if p and q are distinct points of Z, they must differ in one of their coordinates, say their x-coordinates. Choosing an irrational number α between the two x-coordinates, the sets where $x < \alpha$ and $x > \alpha$ give a separation of Z in which p and q lie in different subsets. Therefore, p and q cannot both be contained in any connected subset, so p is not equivalent to q. Thus the components of Z are the one-point subsets.

Proposition 4.14 (Properties of Components). *Let X be any space.*

(*a*) *Each component of X is closed in X.*

(*b*) *Any connected subset of X is contained in a single component.*

Proof. If B is any component of X, it follows from Proposition 4.4(b) that $\overline{B}$ is a connected set containing B. Since components are maximal connected sets, $\overline{B} = B$, so B is closed.

Suppose $A \subset X$ is connected. Since the components cover X, A has a point in common with some component B. By Proposition 4.4(c) $A \cup B$ is connected, so by maximality of B, it must be equal to B. This means that $A \subset B$. □

Although components are always closed, they may not be open in general, so they do not necessarily separate the space. Consider the set Z of rational points in the plane, for example: Its components are single points, which are not open sets.

We can also apply the construction used to define components with path connectedness in place of connectedness. Define the *path connectivity relation* for points p, q in a space X by saying $p \underset{\mathrm{p}}{\sim} q$ if there is a path in X from p to q.

Exercise 4.1. Show that $\underset{\mathrm{p}}{\sim}$ is an equivalence relation.

The equivalence classes under $\underset{\mathrm{p}}{\sim}$ are called the *path components* of X.

Proposition 4.15 (Properties of Path Components). *Let X be any space.*

(*a*) *Each path component is contained in a single component, and each component is a disjoint union of path components.*

(*b*) *If $A \subset X$ is path connected, then A is contained in a single path component.*

Exercise 4.2. Prove Proposition 4.15.

We say that a space X is *locally connected* if it admits a basis of connected open sets, and *locally path connected* if it admits a basis of path connected open sets. To put it more concretely, for any $p \in X$ and any neighborhood U of p, p has a (path) connected neighborhood contained in U. Clearly, any locally path connected space is locally connected.

A space can be connected but not locally connected, as is, for example, the topologist's sine curve (see Problem 4-5); and it can be locally connected but not connected, as is the disjoint union of two closed disks.

Lemma 4.16. *Let X be any space.*

(*a*) *If X is locally connected, then each component of X is open.*

(*b*) *If X is locally path connected, then each path component of X is open, its path components are the same as its components, and it is connected if and only if it is path connected.*

Proof. First assume that X is locally connected, and let A be a component of X. If $p \in A$, then p has a connected neighborhood U by local connectedness, and this neighborhood must lie entirely in A by Lemma 4.14(b). Thus every point of A has a neighborhood in A; in other words, A is open.

Now assume that X is locally path connected. The same argument, with "connected" replaced by "path connected," shows that each path component is open. Let $q \in X$, and let A be the component containing q, and B the path component. By Proposition 4.15(a), we know that $B \subset A$ and A can be written as a disjoint union of path components, each of which is open in X and thus in A. If B is not the only path component in A, then the pair $\{B, A \setminus B\}$ is a separation of A, which is a contradiction because A is connected. This proves that $A = B$. Finally, X being connected means it has only one component, which by the above argument is the same as having only one path component, which in turn is equivalent to being path connected. $\square$

Proposition 4.17. *Every manifold is locally path connected.*

Exercise 4.3. Prove Proposition 4.17.

This proposition means that in our work with manifolds we can use connectedness and path connectedness interchangeably. This will simplify many arguments because path connectedness is so much easier to check.

Compactness

Another fundamental fact about continuous functions is the *extreme value theorem* (Theorem A.10 in the Appendix): A continuous real-valued function on a closed, bounded subset of $\mathbb{R}^n$ attains its maximum and minimum values.

This theorem, of course, fails in general for metric spaces, and "bounded" does not even make sense in topological spaces. But the essential property of closed and bounded subsets of $\mathbb{R}^n$ that makes the proof work, compactness, makes sense in arbitrary topological spaces. This property is the subject of the rest of this chapter.

Definitions and Basic Properties

Recall that an *open cover* of a space X is a collection $\mathcal{U}$ of open subsets of X whose union is X, and a *subcover* of $\mathcal{U}$ is a subcollection of $\mathcal{U}$ that still covers X. A topological space X is said to be *compact* if every open cover of X has a finite subcover; or in other words, if given any open cover $\mathcal{U}$ of X, there are finitely many sets $U_1, \dots, U_k \in \mathcal{U}$ such that $X = U_1 \cup \dots \cup U_k$.

As in the case of connectedness, when we say that a subset A of a topological space X is compact, we always mean with respect to the subspace topology unless otherwise specified. A subspace $A \subset X$ is compact if and only if given any collection of open subsets of X whose union contains A (which we also call an open cover of A), there is a finite subcover.

The most important fact about compact spaces is that continuous images of compact spaces are compact.

Theorem 4.18 (Main Theorem on Compactness). *Let X, Y be topological spaces and let $f\colon X \to Y$ be a continuous map. If X is compact, then $f(X)$ is compact.*

Proof. Let $\mathcal{U}$ be an open cover of $f(X)$. (As noted in the remark above, we can take the elements of $\mathcal{U}$ either to be open subsets of $f(X)$ in the subspace topology, or to be open subsets of Y whose union contains $f(X)$.) For each $U \in \mathcal{U}$, $f^{-1}(U)$ is an open subset of X. Since $\mathcal{U}$ covers $f(X)$, every point of X is in some set $f^{-1}(U)$, so the collection $\{f^{-1}(U) : U \in \mathcal{U}\}$ is an open cover of X. By compactness of X, some finite number of these, say

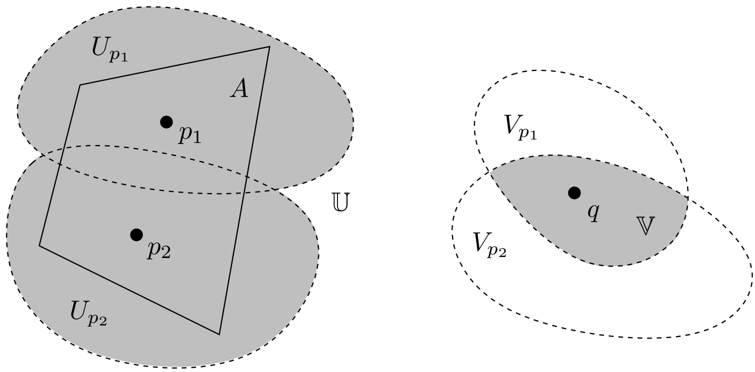


FIGURE 4.5. The case $B = \{q\}$.

$\{f^{-1}(U_1), \dots, f^{-1}(U_k)\}$, cover X. Then it follows that $\{U_1, \dots, U_k\}$ cover $f(X)$. □

Proposition 4.19 (Properties of Compact Spaces).

(*a*) *Every closed subset of a compact space is compact.*

(*b*) *In a Hausdorff space X, compact sets can be separated by open sets. That is, if $A, B \subset X$ are disjoint compact subsets, there exist disjoint open sets $U, V \subset X$ such that $A \subset U$ and $B \subset V$.*

(*c*) *Every compact subset of a Hausdorff space is closed.*

(*d*) *Every finite product of compact spaces is compact.*

(*e*) *Every quotient of a compact space is compact.*

Proof. For part (a), let $\mathcal{U}$ be a cover of A by open subsets of X. Then $\mathcal{U} \cup \{X \smallsetminus A\}$ is an open cover of X, which has a finite subcover $\{U_1, \dots, U_k, X \smallsetminus A\}$. Therefore, A must be covered by the finite collection $\{U_1, \dots, U_k\}$.

To prove (b), first consider the case in which $B = \{q\}$ is a one-point set. For each $p \in A$, there exist disjoint open sets U_p containing p and V_p containing q by the Hausdorff property. The collection $\{U_p : p \in A\}$ is an open cover of A, so it has a finite subcover: Call it $\{U_{p_1}, \dots, U_{p_k}\}$ (Figure 4.5). Let $\mathbb{U} = U_{p_1} \cup \dots \cup U_{p_k}$ and $\mathbb{V} = V_{p_1} \cap \dots \cap V_{p_k}$. Then $\mathbb{U}$ and $\mathbb{V}$ are disjoint open sets with $A \subset \mathbb{U}$ and $\{q\} \subset \mathbb{V}$, so this case is proved.

Next consider the case of a general compact subset B. The argument above shows that for each $q \in B$ there exist disjoint open subsets $\mathbb{U}_q, \mathbb{V}_q \subset X$ such that $A \subset \mathbb{U}_q$ and $q \in \mathbb{V}_q$. By compactness of B, finitely many of these, say $\{\mathbb{V}_{q_1}, \dots, \mathbb{V}_{q_m}\}$, cover B. Then setting $U = \mathbb{U}_{q_1} \cap \dots \cap \mathbb{U}_{q_m}$ and $V = \mathbb{V}_{q_1} \cup \dots \cup \mathbb{V}_{q_m}$ proves the result.

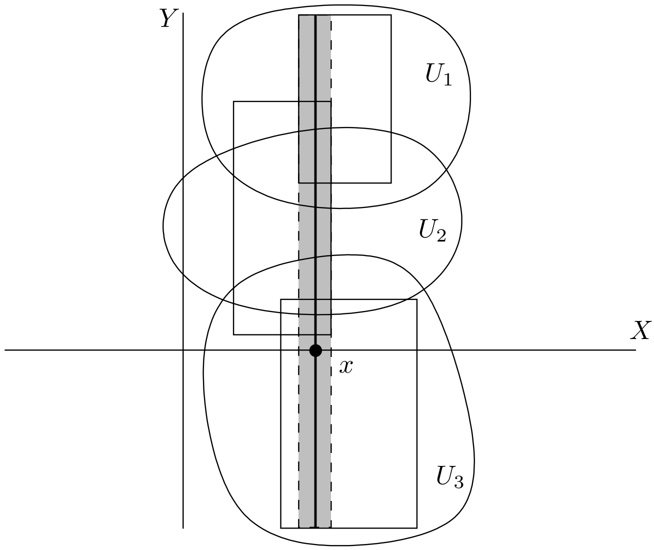


FIGURE 4.6. Finding a finite cover of the strip $Z_x \times Y$.

For (c), suppose X is Hausdorff and $A \subset X$ is compact. For any point $q \in X \smallsetminus A$, by part (b) there exist disjoint open sets U containing A and V containing q. In particular, V is a neighborhood of q disjoint from A, so every such q is exterior to A. This means that A is closed.

To prove (d), it suffices by induction to consider a product $X \times Y$ of two compact spaces. Let $\mathcal{U}$ be an open cover of $X \times Y$. Choose any $x \in X$. The "slice" $\{x\} \times Y$ is homeomorphic to Y, so finitely many of the sets of $\mathcal{U}$ cover it, say $U_1, \dots, U_k$ (Figure 4.6). Because product open sets are a basis for the product topology, for each $y \in Y$ there is a product open set $V \times W \subset X \times Y$ such that $(x, y) \in V \times W \subset U_1 \cup \cdots \cup U_k$. Finitely many of these product sets cover $\{x\} \times Y$, say $V_1 \times W_1, \dots, V_m \times W_m$. If we set $Z_x = V_1 \cap \cdots \cap V_m$, then it is evident that the whole "strip" $Z_x \times Y$ is actually contained in $U_1 \cup \cdots \cup U_k$.

Thus we have shown the following: For each $x \in X$, there exists an open subset $Z_x \subset X$ such that $Z_x \times Y$ is covered by finitely many of the sets in $\mathcal{U}$. The collection $\{Z_x : x \in X\}$ is an open cover of X, which by compactness has a finite subcover, say $\{Z_{x_1}, \dots, Z_{x_k}\}$. Since finitely many sets of $\mathcal{U}$ cover each strip $Z_{x_i} \times Y$, and finitely many such strips cover $X \times Y$, we are done.

Finally, part (e) is immediate from Theorem 4.18, since a quotient of a compact space is the image of a compact space by a continuous map. □

Part (d) is actually true in the more general context of infinite products (see [Sie92] or [Mun75]); in its general form, it is known as *Tychonoff's theorem.*

Exercise 4.4. Let X be a compact space, and suppose $\{F_n\}$ is a countable collection of nonempty closed subsets of X that are *nested*, which means that $F_n \supset F_{n+1}$ for each n. Show that $\bigcap_n F_n$ is nonempty.

One of the main applications of compactness is the following generalization of the extreme value theorem of elementary calculus.

Theorem 4.20 (Extreme Value Theorem). *If X is a compact space and $f\colon X \to \mathbb{R}$ is continuous, then f is bounded and attains its maximum and minimum values on X.*

Proof. By the main theorem on compactness, $f(X)$ is a compact subset of $\mathbb{R}$, so by Proposition A.6 it is closed and bounded. In particular, it contains its supremum and infimum. □

The next lemma expresses an important property of compact metric spaces, which we will use frequently later in the book. Recall that the *diameter* of a set S in a metric space is defined to be $\operatorname{diam}(S) = \sup\{d(x, y) : x, y \in S\}$. If $\mathcal{U}$ is an open cover of a metric space, a number $\delta > 0$ is called a *Lebesgue number* for the cover if any set whose diameter is less than δ is contained in one of the sets $U \in \mathcal{U}$.

Lemma 4.21 (Lebesgue Number Lemma). *Any open cover of a compact metric space has a Lebesgue number.*

Proof. Let $\mathcal{U}$ be an open cover of the compact metric space M. Each point $x \in M$ is in some set $U \in \mathcal{U}$. Since U is open, there is some $r(x) > 0$ such that $B_{2r(x)}(x) \subset U$. The balls $\{B_{r(x)}(x) : x \in M\}$ form an open cover of M, so finitely many of them, say $B_{r(x_1)}(x_1), \dots, B_{r(x_n)}(x_n)$, cover M.

We will show that $\delta = \min\{r(x_1), \dots, r(x_n)\}$ is a Lebesgue number for $\mathcal{U}$. To see why, suppose $S \subset M$ is a nonempty set whose diameter is less than δ. Let y be any point of S; then there is some x_i such that $y \in B_{r(x_i)}(x_i)$ (Figure 4.7). It suffices to show that any other point of S is in $B_{2r(x_i)}(x_i)$, since the latter set is by construction contained in some $U \in \mathcal{U}$. If $z \in S$, the triangle inequality gives

$$d(z, x_i) \le d(z, y) + d(y, x_i) < \delta + r(x_i) \le 2r(x_i),$$

which proves the claim. □

Sequential and Limit Point Compactness

The definition of compactness in terms of open covers lends itself to simple proofs of some rather powerful theorems, but it does not convey much intuitive content. There are two other properties that are equivalent to compactness for manifolds and metric spaces (though not for arbitrary topological spaces), and that give a more vivid picture of what compactness really means. A space X is said to be *limit point compact* if every

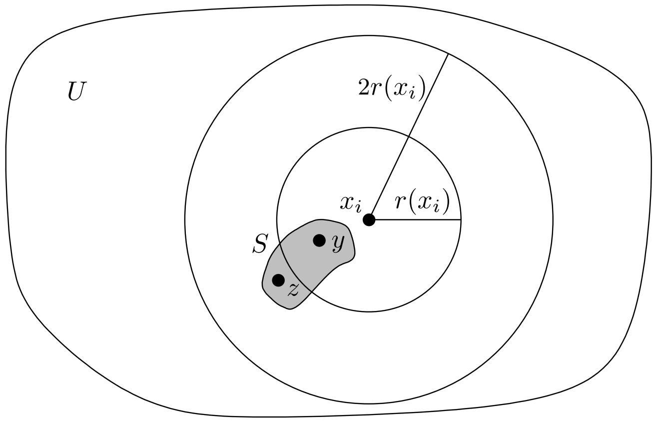


FIGURE 4.7. Proof of the Lebesgue number lemma.

infinite subset of X has a limit point in X, and *sequentially compact* if every sequence of points in X has a subsequence that converges to a point in X.

Proposition 4.22. *Compactness implies limit point compactness.*

Proof. Suppose X is compact, and let $S \subset X$ be an infinite subset. If S has no limit point, then every point $x \in X$ has a neighborhood U such that $U \cap S$ is either empty or $\{x\}$. Finitely many of these neighborhoods cover X. But since each such neighborhood contains at most one point of S, this implies that S is finite, which is a contradiction. □

Problem 4-7 shows that the converse of this proposition is not true in general.

Lemma 4.23. *For first countable Hausdorff spaces, limit point compactness implies sequential compactness.*

Proof. Suppose X is first countable, Hausdorff, and limit point compact, and let $\{p_n\}$ be any sequence of points in X. If the sequence takes on only finitely many values, then it has a constant subsequence, which is certainly convergent. So we may suppose it takes on infinitely many values.

By hypothesis the set of values $\{p_n\}$ has a limit point $q \in X$. If q is actually equal to p_n for infinitely many values of n, again there is a constant subsequence and we are done; so by discarding finitely many terms at the beginning of the sequence if necessary we may assume $p_n \neq q$ for all n. First countability of X means that there is a countable neighborhood basis at q, say $\{B_n : n = 1, 2, \dots\}$. By replacing B_n with $B_1 \cap \cdots \cap B_n$ if necessary, we may assume that the neighborhood basis is *nested*: $B_1 \supset B_2 \supset \cdots$. For such a neighborhood basis, it is easy to see that any subsequence $\{p_{n_i}\}$ such that $p_{n_i} \in B_i$ converges to q.

Since q is a limit point, we can choose n_1 such that $p_{n_1} \in B_1$. Suppose by induction that we have chosen $n_1 < n_2 < \cdots < n_k$ with $p_{n_i} \in B_i$. By the Hausdorff property, q has a neighborhood U disjoint from the finite set $\{p_n : 1 \le n \le n_k\}$, and by definition of limit point there is some n_{k+1} (necessarily greater than n_k) such that $p_{n_{k+1}} \in U \cap B_{k+1}$. This completes the induction, and proves that there is a subsequence $\{p_{n_i}\}$ converging to q. □

The next result shows that for manifolds and most of the other spaces we will be considering in this book, we can use all three notions of compactness interchangeably.

Proposition 4.24. *For metric spaces and second countable Hausdorff spaces, compactness, limit point compactness, and sequential compactness are all equivalent.*

Proof. We have shown that compactness implies limit point compactness for all spaces, and limit point compactness implies sequential compactness for first countable Hausdorff spaces, which include both metric spaces and second countable Hausdorff spaces. So it remains to show that a metric space or second countable Hausdorff space that is sequentially compact is actually compact.

Suppose first that X is second countable and sequentially compact. (For this part we do not need the Hausdorff property.) Any open cover $\mathcal{U}$ of X has a countable subcover $\{U_n : n = 1, 2, \dots\}$ by Lemma 2.15. Suppose that no finite subcollection of U_n's covers X. This means that for each n there exists $q_n \in X$ such that $q_n \notin U_1 \cup \cdots \cup U_n$. By hypothesis, the sequence $\{q_n\}$ has a convergent subsequence $q_{n_k} \to q$. Now, $q \in U_m$ for some m because the U_n's cover X, and then convergence of the subsequence means that there exists some N such that $q_{n_k} \in U_m$ whenever $k \ge N$. But by construction, $q_{n_k} \notin U_1 \cup \cdots \cup U_m$ as soon as $n_k \ge m$, which is a contradiction. This proves that finitely many of the U_n's cover X. Therefore, second countable sequentially compact spaces are compact.

Finally, let M be a sequentially compact metric space. We will show that M is second countable, which by the above argument implies that M is compact. From Problem 2-15, it suffices to show that M has a countable dense subset.

The key idea is to show first that sequential compactness implies the following weak form of compactness for metric spaces: For each $\varepsilon > 0$, the open cover of M consisting of all ε-balls has a finite subcover. Suppose this is not true for some ε. Construct a sequence as follows. Let $q_1 \in M$ be arbitrary. Since $B_\varepsilon(q_1) \neq M$, there is a point $q_2 \notin B_\varepsilon(q_1)$. Similarly, since $B_\varepsilon(q_1) \cup B_\varepsilon(q_2) \neq M$, there is a point q_3 in neither of the two preceding ε-balls. Proceeding by induction, we construct a sequence $\{q_n\}$ such that for each n,

$$q_{n+1} \notin B_\varepsilon(q_1) \cup \cdots \cup B_\varepsilon(q_n). \tag{4.1}$$

Replacing this sequence by a convergent subsequence (which still satisfies (4.1)), we can assume $q_n \to q \in M$. Since convergent sequences are Cauchy, as soon as n is large enough we have $d(q_{n+1}, q_n) < \varepsilon$, which contradicts (4.1).

Now, for each n let $q_1^{(n)}, \dots, q_{k_n}^{(n)}$ be finitely many points such that the balls of radius $1/n$ around these points cover M. The collection of points $\{q_i^{(n)}\}$ is countable, and is easily seen to be dense. This shows that M is second countable and completes the proof. □

Exercise 4.5. Show that every compact metric space is complete.

The Closed Map Lemma

The next lemma, though simple, is among the most useful results in this entire chapter.

Lemma 4.25 (Closed Map Lemma). *Suppose F is a continuous map from a compact space to a Hausdorff space.*

(*a*) *F is a closed map.*

(*b*) *If F is surjective, it is a quotient map.*

(*c*) *If F is injective, it is a topological embedding.*

(*d*) *If F is bijective, it is a homeomorphism.*

Proof. Let $F\colon X \to Y$ be such a map. If $A \subset X$ is closed, it is compact, since any closed subset of a compact space is compact (Proposition 4.19(a)). Therefore, $F(A)$ is compact by the main theorem on compactness, and closed in Y because compact subsets of Hausdorff spaces are closed (Proposition 4.19(c)). This shows that F is a closed map. If in addition F is surjective, it is a quotient map by Lemma 3.19. If it is bijective, the fact that it is closed implies that its inverse is continuous, so it is a homeomorphism (Exercise 2.14). Finally, if F is injective, it is bijective onto its image, so the fact that it is an embedding follows from (d). □

Here are some immediate applications of the closed map lemma.

In Example 3.24 we constructed a quotient space of the square $I \times I$ by gluing the side boundary segments together and the top and bottom boundary segments together, and we claimed that it was homeomorphic to the torus. Here is a proof. Construct another map $q\colon I \times I \to \mathbb{T}^2$ by setting $q(u, v) = (\cos 2\pi u, \sin 2\pi u, \cos 2\pi v, \sin 2\pi v)$. By the closed map lemma, this is a quotient map. Since it makes the same identifications as the quotient map we started with, the original quotient of $I \times I$ must be homeomorphic to the torus by the uniqueness of quotient spaces.

In Lemma 3.15 we showed that the doughnut surface is homeomorphic to the torus by a rather laborious explicit computation. Now that lemma can be proved much more simply, as follows. Consider the map $F\colon \mathbb{R}^2 \to \mathbb{R}^3$ defined in Example 3.7. The restriction of this map to $I \times I$ is a quotient map by the closed map lemma. Since it makes the same identifications as the map q in the preceding paragraph, the two quotient spaces D and $\mathbb{S}^1 \times \mathbb{S}^1$ are homeomorphic. (The homeomorphism is the map that sends $q(u,v)$ to $F(u,v)$.)

Another application of the closed map lemma is the following useful result. You should notice how the closed map lemma, invoked twice in the proof, allows us to avoid ever having to prove continuity directly by ε-δ estimates.

Proposition 4.26. *Let K be a compact convex subset of $\mathbb{R}^n$ with nonempty interior. Then K is homeomorphic to the closed unit ball $\overline{\mathbb{B}^n}$, by a homeomorphism that sends $\mathbb{S}^{n-1}$ to ∂K.*

Proof. Let q be an interior point of K. By replacing K with its image under the translation $x \mapsto x - q$ (which is a homeomorphism of $\mathbb{R}^n$ with itself), we can assume $0 \in \operatorname{Int} K$. Then there is some $\varepsilon > 0$ such that the ball $B_\varepsilon(0)$ is contained in K; using the dilation $x \mapsto x/\varepsilon$, we can assume $\mathbb{B}^n = B_1(0) \subset K$.

The core of the proof is the following claim: *Each ray starting at the origin intersects ∂K in exactly one point.* Since K is compact, its intersection with each closed ray is compact; thus there is a point x_0 in this intersection at which the distance to the origin assumes its maximum. This point is easily seen to lie in the boundary of K. To see that there can be only one such point, we will show that the line segment from 0 to x_0 consists entirely of interior points of K, except for x_0 itself. Since $B_1(0) \subset K$, every line segment from x_0 to a point $y \in B_1(0)$ is contained in K. As y ranges over $B_1(0)$, these line segments sweep out a set C shaped like an ice cream cone (Figure 4.8). Around each point λx_0 for $0 \le \lambda < 1$, it is easy to check that there is a ball of radius $1 - \lambda$ contained in C and hence in K. Thus x_0 is the only boundary point of K on the ray.

Now we define a map $f\colon \partial K \to \mathbb{S}^{n-1}$ by

$$f(x) = \frac{x}{|x|}.$$

In words, $f(x)$ is the point where the line segment from the origin to x intersects the sphere. Since f is the restriction of a continuous map, it is continuous, and the discussion in the preceding paragraph shows that it is bijective. Since ∂K is compact, f is a homeomorphism by the closed map lemma.

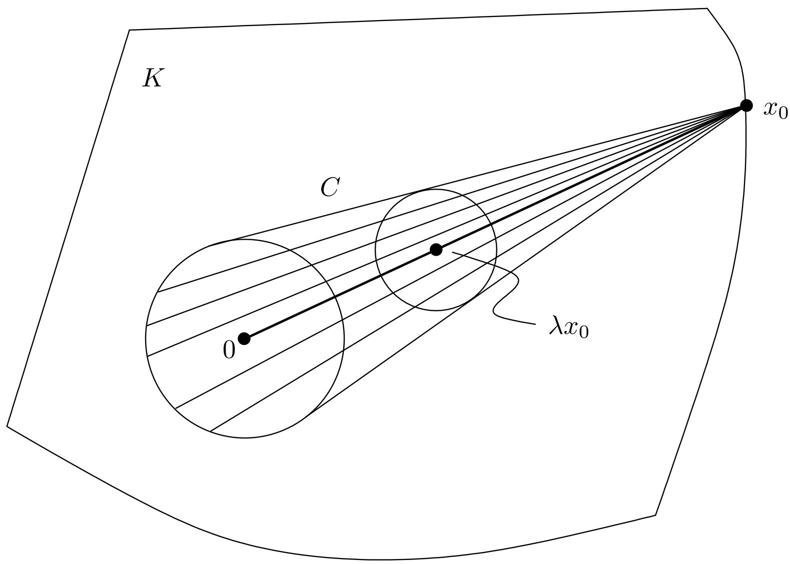


FIGURE 4.8. Proof that there is only one boundary point on a ray.

Finally, define $F\colon \overline{\mathbb{B}^n} \to K$ by

$$F(x) = |x| f^{-1}\left(\frac{x}{|x|}\right).$$

Then F is continuous because f^{-1} is. Geometrically, F takes each radial line segment from 0 to a point $\omega \in \mathbb{S}^n$ linearly onto the radial segment from 0 to the point $f^{-1}(\omega) \in \partial K$. By convexity, F takes its values in K. The map F is injective, since points on distinct rays are mapped to distinct rays, and each radial segment is mapped linearly to its image. It is surjective because each point $y \in K$ is on some ray from 0. By the closed map lemma, F is a homeomorphism. □

Locally Compact Hausdorff Spaces

Compact Hausdorff spaces have many of the familiar properties of subsets of Euclidean spaces. However, while all manifolds are Hausdorff, many interesting manifolds are not compact. Nonetheless, many of the nice properties of compact Hausdorff spaces carry over to a more general class of spaces, which we now define.

A topological space X is said to be *locally compact* if for every $q \in X$ there is a compact subset of X containing a neighborhood of q. In this generality, the definition is not particularly useful, and does not seem parallel to other definitions of what it means for a topological space to possess a property "locally," which usually entails the existence of a basis of open sets with a particular property. But when combined with the Hausdorff property, local compactness is much more useful. A subset A of a topological space X is said to be *precompact* or *relatively compact* if $\overline{A}$ is compact.

Proposition 4.27. *Let X be a Hausdorff space. The following are equivalent.*

(*a*) *X is locally compact.*

(*b*) *Each point of X has a precompact neighborhood.*

(*c*) *X has a basis of precompact open sets.*

Proof. Clearly, (c) $\Longrightarrow$ (b) $\Longrightarrow$ (a), so all we have to prove is (a) $\Longrightarrow$ (c). It suffices to show that if X is locally compact Hausdorff, then each point $x \in X$ has a neighborhood basis of precompact open sets. Let $K \subset X$ be a compact set containing a neighborhood U of x. The collection $\mathcal{V}$ of all neighborhoods of x contained in U is clearly a neighborhood basis at x.

Because X is Hausdorff, K is closed in X. If $V \in \mathcal{V}$, then $\overline{V} \subset K$ (because $V \subset U \subset K$ and K is closed), and therefore $\overline{V}$ is compact (because a closed subset of a compact set is compact). Thus $\mathcal{V}$ is the required neighborhood basis. □

Lemma 4.28 (Shrinking Lemma). *Let X be a locally compact Hausdorff space. If $x \in X$ and U is any neighborhood of x, there exists a precompact neighborhood V of x such that $\overline{V} \subset U$.*

Proof. Suppose $x \in X$ and U is a neighborhood of x. If W is any precompact neighborhood of x, then $\overline{W} \smallsetminus U$ is closed in $\overline{W}$ and therefore compact. Because open sets separate compact sets in a Hausdorff space, there are disjoint open sets Y containing x and Y' containing $\overline{W} \smallsetminus U$ (Figure 4.9). Let $V = Y \cap W$. Because $\overline{V} \subset \overline{W}$, $\overline{V}$ is compact. Because $\overline{V} \subset \overline{Y}$, which is disjoint from Y', we have $\overline{V} \subset \overline{W} \smallsetminus Y'$. Now the fact that $\overline{W} \smallsetminus U \subset Y'$ means that $\overline{W} \smallsetminus Y' \subset U$, so $\overline{V} \subset U$. □

Lemma 4.29. *Any open or closed subset of a locally compact Hausdorff space is locally compact Hausdorff.*

Proof. Let X be a locally compact Hausdorff space. Note that any subspace of X is Hausdorff, so only local compactness needs to be checked. If $Y \subset X$ is open, the shrinking lemma says that any point in Y has a neighborhood whose closure is compact and contained in Y, so Y is locally compact. Suppose $Z \subset X$ is closed. Any $x \in Z$ has a precompact neighborhood U in X. Since $\overline{U \cap Z} = \overline{U} \cap Z$ is a closed subset of the compact set $\overline{U}$, it is compact, so $U \cap Z$ is a precompact neighborhood of x in Z. Thus Z is locally compact. □

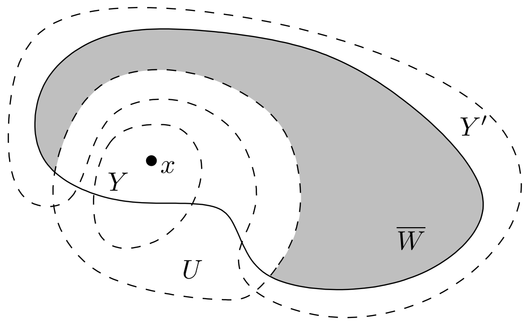


FIGURE 4.9. Proof of the shrinking lemma.

Exercise 4.6. Show that any finite product of locally compact Hausdorff spaces is locally compact Hausdorff.

Example 4.30 (Locally Compact Hausdorff Spaces).

(a) Euclidean space $\mathbb{R}^n$ is locally compact Hausdorff, because any closed ball $\overline{B}_\varepsilon(x)$ is a precompact neighborhood of x. Thus every open or closed subset of $\mathbb{R}^n$ is locally compact Hausdorff.

(b) Let M be a manifold, and let $\mathcal{U}$ be a cover of M by Euclidean balls. Each $U \in \mathcal{U}$ has a basis of open sets that are precompact in U and thus also in M, and the union of all such bases is a basis for M. Thus any manifold is locally compact Hausdorff.

The last example shows that every manifold has a basis of precompact open sets. For later use, we will need the following refinement of that fact. Let M be an n-manifold. A Euclidean ball $B \subset M$ is called *regular* if it has the following properties:

(i) There is a Euclidean ball $B' \subset M$ containing $\overline{B}$.

(ii) For some $r > 0$, there is a chart $\varphi\colon B' \to B_{2r}(0) \subset \mathbb{R}^n$ that sends $\overline{B}$ onto $\overline{B}_r(0)$.

Lemma 4.31. *Every manifold has a countable basis of regular Euclidean balls.*

Proof. Let M be an n-manifold. Every point of M is contained in a Euclidean neighborhood, and since M is second countable, a countable collection $\mathcal{U} = \{U_i : i \in \mathbb{N}\}$ of such neighborhoods covers M by Lemma 2.15. For each of these open sets U_i, choose a homeomorphism φ_i from U_i to an open set $\widetilde{U}_i \subset \mathbb{R}^n$.

Now let $\mathcal{B}$ be the collection of all open subsets of M of the form $\varphi_i^{-1}(B_r(x))$, where $x \in \widetilde{U}_i$ is a point with rational coordinates and r is

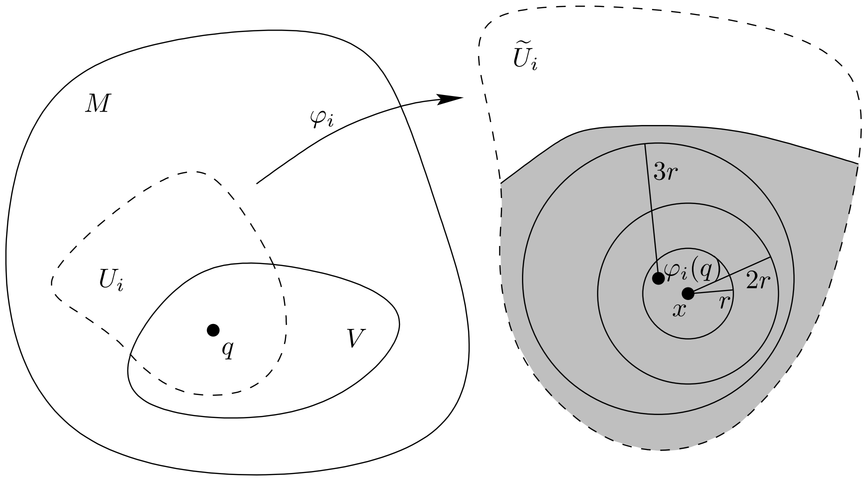


FIGURE 4.10. Showing that the collection $\mathcal{B}$ is a basis.

any positive rational number such that $B_{2r}(x) \subset \widetilde{U}_i$. Since there are only countably many such balls for each U_i, the collection $\mathcal{B}$ is countable. For any such set $B = \varphi_i^{-1}(B_r(x)) \in \mathcal{B}$, let $B' = \varphi_i^{-1}(B_{2r}(x))$. To show that B is a regular ball, we need to show that $\overline{B} \subset B'$ and $\varphi_i(\overline{B}) = \overline{B}_r(x)$. We will show, equivalently, that $\varphi_i^{-1}(\overline{B}_r(x)) = \overline{B}$. Now, $\varphi_i^{-1}(\overline{B}_r(x))$ is compact and therefore closed in M (because M is Hausdorff), so $\overline{B} \subset \varphi_i^{-1}(\overline{B}_r(x)) \subset B'$. This means that the closure of B in M is equal to its closure in B', which is $\varphi_i^{-1}(\overline{B}_r(x))$, since φ_i is a homeomorphism.

To show that the collection $\mathcal{B}$ is a basis, it suffices to show that each open subset of M satisfies the basis criterion with respect to it. Let $V \subset M$ be any open subset and $q \in V$. Then $q \in U_i$ for some i, and $\varphi_i(V \cap U_i)$ is an open subset of $\widetilde{U}_i$ containing $\varphi_i(q)$ (Figure 4.10). Choose a rational number $r > 0$ small enough that $B_{3r}(\varphi_i(q)) \subset \varphi_i(V \cap U_i)$, and then choose a point $x \in \varphi_i(V \cap U_i)$ with rational coordinates such that $|x - \varphi_i(q)| < r$. Then $\varphi_i(q) \in B_r(x)$, and it follows from the triangle inequality that $B_{2r}(x) \subset \varphi_i(V \cap U_i)$. Therefore, $B = \varphi_i^{-1}(B_r(x))$ is in $\mathcal{B}$, contains q, and is contained in V, thus completing the proof. $\square$

The closed map lemma is powerful, but it applies only when the domain is compact. The following proposition provides a useful generalization of the closed map lemma to noncompact spaces. A continuous map is said to be *proper* if the inverse image of each compact subset of Y is compact.

Proposition 4.32. *Suppose $f\colon X \to Y$ is a continuous map between locally compact Hausdorff spaces. If f is proper, it is a closed map.*

Proof. Let $K \subset X$ be a closed set. We will show that $f(K)$ contains all of its boundary points, which means that it is closed in Y.

If $y \in Y$ is a boundary point of $f(K)$, let U be a precompact neighborhood of y. An easy verification shows that y is also a boundary point of $f(K) \cap \overline{U}$. Because f is proper, $f^{-1}(\overline{U})$ is compact, which implies that $K \cap f^{-1}(\overline{U})$ is compact. By continuity, $f(K \cap f^{-1}(\overline{U})) = f(K) \cap \overline{U}$ is compact and therefore closed in Y. In particular, $y \in f(K) \cap \overline{U} \subset f(K)$, so $f(K)$ is closed. □

Theorem 4.33 (Baire Category Theorem). *In a locally compact Hausdorff space or a complete metric space, any countable collection of dense open subsets has dense intersection.*

Proof. Suppose $\{V_n\}_{n\in\mathbb{N}}$ is a countable collection of dense open subsets of such a space X. We need to show that if $U \subset X$ is a nonempty open subset, the intersection of U with $\bigcap_n V_n$ is nonempty.

First consider the case in which X is locally compact Hausdorff. Since V_1 is dense, $U \cap V_1$ is nonempty, so by the shrinking lemma there is a nonempty precompact open set W_1 such that $\overline{W}_1 \subset U \cap V_1$. Similarly, there is a nonempty precompact open set W_2 such that $\overline{W}_2 \subset W_1 \cap V_2 \subset U \cap V_1 \cap V_2$. Continuing by induction, we obtain a sequence of nested nonempty compact sets $\overline{W}_1 \supset \overline{W}_2 \supset \cdots \supset \overline{W}_n \supset \cdots$ such that $\overline{W}_n \subset U \cap V_1 \cap \cdots \cap V_n$. By Exercise 4.4, there is a point $x \in \bigcap_n \overline{W}_n$, which is clearly in $U \cap \bigcap_n V_n$ as well.

In the case that X is a complete metric space, we modify the above proof as follows. At the inductive step, since $W_{n-1} \cap V_n$ is open and nonempty, there is some ball $B_{\varepsilon_n}(x_n)$ contained in the intersection. Choosing $r_n < \min(\varepsilon_n, 1/n)$, we obtain a sequence of nested closed balls such that $\overline{B}_{r_n}(x_n) \subset U \cap V_1 \cap \cdots \cap V_n$. Because $r_n \to 0$, the centers $\{x_n\}$ form a Cauchy sequence, which converges to a point $x \in U \cap \bigcap_n V_n$. □

The Baire category theorem has a useful complementary reformulation. A subset F of a topological space X is said to be *nowhere dense* if its closure contains no nonempty open set.

Corollary 4.34. *In a locally compact Hausdorff space or a complete metric space, any countable collection of nowhere dense sets has empty interior.*

Proof. Let X be such a space, and let $\{F_n\}$ be a countable collection of nowhere dense subsets of X. Replacing each F_n by its closure, we may assume that the sets are closed. Then their complements U_n are open and dense, so by the Baire category theorem $\bigcap_n U_n$ is dense. It follows that $\bigcup_n F_n = X \smallsetminus \bigcap_n U_n$ cannot contain any nonempty open set. □

For example, it is easy to show that the solution set to any polynomial equation in two variables is nowhere dense in $\mathbb{R}^2$. Since there are only countably many polynomials with rational coefficients, this corollary implies that there are points in the plane (a dense set of them, in fact) that satisfy no rational polynomial equation.

The name of the theorem derives from the (astonishingly unedifying) terminology used by Baire: He defined a set of the *first category* to be a countable union of nowhere dense sets, and a set of the *second category* to be any set that is not of the first category. The theorem proved by Baire was that for spaces satisfying the hypothesis, every open set is of the second category. Although the category terminology is mostly ignored nowadays, the name of the theorem has stuck.

As we mentioned in Chapter 3, quotient maps do not generally behave well with respect to products. In particular, it is not always true that the product of two quotient maps is again a quotient map. However, it turns out that the product of a quotient map with the identity map of a locally compact Hausdorff space is indeed a quotient map, as the next lemma shows. This will be used in Chapter 7; the proof is rather technical and can safely be skipped on first reading.

Lemma 4.35. *Suppose $\pi\colon X \to Y$ is a quotient map and K is a locally compact Hausdorff space. The map $\pi \times \mathrm{Id}\colon X \times K \to Y \times K$ is a quotient map.*

Proof. We need to show that $\pi \times \mathrm{Id}$ takes saturated open sets in $X \times K$ to open sets in $Y \times K$. Let $U \subset X \times K$ be a saturated open set. Given $(x_0, k_0) \in U$, we will show that (x_0, k_0) has a saturated product neighborhood $W \times J$ contained in U. It then follows that $\pi(W) \times J$ contains $(\pi(x_0), k_0)$, is contained in $\pi \times \mathrm{Id}(U)$, and is open (since $\pi(W)$ is the image of a saturated open set under the quotient map π). Thus $\pi \times \mathrm{Id}(U)$ is open in $Y \times K$.

Now we proceed to prove the existence of the desired saturated product neighborhood. For any subset $W \subset X$, we define its *saturation* to be $\mathrm{Sat}(W) = \pi^{-1}(\pi(W))$; it is the smallest saturated subset containing W.

By definition of the product topology, (x_0, k_0) has a product neighborhood $W_0 \times J_0 \subset U$. By the shrinking lemma, there is a precompact neighborhood J of k_0 such that $\overline{J} \subset J_0$, and thus $(x_0, k_0) \in W_0 \times \overline{J} \subset W_0 \times J_0 \subset U$ (Figure 4.11). Because U is saturated, it follows that $\mathrm{Sat}(W_0) \times \overline{J} \subset U$. Now, $\mathrm{Sat}(W_0) \times J$ is a saturated subset of $X \times K$, but not necessarily open (since π may not be an open map).

We will show that there exists an open set $W_1 \subset X$ containing $\mathrm{Sat}(W_0)$ such that $W_1 \times \overline{J} \subset U$. To prove this, fix some $x \in \mathrm{Sat}(W_0)$. For any $k \in \overline{J}$, (x, k) has a product neighborhood in U. Finitely many of these cover the compact set $\{x\} \times \overline{J}$; call them $V_1 \times J_1, \ldots, V_m \times J_m$. If we set $V_x = V_1 \cap \cdots \cap V_m$, then V_x is a neighborhood of $\{x\}$ such that $V_x \times \overline{J} \subset U$. Taking W_1 to be the union of all such sets V_x for $x \in \mathrm{Sat}(W_0)$ proves the claim.

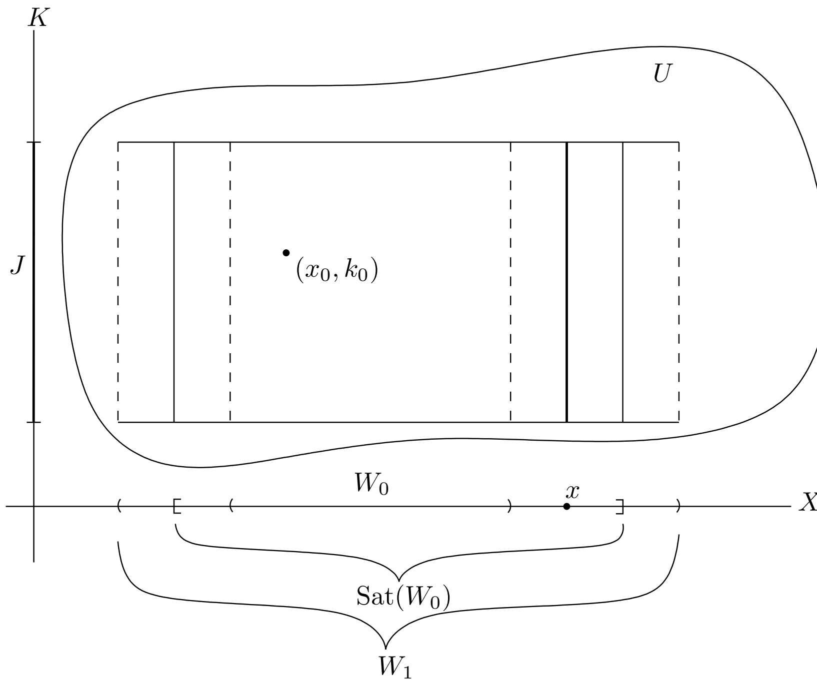


FIGURE 4.11. Finding a saturated product neighborhood.

Repeating this construction, we obtain a sequence of open sets $W_i \subset X$ such that

$$W_0 \subset \mathrm{Sat}(W_0) \subset W_1 \subset \mathrm{Sat}(W_1) \subset \cdots$$

and $W_i \times \overline{J} \subset U$. Let W be the union of all the W_i's. Then W is open because it is a union of open sets, and $W \times J \subset U$. Moreover, $W \times J$ is saturated: If $(x, k) \in W \times J$, then x is in some W_i; and if (x', k) is any point in the same fiber, then $x' \in W_{i+1}$, so $(x', k) \in W \times J$ as well. Thus $W \times J$ is the required saturated product neighborhood of (x_0, k_0). $\square$

Problems

4-1. (a) If U is any open subset of $\mathbb{R}$ and $x \in U$, show that $U \smallsetminus \{x\}$ is disconnected.

(b) Show that a topological space cannot be both a 1-manifold and an n-manifold for any $n > 1$.

4-2. Show that the union of the x-axis and the y-axis in $\mathbb{R}^2$ is not a manifold in the subspace topology.

4-3. Show that any n-manifold is a disjoint union of countably many connected n-manifolds.

4-4. Suppose $f\colon X \to Y$ is a surjective local homeomorphism. If X is locally connected, locally path connected, or locally compact, show that Y has the same property.

4-5. Let X be the topologist's sine curve (Example 4.10).

(a) Show that X is connected but not path connected or locally connected.

(b) Determine the components and the path components of X.

4-6. Like Problem 3-7, this problem constructs a space that is locally Euclidean and Hausdorff but not second countable. Unlike that example, however, this one is connected.

(a) Recall that a totally ordered set is said to be well-ordered if every nonempty subset has a smallest element (see the Appendix). Show that the well-ordering theorem (Theorem A.2) implies that there exists an uncountable well-ordered set Y such that for every $p \in Y$, there are only countably many $q < p$. [Hint: Let X be any uncountable well-ordered set. If X does not satisfy the desired condition, let Y be an appropriate subset of X.]

(b) Now let

$$\mathfrak{L} = \big(Y \times [0,1)\big) \smallsetminus \{(a_0, 0)\},$$

where a_0 is the smallest element of Y. We give $\mathfrak{L}$ the *dictionary order*: This means that $(p,q) < (r,s)$ if either $p < r$, or $p = r$ and $q < s$. With the order topology, $\mathfrak{L}$ is called the *long line*. Show that $\mathfrak{L}$ is locally Euclidean and Hausdorff but not second countable.

(c) Show that $\mathfrak{L}$ is path connected.

4-7. Define a topology on $\mathbb{Z}$ by declaring a set A to be open if and only if $n \in A$ implies $-n \in A$. Show that $\mathbb{Z}$ with this topology is second countable and limit point compact but not compact.

4-8. Let V be a finite-dimensional real vector space. A *norm* on V is a real-valued function on V, written $v \mapsto |v|$, satisfying

- Positivity: $|v| \geq 0$, and $|v| = 0$ if and only if $v = 0$.
- Homogeneity: $|cv| = |c|\,|v|$ for any $c \in \mathbb{R}$ and $v \in V$.
- Triangle inequality: $|v + w| \leq |v| + |w|$.

A norm determines a metric by $d(v, w) = |v - w|$. Show that all norms determine the same topology on V. [Hint: Consider the restriction of the norm to the unit sphere.]

4-9. Suppose K and L are compact convex sets in $\mathbb{R}^n$, both with nonempty interior. Show that any continuous map $f\colon \partial K \to \partial L$ has a continuous extension to a map $F\colon K \to L$. If f is a homeomorphism, show that F can be chosen to be a homeomorphism also.

4-10. Let X be a noncompact, locally compact Hausdorff space. The *one-point compactification* of X is the topological space X^* defined as follows. Let ∞ be some object not in X, and let $X^* = X \amalg \{\infty\}$ with the following topology:

$$\begin{aligned}\mathcal{T} = \{&\text{open subsets of } X\}\\ &\cup \{U \subset X^* : X^* \smallsetminus U \text{ is a compact subset of } X\}.\end{aligned}$$

(a) Show that $\mathcal{T}$ is a topology.

(b) Show that X^* is a compact Hausdorff space.

(c) Show that X is open and dense in X^* and has the subspace topology.

4-11. If X and Y are noncompact, locally compact Hausdorff spaces, show that a continuous map $f\colon X \to Y$ extends to a continuous map $f^*\colon X^* \to Y^*$ if and only if it is proper.

4-12. Let $\sigma : \mathbb{S}^n \smallsetminus \{N\} \to \mathbb{R}^n$ be stereographic projection, as defined in Example 3.6. Show that σ extends to a homeomorphism of $\mathbb{S}^n$ with the one-point compactification of $\mathbb{R}^n$.

4-13. If M is a noncompact n-manifold, show that its one-point compactification is an n-manifold if and only if there exists a precompact open subset $U \subset M$ such that $M \smallsetminus U$ is homeomorphic to $\mathbb{R}^n \smallsetminus \mathbb{B}^n$. [Hint: You may find the inversion map $\mathcal{I} : \mathbb{R}^n \smallsetminus \mathbb{B}^n \to \overline{\mathbb{B}^n}$ defined by $\mathcal{I}(x) = x/|x|^2$ useful.]

4-14. Suppose M is a 1-dimensional manifold with boundary. Show that the interior and boundary of M are disjoint. Use this to conclude that M is a manifold if and only if $\partial M = \varnothing$.

5

Simplicial Complexes

In this chapter we give a brief introduction to simplicial complexes. These are spaces constructed from building blocks called simplices, which are points, line segments, filled-in triangles, solid tetrahedra, and their higher-dimensional analogues. They provide a highly useful way of constructing topological spaces, and play a fundamental role in geometry and algebraic topology.

As we did with manifolds, we will define simplicial complexes in two stages, starting with a very concrete version and proceeding to the most general definition. Concretely, we think of a simplicial complex as a collection of simplices in some Euclidean space that overlap "nicely." More abstractly, a simplicial complex is an abstract "vertex scheme," specifying which sets of vertices are supposed to span simplices. We will see that any abstract simplicial complex determines a topological space, called a polyhedron, in a natural way.

Then we apply these ideas to manifolds by asking which manifolds are homeomorphic to polyhedra. Any such homeomorphism is called a triangulation of the manifold, and any manifold that admits such a homeomorphism is said to be triangulable. We will give a complete proof that every 1-manifold is triangulable, and will give a brief sketch of the proof for 2-manifolds. These results will be used in the next chapter as stepping stones toward classifying curves and surfaces up to homeomorphism.

At the end of the chapter we explore two combinatorial properties of simplicial complexes that are important in the study of manifolds. The first is the concept of an orientation of a complex, which generalizes and systematizes the intuitive notions of "direction" in 1-dimensional complexes,

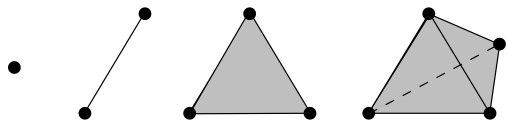


FIGURE 5.1. Simplices.

"clockwise" and "counterclockwise" in 2-dimensional ones, and "handedness" in 3-dimensional ones. The second is the Euler characteristic, which is the alternating sum of the numbers of simplices in different dimensions, and generalizes Euler's classical formula for compact convex polyhedra in $\mathbb{R}^3$.

Euclidean Simplicial Complexes

We begin with a little linear algebra. An *affine map* between vector spaces is a map $f\colon V \to W$ of the form $f(x) = a(x) + b$, where a is a linear map and $b \in W$. An *affine subspace* of a vector space is the zero set of some affine map: $\{x : a(x) + b = 0\}$. Its *dimension* is the dimension of the kernel of the linear part of the affine map. The special case of an affine subspace of V whose dimension is one less than that of V is called an *affine hyperplane* in V. Elementary linear algebra shows that if $n \geq k$, any $k+1$ points $v_0, \dots, v_k$ in $\mathbb{R}^n$ are contained in some k-dimensional affine subspace (just choose a linear map $a\colon \mathbb{R}^n \to \mathbb{R}^{n-k}$ whose kernel contains $\{v_1 - v_0, \dots, v_k - v_0\}$ and let $b = -a(v_0)$). We say that $k+1$ points are in *general position* if they are not contained in any $(k-1)$-dimensional affine subspace, or equivalently if $\{v_1 - v_0, \dots, v_k - v_0\}$ are linearly independent.

Given points $v_0, \dots, v_k$ in general position in $\mathbb{R}^n$, the *simplex* (plural: simplices) spanned by them is the set of all points in $\mathbb{R}^n$ of the form

$$\sum_{i=0}^{k} t_i v_i, \quad \text{where } 0 \leq t_i \leq 1 \text{ and } \sum_{i=0}^{k} t_i = 1, \tag{5.1}$$

with the subspace topology. Each of the points v_i is called a *vertex* of the simplex. We will sometimes use the notation $\langle v_0, \dots, v_k \rangle$ to denote the simplex spanned by $v_0, \dots, v_k$. The integer k (one less than the number of vertices) is called its *dimension*, and a k-dimensional simplex is often called a *k-simplex.* A 0-simplex is a single point, a 1-simplex is a line segment, a 2-simplex is a (filled-in) triangle, and a 3-simplex is a solid tetrahedron (Figure 5.1).

For any subset $A \subset \mathbb{R}^n$, the *convex hull* of A is defined to be the intersection of all convex sets containing A. It is immediate that the convex hull is itself a convex set, in fact, the smallest convex set containing A.

Lemma 5.1. *A simplex is the convex hull of its vertices.*

Exercise 5.1. Prove Lemma 5.1.

Let σ be a simplex. Each simplex spanned by a nonempty subset of the vertices is called a *face* of σ. The faces that are not equal to σ itself are called its *proper faces*. The 0-dimensional faces of σ are just its vertices, and the 1-dimensional faces are called its *edges*. The $(k-1)$-dimensional faces of a k-simplex are sometimes called its *boundary faces*.

A map $f\colon \sigma \to \tau$ between simplices is called a *simplicial map* if it is the restriction of an affine map that takes vertices of σ to vertices of τ. As the next exercise shows, simplicial maps between a given pair of simplices are in one-to-one correspondence with maps between their vertices.

Exercise 5.2.

(a) Show that given any map f_0 from the set of vertices of σ to the set of vertices of τ, there is a unique simplicial map $f\colon \sigma \to \tau$ whose restriction to the vertices of σ is f_0.

(b) Show that any two k-simplices are homeomorphic by a simplicial homeomorphism.

(c) Show that every k-simplex is homeomorphic to $\overline{\mathbb{B}^k}$. [Hint: Work with a particular simplex in $\mathbb{R}^k$ and use Proposition 4.26.]

It follows from part (c) of the preceding exercise that a k-simplex is a k-dimensional manifold with boundary. Thus we define the *boundary* of a simplex to be the union of its boundary faces (which is the same as the union of all of its proper faces), and its *interior* to be the simplex minus its boundary. The interior of a k-simplex is sometimes called an *open k-simplex*; it is the set of points of the form $\sum t_i v_i$ where $\{v_0, \dots, v_k\}$ are the vertices of σ and none of the coefficients t_i are zero. For example, if σ is a 0-simplex, $\operatorname{Int}\sigma = \sigma$, and if σ is a 1-simplex, $\operatorname{Int}\sigma$ is σ minus its vertices. Note that an open simplex is generally not an open subset of $\mathbb{R}^n$, and the interior and boundary of σ as a simplex may not be equal to its topological interior and boundary as a subset of $\mathbb{R}^n$.

A *Euclidean simplicial complex* is a collection K of simplices in some Euclidean space $\mathbb{R}^n$ satisfying the following conditions:

(i) If $\sigma \in K$, then every face of σ is in K.

(ii) The intersection of any two simplices in K is either empty or a face of each.

(iii) Local Finiteness: Every point in a simplex of K has a neighborhood that intersects at most finitely many simplices of K.

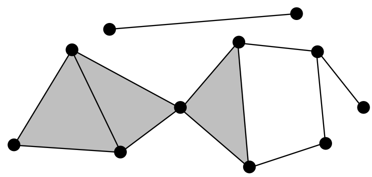


FIGURE 5.2. A complex in $\mathbb{R}^2$.

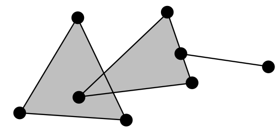


FIGURE 5.3. Not a complex.

The *dimension* of K is defined to be the maximum dimension of any simplex in K (which is well-defined, since simplices in $\mathbb{R}^n$ have dimension at most n). Figure 5.2 shows an example of a 2-dimensional simplicial complex in $\mathbb{R}^2$. The set of simplices shown in Figure 5.3 is not a simplicial complex, because the intersection condition is violated.

Given a Euclidean complex K, the union of all the simplices in K, with the subspace topology inherited from $\mathbb{R}^n$, is a topological space denoted by $|K|$ and called the (*Euclidean*) *polyhedron* of K.

Many of the spaces we have seen so far are homeomorphic to Euclidean polyhedra. Here are some simple examples.

Example 5.2 (Euclidean Polyhedra).

(a) Any n-simplex together with its faces is a simplicial complex whose polyhedron is homeomorphic to $\overline{\mathbb{B}^n}$.

(b) The proper faces of an n-simplex constitute an $(n-1)$-dimensional complex whose polyhedron is homeomorphic to $\mathbb{S}^{n-1}$.

(c) The set of all unit-length intervals $[n, n+1] \subset \mathbb{R}$ for $n \in \mathbb{Z}$, together with their endpoints, is a simplicial complex whose polyhedron is $\mathbb{R}$.

(d) For any integer $m \geq 3$, let P_m be a regular m-sided polygon in the plane. The set of edges and vertices of P_m is a simplicial complex whose polyhedron is homeomorphic to $\mathbb{S}^1$.

Example 5.3. The set of closed line segments in the plane from the origin to the points $(1, 1/n)$ for $n \in \mathbb{N}$, together with their vertices (Figure 5.3), is *not* a simplicial complex, because the local finiteness condition fails at the origin.

Exercise 5.3. Prove the claims made in the two preceding examples.

You might wonder why we should focus on building spaces out of simplices, and not out of cubes or some other sort of geometric object. The simple answer is that simplicial complexes are the most general: It is not hard to show that a locally finite "polyhedral complex" (under any reasonable definition) can be subdivided to form a simplicial complex.

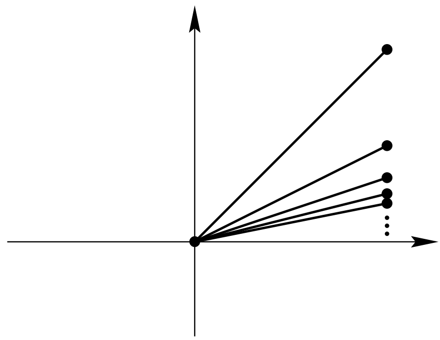


FIGURE 5.4. Failure of local finiteness.

Let K be a Euclidean simplicial complex. Any subset $K' \subset K$ that is itself a simplicial complex is called a *subcomplex* of K. It is clear that the only condition that needs to be checked is that the faces of each simplex in K' are in K'. In particular, for any nonnegative integer k, the subset $K^{(k)} \subset K$ consisting of all simplices of dimension less than or equal to k is a subcomplex, called the *k-skeleton* of K.

Let K and L be two Euclidean simplicial complexes. A continuous map $f\colon |K| \to |L|$ whose restriction to each simplex of K is a simplicial map to a simplex of L is called a *simplicial map*, and is denoted by $f\colon K \to L$. The restriction of f to $K^{(0)}$ is called the *vertex map* of f.

Exercise 5.4. Let K and L be Euclidean simplicial complexes.

(a) Let $f_0\colon K^{(0)} \to L^{(0)}$ be any map with the property that whenever $\{v_0, \dots, v_k\}$ are the vertices of a simplex of K, $\{f_0(v_0), \dots, f_0(v_k)\}$ are the vertices of a simplex of L (possibly with repetitions). Show that there is a unique simplicial map $f\colon K \to L$ whose vertex map is f_0.

(b) Now let f_0 be as in (a), and assume in addition that f_0 is bijective and $\{v_0, \dots, v_k\}$ are the vertices of a simplex of K if and only if $\{f_0(v_0), \dots, f_0(v_k)\}$ are the vertices of a simplex of L. Show that $|K|$ and $|L|$ are homeomorphic by a simplicial map.

All the considerations of this section carry over without change if we replace $\mathbb{R}^n$ by an arbitrary finite-dimensional vector space V. We give V the metric topology induced by any norm; Problem 4-8 shows that the resulting topology is independent of the norm. The only properties of $\mathbb{R}^n$ that we use are its vector space structure and its topology, and since any choice of basis gives a linear homeomorphism of V with $\mathbb{R}^n$, all the results of this section are true with $\mathbb{R}^n$ replaced by V. We will use this slightly more general setting in the next section.

Abstract Simplicial Complexes

Just as it is too restrictive to define manifolds to be subsets of Euclidean spaces, Euclidean simplicial complexes are not sufficiently general for many important applications. In this section we will define a more general kind of simplicial complex. The key idea is already implicit in Exercise 5.4(b), which says that a simplicial complex is completely determined, up to simplicial homeomorphism, by knowledge of which sets of vertices span simplices.

Motivated by this observation, we define an *abstract simplicial complex* to be a collection $\mathcal{K}$ of nonempty finite sets called (*abstract*) *simplices*, subject only to one condition: If $\sigma \in \mathcal{K}$, then every nonempty subset of σ is in $\mathcal{K}$. Any element of a simplex $\sigma \in \mathcal{K}$ is called a *vertex* of σ, and any nonempty subset of σ is called a *face* of σ. (We make no distinction between a vertex v and the corresponding face $\{v\}$.) To distinguish the simplices we defined earlier (as convex subsets of some Euclidean space) from abstract simplices in this sense, we will sometimes refer to the former as *Euclidean simplices.*

The *dimension* of an abstract simplex consisting of $k+1$ vertices is defined to be k. The dimension of $\mathcal{K}$ is the maximum dimension of any simplex in $\mathcal{K}$, if it exists; if there are simplices of arbitrarily high dimensions, $\mathcal{K}$ is said to be *infinite-dimensional.* We say that $\mathcal{K}$ is a *finite complex* if $\mathcal{K}$ is a finite set, and *locally finite* if every vertex belongs to only finitely many simplices.

A subset of $\mathcal{K}$ that is itself a simplicial complex (i.e., that contains all the faces of each of its simplices) is called a *subcomplex* of $\mathcal{K}$. The set $\mathcal{K}^{(k)}$ of all simplices of dimension at most k is a k-dimensional subcomplex called the *k-skeleton* of $\mathcal{K}$.

Given two abstract complexes $\mathcal{K}, \mathcal{L}$, a map $f\colon \mathcal{K} \to \mathcal{L}$ is called a *simplicial map* if it is of the form $f(\{v_0, \dots, v_k\}) = \{f_0(v_0), \dots, f_0(v_k)\}$ for some map $f_0\colon \mathcal{K}^{(0)} \to \mathcal{L}^{(0)}$, called the *vertex map* of f (which must have the property that $\{f(v_0), \dots, f(v_k)\} \in \mathcal{L}$ whenever $\{v_0, \dots, v_k\} \in \mathcal{K}$). A simplicial map f is called an *isomorphism* if f_0 is a bijection and $\{v_0, \dots, v_k\}$ is a simplex of $\mathcal{K}$ if and only if $\{f_0(v_0), \dots, f_0(v_k)\}$ is a simplex of $\mathcal{L}$.

One way of constructing an abstract simplicial complex, as you have probably already guessed, is the following. Given a Euclidean simplicial complex K, let $\mathcal{K}$ denote the set of all those finite subsets $\{v_0, \dots, v_k\} \subset K^{(0)}$ that consist of the vertices of some simplex of K. It is immediate that $\mathcal{K}$ is an abstract simplicial complex, called the *vertex scheme* of K. It is an immediate consequence of Exercise 5.4(b) that two Euclidean complexes are simplicially homeomorphic if and only if their vertex schemes are isomorphic.

Exercise 5.5. Show that every finite abstract complex is the vertex scheme of a Euclidean simplicial complex. [Hint: Use basis vectors $e_i = (0, \dots, 1, \dots, 0)$ as vertices.]

Not all abstract simplicial complexes are vertex schemes of Euclidean complexes, however. Such an abstract complex must obviously be finite-dimensional and locally finite. Moreover, since the local finiteness condition forces the vertex set of a Euclidean complex to be a discrete subset of $\mathbb{R}^n$, its vertex scheme can have only countably many simplices. Problem 5-5 shows that these conditions are also sufficient.

The theory of quotient spaces gives a useful way of constructing topological spaces out of abstract complexes without these restrictions. The first step is to construct a canonical Euclidean k-simplex for each abstract k-simplex. Using equation (5.1) as a guide, we wish to think of our simplex as "a set of points of the form $\sum t_i v_i$." The trouble is that the vertices v_i of an abstract simplex are just abstract objects and not points in some Euclidean space, so this expression no longer makes literal sense as a vector sum. Instead, we consider such a sum as a "formal linear combination" of the vertices v_i. (The word "formal" is used here to indicate that the expression has the form of a linear combination, but may not actually represent addition of vectors in a vector space.) To make this precise, we introduce a bit of algebraic terminology.

Given a set S, we wish to define a vector space whose elements we can think of as "formal linear combinations" of the elements of S. The main property of such a linear combination is that it is completely determined by the coefficient t attached to each $v \in S$. Thus we are led to the following definition: A *formal linear combination* of elements of S is a function $t\colon S \to \mathbb{R}$ such that $t(v) = 0$ for all but finitely many $v \in S$. Under the operations of pointwise addition and multiplication by constants, the set of all such functions is a vector space, denoted by $\mathbb{R}\langle S\rangle$ and called the *free vector space* on S.

Any element $t \in \mathbb{R}\langle S\rangle$ can be represented symbolically as

$$t = \sum_{i=0}^{k} t_i v_i, \tag{5.2}$$

where v_i are the (finitely many) elements of S for which $t(v) \neq 0$, and $t_i = t(v_i)$. To be a bit more precise, each $v \in S$ determines in a natural way a function from S to $\mathbb{R}$, also denoted by v for simplicity, given by

$$v(w) = \begin{cases} 1, & w = v, \\ 0, & w \neq v. \end{cases}$$

It is easy to check that each $t \in \mathbb{R}\langle S\rangle$ has a unique expression as a finite linear combination of these functions; this is the appropriate way to interpret (5.2).

Now consider any abstract simplex $\{v_0, \dots, v_k\}$. We define its *geometric realization* to be the k-simplex $\langle v_0, \dots, v_k\rangle$ in the finite-dimensional vector space $\mathbb{R}\langle v_0, \dots, v_k\rangle$. With the topology on $\mathbb{R}\langle v_0, \dots, v_k\rangle$ induced by any norm, this geometric realization is homeomorphic to a Euclidean k-simplex.

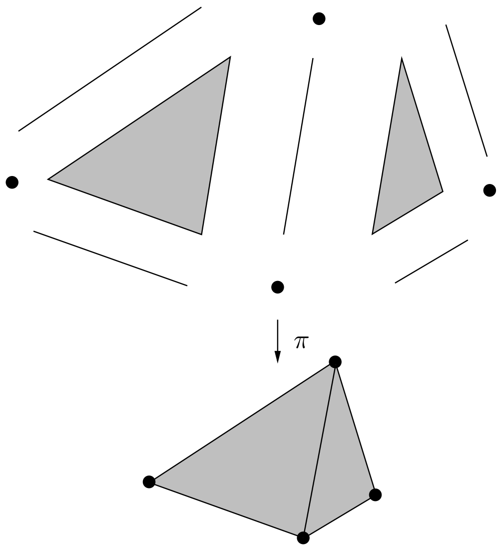


FIGURE 5.5. The quotient map defining the topology of $|\mathcal{K}|$.

Since each abstract simplex determines its geometric realization and vice versa, we will sometimes use the term "simplex" and the notation $\langle v_0, \ldots, v_k \rangle$ interchangeably to refer either to an abstract simplex or to its geometric realization. When we need to distinguish between the two, we will use the notation $|\sigma|$ for the geometric realization of σ. As in the Euclidean case, the open simplex $\operatorname{Int}|\sigma|$ is the subset of $|\sigma|$ consisting of points all of whose coefficients t_i are nonzero.

For any abstract simplicial complex $\mathcal{K}$, let $|\mathcal{K}|$ denote the set of *all* formal linear combinations of the form $\sum_{i=0}^k t_i v_i$ with $\langle v_0, \ldots, v_k \rangle$ a simplex of $\mathcal{K}$ and with coefficients satisfying $0 \le t_i \le 1$ and $\sum_{i=0}^k t_i = 1$. This can be thought of abstractly as a subset of the free vector space $\mathbb{R}\langle \mathcal{K}^{(0)} \rangle$ on the vertex set of $\mathcal{K}$. More concretely, $|\mathcal{K}|$ is just the union of all the geometric realizations of the simplices of $\mathcal{K}$, with points in two simplices identified whenever they have the same expression as linear combinations of vertices.

We topologize $|\mathcal{K}|$ in the following way. Let $\coprod_{\sigma \in \mathcal{K}} |\sigma|$ be the *disjoint* union of the geometric realizations of all the simplices of $\mathcal{K}$, with the disjoint union topology as in Problem 2-9 (this just means that a set is open in $\coprod_{\sigma \in \mathcal{K}} |\sigma|$ if and only if its intersection with each $|\sigma|$ is open in $|\sigma|$), and let

$$\pi \colon \coprod_{\sigma \in \mathcal{K}} |\sigma| \to |\mathcal{K}|$$

be the natural map that sends each simplex $|\sigma|$ to itself (Figure 5.5). We give $|\mathcal{K}|$ the quotient topology with respect to π. Unwinding the definitions, this is the same as saying that a subset of $|\mathcal{K}|$ is open (or closed) if and only if its intersection with each simplex is open (or closed). With this topology, $|\mathcal{K}|$ is called the *geometric realization* of $\mathcal{K}$.

This way of characterizing a topology turns out to be of great importance, so it has a name. Given any collection $\{S_\alpha\}_{\alpha\in A}$ of subspaces of a topological space X whose union is X, the topology of X is said to be *coherent* with the subspaces S_α if a set is open in X if and only if its intersection with each S_α is open in S_α. It is easy to check that this is equivalent to saying that a set is closed in X if and only if its intersection with each S_α is closed in S_α.

Lemma 5.4. *Let $\mathcal{K}$ be an abstract simplicial complex and $|\mathcal{K}|$ its geometric realization.*

(*a*) *Each simplex $|\sigma|$ is a closed, compact subset of $|\mathcal{K}|$.*

(*b*) *If $\dim \mathcal{K} = n$, then each open n-simplex $\operatorname{Int}|\sigma|$ is an open subset of $|\mathcal{K}|$.*

(*c*) *The topology of $|\mathcal{K}|$ is the unique topology coherent with the collection of subspaces $\{|\sigma| : \sigma \in \mathcal{K}\}$.*

(*d*) *A map $F\colon |\mathcal{K}| \to |\mathcal{L}|$ is continuous if and only if its restriction to $|\sigma|$ is continuous for each $\sigma \in \mathcal{K}$.*

Exercise 5.6. Prove Lemma 5.4.

Any simplicial map $f\colon \mathcal{K} \to \mathcal{L}$ between abstract complexes induces in an obvious way a map $|f| : |\mathcal{K}| \to |\mathcal{L}|$. (On each simplex $|\sigma|$, $|f|$ is just the Euclidean simplicial map determined by the vertex map of f.) Since the restriction of $|f|$ to each simplex is continuous, $|f|$ is a continuous map.

Lemma 5.5. *Let $\mathcal{K}$, $\mathcal{L}$, and $\mathcal{M}$ be simplicial complexes.*

(*a*) *If $\mathrm{Id}\colon \mathcal{K} \to \mathcal{K}$ denotes the identity map of $\mathcal{K}$, then $|\,\mathrm{Id}\,|$ is the identity map of $|\mathcal{K}|$.*

(*b*) *If $f\colon \mathcal{K} \to \mathcal{L}$ and $g\colon \mathcal{L} \to \mathcal{M}$ are simplicial maps, then $|g\circ f| = |g|\circ|f|$.*

(*c*) *Isomorphic complexes have homeomorphic geometric realizations.*

Exercise 5.7. Prove Lemma 5.5.

Lemma 5.6. *If $\mathcal{K}$ is the vertex scheme of a Euclidean simplicial complex K, then the geometric realization of $\mathcal{K}$ is homeomorphic to $|K|$.*

Proof. Recall that an abstract simplex $\sigma \in \mathcal{K}$ is just the set of vertices of some Euclidean simplex $\widetilde{\sigma} \in K$. Let π' denote the natural map

$$\pi'\colon \coprod_{\sigma\in\mathcal{K}} |\sigma| \to |K|,$$

which, restricted to each simplex $|\sigma|$, is the obvious simplicial homeomorphism $|\sigma| \to \widetilde{\sigma}$. This map makes the same identifications as the map π we used above to define the topology of $|\mathcal{K}|$. If we can show that π' is a quotient map, the lemma will follow from the uniqueness of quotient spaces. To show that it is a quotient map is the same as showing that $|K|$ has the topology coherent with its simplices.

To verify this, let $G \subset |K|$ be an arbitrary subset. If G is closed, then clearly its intersection with any simplex is closed, because it is an intersection of closed sets. Conversely, suppose the intersection of G with each simplex is closed. If $x \in |K|$ is any limit point of G, by local finiteness x has a neighborhood U that intersects only finitely many simplices. Thus $G \cap U$ is the union of finitely many closed subsets of U and hence closed in U. This implies $x \in G$, so G is closed in $|K|$. (This is the reason we insisted on local finiteness in the definition of Euclidean simplicial complexes.) □

Any topological space that is homeomorphic to the geometric realization of some simplicial complex is called a *polyhedron*. A particular such homeomorphism is called a *triangulation* of X. Any space that admits a triangulation (i.e., any polyhedron) is said to be *triangulable*. Sometimes one can obtain a better understanding of the topology of an unknown space by first showing that it is triangulable; this will be our approach, for example, to the classification of 1-dimensional and 2-dimensional manifolds.

Example 5.7. The following abstract complexes are isomorphic to the vertex schemes of the Euclidean complexes of Example 5.2, and therefore yield triangulations of the indicated spaces.

(a) The set of all nonempty subsets of $\{0, 1, 2, \ldots, n\}$ is an abstract complex whose geometric realization is homeomorphic to $\overline{\mathbb{B}^n}$.

(b) The set of all proper nonempty subsets of $\{0, 1, 2, \ldots, n\}$ is an abstract complex whose geometric realization is homeomorphic to $\mathbb{S}^{n-1}$.

(c) Let $\mathcal{K}_\infty$ be the abstract complex consisting of 0-simplices $\{\{n\} : n \in \mathbb{Z}\}$ and 1-simplices $\{\{n, n+1\} : n \in \mathbb{Z}\}$. Its geometric realization is homeomorphic to $\mathbb{R}$.

(d) For any integer $m \geq 3$, let $\mathcal{K}_m$ be the abstract complex whose 0-simplices are $\{\{1\}, \{2\}, \ldots, \{m\}\}$, and whose 1-simplices are $\{\{1, 2\}, \{2, 3\}, \ldots, \{m-1, m\}, \{m, 1\}\}$. Its geometric realization is homeomorphic to $\mathbb{S}^1$.

Example 5.8 (Graphs). We define a *graph* to be a 1-dimensional polyhedron with a given triangulation. (For some applications, it is useful to have a more general definition, allowing two edges to share more than one vertex, or one edge to begin and end at the same vertex; but this will suffice for our purposes. The kind of graph we have defined is sometimes called a *simple graph* to distinguish it from other more general types. Note that this use of the word graph has no relation to the graph of a function as defined in Chapter 3.) A *subgraph* of a graph is the polyhedron of a 1-dimensional subcomplex. A graph is said to be *finite* if its associated simplicial complex is finite.

We will illustrate the utility of simplicial complexes by showing how a topological property of polyhedra—connectedness—can be detected by purely combinatorial means. Let $\mathcal{K}$ be a simplicial complex. An *edge path* in $\mathcal{K}$ is a finite or infinite sequence of vertices such that any two consecutive vertices span an edge. An edge path is said to be *reduced* if in addition any three consecutive vertices are all distinct. (The idea is that a reduced edge path contains no "dead-end excursions" like v, w, v.) We say that $\mathcal{K}$ is *edge path connected* if any two vertices can be joined by a finite edge path.

Proposition 5.9. *Let $\mathcal{K}$ be a simplicial complex. Then $|\mathcal{K}|$ is connected if and only if $\mathcal{K}$ is edge path connected, in which case any two vertices can be joined by a reduced edge path.*

Proof. Suppose $\mathcal{K}$ is edge path connected. Because simplices are connected, any two vertices that span an edge lie in the same component of $|\mathcal{K}|$. It follows easily by induction that any two vertices joined by a finite edge path lie in the same component. Thus if $\mathcal{K}$ is edge path connected, all the vertices lie in the same component V_0 of $|\mathcal{K}|$. Any point $x \in |\mathcal{K}|$ lies in some simplex $|\sigma|$, and since $|\sigma|$ contains at least one vertex in common with V_0, it must be contained in V_0. This shows that $V_0 = |\mathcal{K}|$, so $|\mathcal{K}|$ is connected.

Conversely, suppose $|\mathcal{K}|$ is connected. Choose a vertex $v \in \mathcal{K}$, and let $\mathcal{C}$ denote the subcomplex of $\mathcal{K}$ consisting of all vertices that are traversed in edge paths starting from v, together with all the simplices of $\mathcal{K}$ they span. If σ is a simplex of $\mathcal{K}$ that has a vertex $w \in \mathcal{C}$, then every vertex w' of σ must lie in $\mathcal{C}$, because we can form an edge path from v to w' by starting with an edge path to w and then appending w' (since $\langle w, w' \rangle$ is an edge of σ). Thus $\sigma \in \mathcal{C}$ as well. It follows that $|\mathcal{C}|$ is both open and closed in $|\mathcal{K}|$, because its intersection with each simplex is either empty or the entire simplex. Thus $\mathcal{C} = \mathcal{K}$, which shows that $\mathcal{K}$ is edge path connected.

Now suppose $\mathcal{K}$ is edge path connected. Given any two vertices $v, v' \in \mathcal{K}$, there is an edge path $(v, \ldots, v')$ connecting them. If this edge path is not reduced, it must have three consecutive vertices of the form w, w', w for some pair of vertices w, w' that span an edge. It is easy to see that the sequence obtained by replacing these three vertices with the single vertex w is still an edge path connecting the same two vertices. Repeatedly shortening the edge path in this way until it is impossible to shorten it any more, we obtain a reduced edge path joining the same two vertices. $\square$

Triangulation Theorems

In the next chapter we will begin to study the problem of classifying manifolds up to homeomorphism. Our approach to classifying 1-manifolds and 2-manifolds will be to start with a triangulated manifold and study the combinatorial properties of the triangulation. For this we will need to know that all manifolds of dimensions 1 and 2 are triangulable.

Theorem 5.10 (Triangulation Theorem for 1-Manifolds). *Every* 1-*manifold can be triangulated by a* 1-*dimensional simplicial complex.*

Proof. We begin by showing that there exists a sequence of compact subspaces $G_n \subset M$, $n = 1, 2, \dots$, whose union is M, satisfying the following conditions:

(i) Each G_n is a finite graph.

(ii) For each n, G_n is a subgraph of G_{n+1}.

(iii) For each n there exists $m > n$ such that $G_n \subset \operatorname{Int} G_m$.

By Lemma 4.31, M admits a countable cover $\{B_i\}$ by regular Euclidean balls. From the definition of regular balls, it is evident that the closure of each regular ball is homeomorphic to a 1-simplex whose topological interior in M is equal to its interior as a simplex.

Begin by letting G_1 be the graph consisting of the single 1-simplex $\overline{B}_1$ and its vertices. Now let $n > 1$, and assume by induction that we have found finite graphs $G_1 \subset G_2 \subset \cdots \subset G_n$ satisfying (i) and (ii) with $G_n = \overline{B}_1 \cup \cdots \cup \overline{B}_n$. Consider the next 1-simplex $\overline{B}_{n+1}$. Some of the vertices of G_n may lie in B_{n+1} (the interior of $\overline{B}_{n+1}$); the ones that do define a subdivision of $\overline{B}_{n+1}$ into a finite graph S, with the property that no vertex of G_n lies in the interior of any edge of S.

For each of the edges $e \subset S$, we will prove the following claim: Either e intersects each of the edges of G_n only at vertices, or e is entirely contained in one of the edges of G_n. To prove this, suppose e has an interior point that lies in some edge $e' \subset G_n$ (Figure 5.6). By the remark above, it must be an interior point of e' as well.

Note that $\operatorname{Int} e \cap \operatorname{Int} e'$ is open in $\operatorname{Int} e$. On the other hand, e' is a compact subset of the Hausdorff space M, so it is closed in M, and therefore $\operatorname{Int} e \cap \operatorname{Int} e' = \operatorname{Int} e \cap e'$ is closed in $\operatorname{Int} e$. By connectedness, therefore, $\operatorname{Int} e \cap \operatorname{Int} e'$ is all of $\operatorname{Int} e$. In other words, $\operatorname{Int} e \subset \operatorname{Int} e'$, which implies $e \subset e'$ and proves the claim.

Now simply throw away those edges of S that are contained in G_n, and redefine S to be the graph consisting of the remaining edges and their vertices. Let $G_{n+1} = G_n \cup S$. We wish to show that G_{n+1} is a finite graph containing G_n as a subgraph. The edges of S intersect each of the edges of G_n only at vertices; but it may happen that both vertices of some edge

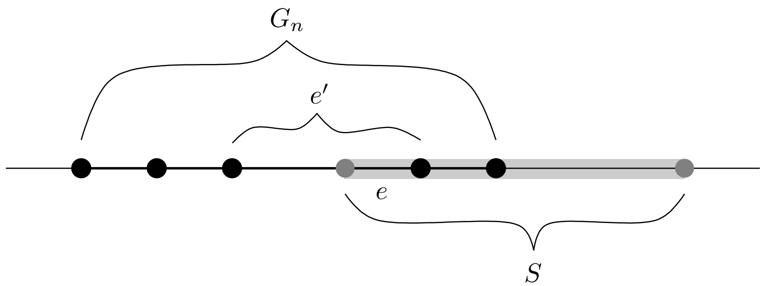


FIGURE 5.6. Proof that $e \subset e'$.

e in S intersect a single edge e' in G_n. If so, simply subdivide e into two edges by adding a new vertex in its interior. With this modification, each of the edges of S intersects each edge of G_n at most in a single vertex.

Since G_{n+1} has only finitely many simplices, its topology is coherent with the simplices by Problem 5-1. Thus the foregoing argument proves that G_{n+1} is a finite graph whose polyhedron is $\overline{B}_1 \cup \cdots \cup \overline{B}_{n+1}$ and that contains G_n as a subgraph. Continuing by induction, we obtain an increasing sequence $\{G_n : n = 1, 2, \dots\}$ of graphs such that every point of M is contained in G_n for some n. Since the interiors of the Euclidean balls B_i cover M, for each n there is some $m > n$ such that the compact set G_n is covered by $B_1 \cup \cdots \cup B_m$, and therefore $G_n \subset \operatorname{Int} G_m$. This completes the induction.

Let $\mathcal{K}$ be the abstract simplicial complex whose 0-skeleton is the union of the 0-skeletons of G_n for all n, and whose 1-simplices are the pairs $\{v, v'\}$ that span an edge in some G_n. There is an obvious bijective map $|\mathcal{K}| \to M$, defined by choosing a homeomorphism from each 1-simplex of $|\mathcal{K}|$ onto the corresponding edge in M. To see that this map is a homeomorphism, we need only show that M has the topology coherent with the simplices. Clearly, any set that is closed in M has closed intersection with each simplex, because the simplices have the subspace topology. Conversely, suppose $K \subset M$ is a subset whose intersection with each simplex is closed. If $x \in M$ is a limit point of K, choose n large enough that $x \in \operatorname{Int} G_n$. Since the intersection of K with each of the (finitely many) simplices of G_n is closed in G_n, it follows that $K \cap \operatorname{Int} G_n$ is closed in $\operatorname{Int} G_n$. In particular, $x \in K$, which proves that K is closed in M and thus the topology of M is coherent with the simplices. □

For use in the next chapter, we will need the following property of triangulated 1-manifolds.

Proposition 5.11. *If $\mathcal{K}$ is a simplicial complex whose geometric realization is a 1-manifold, each vertex of $\mathcal{K}$ lies on exactly two edges.*

Proof. Let v be any vertex, and let V be the union of $\{v\}$ together with the interiors of all the edges that have v as a vertex. Since the intersection of V with each simplex is open in the simplex, V is open in $|\mathcal{K}|$.

Because $|\mathcal{K}|$ is a 1-manifold, v has a neighborhood $U \subset V$ homeomorphic to an open interval. It follows that $U \smallsetminus \{v\}$ has exactly two components. For each edge e containing v, $\operatorname{Int} e \cap (U \smallsetminus \{v\})$ is an open subset of $U \smallsetminus \{v\}$. These sets are disjoint (because all the edges have disjoint interiors), and nonempty (because $e \cap U$ is nonempty and open in e and thus must contain some interior points of e). Therefore, if v lies on more than two edges, we have a separation of $U \smallsetminus \{v\}$ into more than two nonempty disjoint open subsets, contradicting the fact that it has only two components. This shows that each vertex lies on at most two edges.

On the other hand, if some vertex v lies on only one edge e, the construction above shows that v has a Euclidean neighborhood U contained entirely in e. This means that v has a neighborhood $Y \subset U$ such that $Y \smallsetminus \{v\}$ is connected—just take Y to be the image of $[0, \varepsilon)$ under some homeomorphism $\varphi\colon [0,1] \to e$ taking 0 to v. But any Euclidean neighborhood minus a point is disconnected by Problem 4-1(a), so there can be no such vertex. □

We turn our attention next to 2-manifolds. The following theorem was proved by Tibor Radó [Rad25] in 1925.

Theorem 5.12 (Triangulation Theorem for Surfaces). *Every 2-manifold admits a triangulation by a 2-dimensional simplicial complex, in which each edge lies on exactly two 2-simplices.*

Sketch of proof. The basic approach is analogous to the proof of triangulability of 1-manifolds: Cover the manifold with countably many regular disks, and inductively show that each successive disk can be triangulated in a way that is compatible with the triangulations that have already been defined, so that the manifold is ultimately written as an increasing union of polyhedra. In the case of surfaces, however, finding a triangulation of each successive disk that is compatible with the previous ones is much more difficult, primarily because the boundary of the new disk might intersect the boundaries of the already-defined simplices infinitely many times. Even if there are only finitely many intersections, showing that the regions defined by the intersecting curves are homeomorphic to closed disks, and therefore triangulable, requires a delicate topological result known as the *Schönflies theorem*, which asserts that any topological embedding of the circle into $\mathbb{R}^2$ extends to an embedding of the closed disk. The details of the proof are long and intricate and would take us too far from our main goals, so we leave it to the reader to look it up. A readable presentation can be found in [Moi77]. □

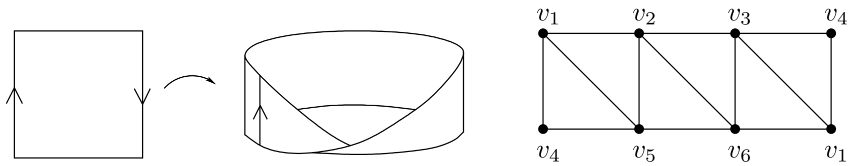


FIGURE 5.7. The Möbius band.

Finally, although we will not use it, we mention the following more recent result, proved by Edwin Moise in 1977 [Moi77].

Theorem 5.13 (Triangulation Theorem for 3-Manifolds). *Every 3-manifold is triangulable.*

Beyond dimension 3, matters are not nearly so nice. It has recently been shown that there are manifolds of dimension 4 that admit no triangulations; and it is still not known whether all manifolds of dimension greater than 4 can be triangulated. See [Ran96] for a history of the subject of triangulations and a summary of the current state of the art.

Orientations

The *Möbius band* is the famous topological space obtained by identifying two edges of the square $I \times I$ according to the relation $(0,t) \sim (1, 1-t)$ (Figure 5.7). It is a manifold with boundary (though not a manifold), and it is triangulable (one triangulation is shown in Figure 5.7). If you have ever made a paper model (it is best to start with a long, narrow rectangle instead of a square), you have undoubtedly noticed that it has the curious property that it is impossible to consistently pick out which is the "front" side and which is the "back"—you cannot continuously color one side gray and the other side white.

By using simplicial theory, we can make this notion precise and extend it to complexes of other dimensions as well. Instead of choosing which side of each triangle to call the front, we will, in effect, choose which direction of travel around the vertices to consider "counterclockwise."

Let σ be an abstract k-simplex. Given any two orderings $(v_{i_0}, \dots, v_{i_k})$ and $(v_{j_0}, \dots, v_{j_k})$ of the vertices of σ, there is a permutation s of the set $\{0, \dots, k\}$ such that $s(i_p) = j_p$ for $p = 0, \dots, k$. Define an equivalence relation on the set of all orderings by saying that two orderings are equivalent if they differ by an even permutation (see the Appendix). A choice of an equivalence class of vertex orderings is called an *orientation* of σ. For example, an orientation of a 1-simplex is just a choice of initial and terminal vertices, which can be indicated schematically by drawing an arrow along

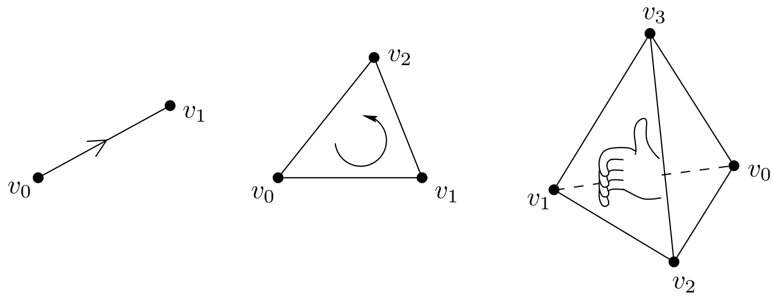


FIGURE 5.8. Orientations of simplices.

the simplex (see Figure 5.8). An orientation of a 2-simplex is a choice of a preferred direction of rotation, which can be indicated by a circular arrow; and an orientation of a 3-simplex is a choice of "handedness": The preferred hand is the one whose fingers curl around the first three vertices in order while the thumb points toward the fourth. Since there is only one way to order a single vertex, by convention an orientation for a 0-simplex is just a choice of a plus or minus sign.

An *oriented simplex* is a simplex together with a choice of orientation. We will write $[v_0, \dots, v_k]$ for the k-simplex $\langle v_0, \dots, v_k \rangle$ oriented by the vertex ordering $(v_0, \dots, v_k)$, and we will let $-[v_0, \dots, v_k]$ denote the same simplex with the opposite orientation. Thus, for example, for 1-simplices and 2-simplices we have

$$\begin{aligned}
[v, w] &= -[w, v], \\
[v, w, x] &= [w, x, v] = [x, v, w] = -[v, x, w] = -[x, w, v] = -[w, v, x].
\end{aligned}$$

Any n-simplex in $\mathbb{R}^n$ automatically gets an orientation, which we call the *natural orientation*, by declaring $[v_0, \dots, v_n]$ to be oriented if and only if $\det(v_1 - v_0, v_2 - v_0, \dots, v_n - v_0) > 0$. To see that this is well-defined, first note that the n vectors $\{v_1 - v_0, \dots, v_n - v_0\}$ are independent precisely when the vertices $\{v_0, \dots, v_n\}$ are in general position. Interchanging two vertices other than v_0 has the same effect as interchanging two rows of the determinant, which changes its sign. If v_0 is interchanged with another vertex v_i, the determinant becomes $\det(v_1 - v_i, \dots, v_0 - v_i, \dots, v_n - v_i)$; multiplying the ith row by -1 (which changes the sign of the determinant) and then adding the ith row to each other row (which leaves the determinant unchanged) transforms the new determinant back to the original one. Thus a transposition of two vertices always changes the sign of the determinant, so an arbitrary permutation of the vertices changes the sign of the determinant if and only if it is even, which shows that this rule gives a well-defined orientation. Geometrically, for a 1-simplex in $\mathbb{R}$ the natural orientation is from the smaller to the larger vertex; for a 2-simplex in $\mathbb{R}^2$ it

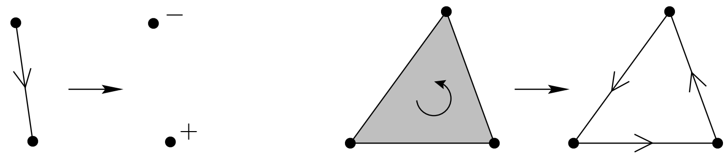


FIGURE 5.9. Induced orientations of boundary faces.

is the counterclockwise direction of rotation; and for a 3-simplex in $\mathbb{R}^3$ it is the right-handed orientation. (These statements can be taken as mathematical definitions of the terms "counterclockwise" and "right-handed.")

If $\sigma = [v_0, \ldots, v_k]$ is an oriented k-simplex, the orientation of σ determines an orientation on each of its boundary faces (i.e., faces of dimension $k-1$), called the *induced orientation*, by the following rule: The induced orientation on the face $\tau_i = \langle v_0, \ldots, \widehat{v}_i, \ldots, v_k \rangle$ (where the hat indicates that v_i is omitted) is defined to be $(-1)^i[v_0, \ldots, \widehat{v}_i, \ldots, v_k]$. To check that this is well-defined, we need to show that the induced orientation of τ_i is unchanged if the vertices of σ are subjected to an even permutation. Because every permutation can be written as a composition of transpositions of adjacent vertices (see Exercise A.19 in the Appendix), it suffices to show that the induced orientation is reversed if two adjacent vertices of σ are transposed. This is clear if neither of the vertices is v_i. If v_i is transposed with $v_{i\pm1}$, the induced orientation becomes $(-1)^{i\pm1}[v_0, \ldots, v_{i-1}, v_{i+1}, \ldots, v_k]$, which is the opposite of what it was originally.

For an oriented 1-simplex $[v_0, v_1]$, the induced orientation gives a minus sign to the initial vertex v_0 and a plus sign to the terminal vertex v_1 (Figure 5.9). For an oriented 2-simplex $[v_0, v_1, v_2]$, the induced orientations on the edges are $[v_1, v_2]$, $-[v_0, v_2] = [v_2, v_0]$, and $[v_0, v_1]$. Thus the arrow on each edge points in the preferred direction of rotation.

Now suppose K is an n-dimensional simplicial complex in which every $(n-1)$-simplex is a face of no more than two n-simplices. (It can be shown, though we will not do so, that any triangulated manifold has this form.) If σ and σ' are two n-simplices that share a boundary face τ, we say that orientations of σ and σ' are *consistent* if they induce opposite orientations on τ. An orientation of K is a choice of orientation of each n-simplex in such a way that any two simplices that intersect in an $(n-1)$-face are consistently oriented. Figure 5.10 gives schematic indications of orientations of 1-dimensional and 2-dimensional complexes. If a complex K admits an orientation, it is said to be *orientable.*

Example 5.14. The triangulation of the Möbius band shown in Figure 5.7 is not orientable. To see why, suppose there exists an orientation. Reversing the orientations of all the 2-simplices if necessary, we may assume that the leftmost triangle is oriented as $[v_1, v_4, v_5]$ (i.e., in the counterclockwise

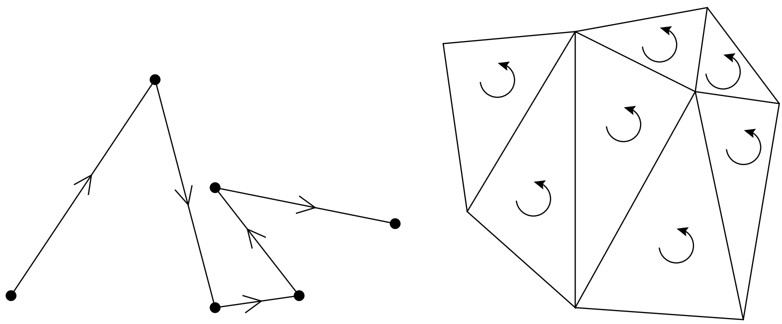


FIGURE 5.10. Orientations of simplicial complexes.

direction). Then the consistency condition implies that the next simplex is oriented as $[v_1, v_5, v_2]$. Similarly, each of the succeeding simplices must be oriented in the counterclockwise direction. But then both the leftmost and the rightmost 2-simplices induce the same orientation $[v_1, v_4]$ on their common edge, which contradicts the consistency condition. Therefore, there exists no orientation.

Example 5.15. Let $\sigma = \langle v_0, \ldots, v_{n+1} \rangle$ be an $(n+1)$-simplex in $\mathbb{R}^{n+1}$, and let K be the set of proper faces of σ, which is a triangulation of $\mathbb{S}^n$. Give σ the natural orientation inherited from $\mathbb{R}^{n+1}$, and give the n-simplices of K the induced orientation. Each $(n-1)$-simplex of K is of the form $\tau = \langle v_0, \ldots, \widehat{v}_i, \ldots, \widehat{v}_j, \ldots, v_{n+1} \rangle$, and belongs to two n-simplices: the one opposite v_i and the one opposite v_j. It is easy to check that the orientations induced on τ by these two faces are opposite, because in one case v_i is removed first and then v_j, while in the other case the order is reversed. Thus we have produced an orientation of K.

Proposition 5.16. *Let K be any n-dimensional Euclidean complex in $\mathbb{R}^n$. The natural orientation of each n-simplex determines an orientation of K.*

Proof. First we show that no more than two n-simplices in K can have an $(n-1)$-face in common. Let τ be an $(n-1)$-simplex in K. The vertices of τ determine a unique affine hyperplane (i.e., $(n-1)$-dimensional affine subspace) in $\mathbb{R}^n$, whose complement has exactly two components, which we call the *sides* of τ. Because the vertices of σ are in general position, the additional vertex of σ that is not in τ must lie on one side of τ or the other, and therefore all of $\sigma \smallsetminus \tau$ must lie on the same side (since it is connected). For any $x \in \operatorname{Int} \tau$ and any sufficiently small $\varepsilon > 0$, $B_\varepsilon(x) \smallsetminus \tau$ has two components, one lying on each side of τ. Any n-simplex that contains τ must contain exactly one of these components for ε small enough. Thus if more than two n-simplices contain τ, two of them must contain the same component of $B_\varepsilon(x) \smallsetminus \tau$ and therefore have interior points in common, which contradicts the definition of a Euclidean complex.

Now we must show that the natural orientations of any two simplices $\sigma, \sigma' \in K$ that share an $(n-1)$-face τ are consistent. Write $\tau = \langle v_0, \dots, v_{n-1}\rangle$, $\sigma = \langle v_0, \dots, v_{n-1}, v_n\rangle$, and $\sigma' = \langle v_0, \dots, v_{n-1}, v_n'\rangle$.

The function $f(v) = \det(v_1 - v_0, \dots, v - v_0)$ is an affine function of v that is zero precisely when v lies in the affine hyperplane determined by τ, so it must be positive on one side of τ and negative on the other side. Since the argument above implies that v_n and v_n' lie on opposite sides of τ, it follows that $\det(v_1 - v_0, \dots, v_n' - v_0)$ has the opposite sign from $\det(v_1 - v_0, \dots, v_n - v_0)$. If the vertices have been ordered so that the natural orientation of σ is $[v_0, \dots, v_n]$, then the natural orientation of σ' is $-[v_0, \dots, v_n']$, and it is immediate that they induce opposite orientations on τ. □

Combinatorial Invariants

Simplicial complexes were invented in the hope that they would enable topological questions about manifolds to be reduced to combinatorial questions about simplicial complexes. To make sense of this, we need a notion of equivalence of complexes that is weaker than simplicial isomorphism but strong enough to imply that they have homeomorphic geometric realizations, and that can be detected purely from the combinatorial structure of the abstract complexes.

The most natural way to modify a simplicial complex to obtain another one with a homeomorphic geometric realization is to "subdivide" the simplices of the original complex into smaller ones. We can then consider two complexes to be equivalent if they both have a common subdivision. In this section we make this notion precise, and study one important property of complexes that is preserved by this kind of equivalence. For our purposes, it is sufficient and simpler to restrict our attention to finite complexes, although many of the definitions can be extended to the general case.

Let K be a finite Euclidean simplicial complex. A *subdivision* of K is a simplicial complex K' with the following properties:

- $|K'| = |K|$.
- Each simplex of K' is contained in a simplex of K.
- Each simplex of K is a finite union of simplices of K'.

Some examples of subdivisions are shown in Figure 5.11.

Example 5.17 (Barycentric Subdivision). A particularly useful kind of subdivision is obtained in the following way. Let $\sigma = \langle v_0, \dots, v_k\rangle$ be a

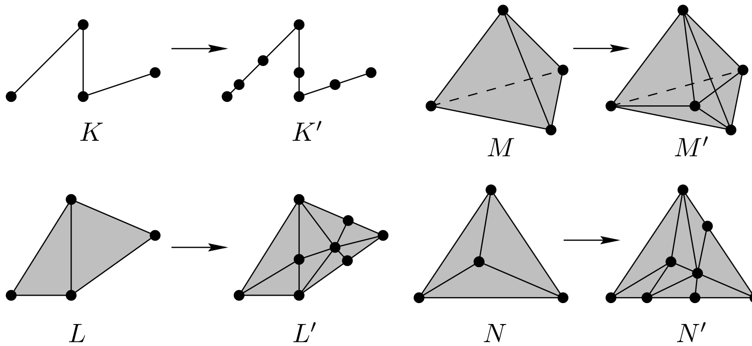


FIGURE 5.11. Subdivisions.

Euclidean k-simplex in $\mathbb{R}^n$. If v is a point not in the k-dimensional affine subspace determined by σ, we define

$$v * \sigma = \langle v, v_0, \dots, v_k \rangle.$$

This is a $(k+1)$-simplex, called the *cone* on σ from v.

Now let K be a finite Euclidean complex. For each k-simplex $\sigma = \langle v_0, \dots, v_k \rangle \in K$, define the *barycenter* of σ to be the point

$$b_\sigma = \sum_{i=0}^{k} \frac{1}{k+1} v_i \in \operatorname{Int} \sigma.$$

It is the "center of gravity" of the vertices of σ. (The name comes from Greek *barys*, meaning "heavy.") For example, the barycenter of a 1-simplex is just its midpoint; the barycenter of a vertex v is v itself.

We will define a complex SK, called the *barycentric subdivision* of K, whose vertices are the barycenters of all the simplices in K. It is easiest to define by induction on the dimension of K. If $\dim K = 0$, then we set $SK = K$ (you cannot subdivide a point!). Assuming that we have defined SK for all finite complexes of dimension less than n, we define SK for a complex of dimension n as the union of $S(K^{(n-1)})$ with the set of all simplices of the form $b_\sigma * \tau$ where σ is an n-simplex of K and τ is any simplex of $S(K^{(n-1)})$ contained in a face of σ. It is straightforward to check that SK is indeed a subdivision of K (see Problem 5-7). Examples of barycentric subdivisions are pictured in Figure 5.12.

The key fact about barycentric subdivision is that it reduces the sizes of all the simplices by a uniform ratio, as the following lemma shows.

Lemma 5.18. *If σ is a Euclidean k-simplex in $\mathbb{R}^m$, the diameter of each simplex in the barycentric subdivision of σ is at most $k/(k+1)$ times that of σ.*

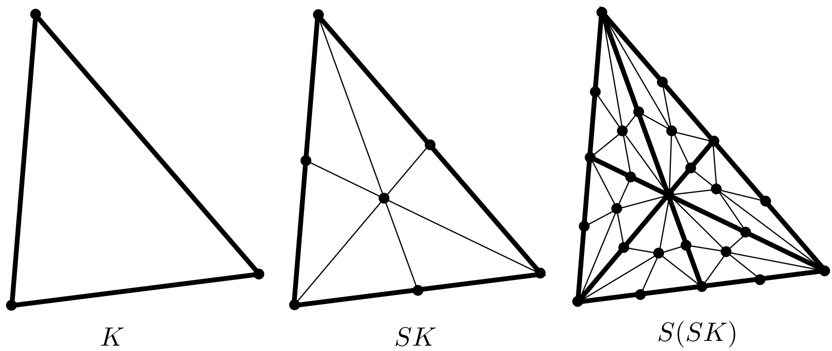


FIGURE 5.12. Barycentric subdivisions.

Proof. Note first that for any point $x \in \sigma$, the maximum of the function $|x - y|$ for $y \in \sigma$ is achieved when y is a vertex. To see why, let R be the maximum distance from x to any vertex of σ; since σ is the convex hull of its vertices and the closed ball $\overline{B}_R(x)$ is a convex set containing the vertices, $\sigma \subset \overline{B}_R(x)$, which proves the claim. It follows immediately that the diameter of σ is the maximum of the distances between its vertices.

The following computation shows that the distance from the barycenter b_τ of a q-simplex $\tau = \langle v_0, \ldots, v_q \rangle$ to any of its vertices v_j is at most $q/(q+1)$ times the diameter of τ:

$$\begin{aligned}
|b_\tau - v_j| &= \left| \sum_{i=0}^{q} \frac{1}{q+1} v_i - v_j \right| \\
&= \left| \sum_{i=0}^{q} \frac{1}{q+1} v_i - \sum_{i=0}^{q} \frac{1}{q+1} v_j \right| \\
&\le \sum_{i=0}^{q} \frac{1}{q+1} |v_i - v_j| \\
&\le \frac{q}{q+1} \operatorname{diam} \tau.
\end{aligned}$$

Now if σ' is any face of the barycentric subdivision of σ and w_1, w_2 are any two vertices of σ', by Problem 5-7 each w_j is the barycenter of a k_j-dimensional face τ_j of σ, and we may assume that τ_1 is a face of τ_2. By the computation above, the distance from w_2 to any point of τ_2 is at most $k_2/(k_2 + 1)$ times the diameter of τ_2. Since w_1, in particular, is a point in τ_2, we have

$$|w_1 - w_2| \le \frac{k_2}{k_2 + 1} \operatorname{diam} \tau_2 \le \frac{k}{k+1} \operatorname{diam} \sigma.$$

It follows from the remark at the beginning of the proof that $\operatorname{diam}\sigma'$ satisfies the same inequality. □

To express subdivisions in terms of the combinatorics of abstract complexes, we will show how to decompose an arbitrary subdivision into a sequence of subdivisions that are combinatorially simpler. Suppose K' is a subdivision of K. We say that it is an *elementary subdivision* if K' contains precisely one more vertex than K. (For example, the 3-dimensional complex M' in Figure 5.11 is an elementary subdivision of M, obtained by adding one vertex in the bottom face of M.)

Suppose we start with a finite Euclidean complex K and choose a k-simplex $\sigma = \langle v_0, \dots, v_k \rangle \in K$, choose a point $v \in \operatorname{Int}\sigma$, and replace each simplex $\langle v_0, \dots, v_k, w_1, \dots, w_m \rangle$ that has σ as a face (including σ itself) by the set of all simplices of the form $\langle v, v_{i_1}, \dots, v_{i_j}, w_1, \dots, w_m \rangle$ as $\{v_{i_1}, \dots, v_{i_j}\}$ ranges over proper subsets of $\{v_0, \dots, v_k\}$. Then it is easy to check that K' is an elementary subdivision of K, and that every elementary subdivision is of this form. Moreover, if K'' is any subdivision of K, there is a finite sequence $K = K_0, K_1, \dots, K_m = K''$ of complexes such that K_{i+1} is an elementary subdivision of K_i. One advantage of working with elementary subdivisions is that the effect of an elementary subdivision on the vertex scheme of K is clearly determined solely by the choice of σ, and so elementary subdivisions can be defined by the recipe above for arbitrary abstract complexes as well.

Two finite simplicial complexes are said to be *combinatorially equivalent* if they become isomorphic after finitely many elementary subdivisions. It was conjectured by Ernst Steinitz and Heinrich Tietze in 1908 that if two finite simplicial complexes have homeomorphic polyhedra, they are combinatorially equivalent; this conjecture became known as the *Hauptvermutung* (main conjecture) of combinatorial topology. It is now known to be true for all finite complexes of dimension 2 and for triangulated compact manifolds of dimension 3, but false in all higher dimensions. (See [Ran96] for a nice discussion of the history of this problem.) Thus the hope of reducing topological questions about manifolds to combinatorial ones about simplicial complexes has not been realized. Nonetheless, simplicial theory has provided us with a number of extremely useful combinatorial invariants that turn out to have important topological ramifications. We conclude this chapter with an introduction to one of them, called the Euler characteristic.

The Euler Characteristic

One of the oldest results in global surface theory is *Euler's formula*: If $P \subset \mathbb{R}^3$ is a compact polyhedral surface that is the boundary of a convex open set, and P has F faces, E edges, and V vertices, then $V - E + F = 2$. This quantity has a natural generalization to arbitrary finite simplicial complexes: If $\mathcal{K}$ is a finite simplicial complex of dimension n, we define the

Euler characteristic of $\mathcal{K}$, denoted by $\chi(\mathcal{K})$, by

$$\chi(\mathcal{K}) = \sum_{k=0}^{n} (-1)^k n_k,$$

where n_k is the number of k-dimensional simplices in $\mathcal{K}$. Although we are not yet in a position to prove Euler's formula in full generality, we can at least show that the Euler characteristic of a simplicial complex is a combinatorial invariant.

Theorem 5.19. *If $\mathcal{K}$ and $\mathcal{L}$ are combinatorially equivalent finite simplicial complexes, then $\chi(\mathcal{K}) = \chi(\mathcal{L})$.*

Proof. It clearly suffices to prove that the Euler characteristic is unchanged by an elementary subdivision. Let $\mathcal{K}'$ be an elementary subdivision of $\mathcal{K}$ obtained by adding a vertex v in the k-simplex $\sigma = \langle v_0, \dots, v_k \rangle$, and let $\Delta\chi = \chi(\mathcal{K}') - \chi(\mathcal{K})$. We must show that $\Delta\chi = 0$.

For each simplex $\tau = \langle v_0, \dots, v_k, w_1, \dots, w_m \rangle$ of $\mathcal{K}$ that has σ as a face, $\mathcal{K}'$ has one less $(k+m)$-simplex. In its place, for each j-element proper subset $\{v_{i_1}, \dots, v_{i_j}\} \subset \{v_0, \dots, v_k\}$, $\mathcal{K}'$ has a new $(j+m)$-simplex $\langle v, v_{i_1}, \dots, v_{i_j}, w_1, \dots, w_m \rangle$. There are $\binom{k+1}{j} = \frac{(k+1)!}{j!(k+1-j)!}$ such subsets, so each such τ makes a contribution to $\Delta\chi$ of

$$-(-1)^{k+m} + \sum_{j=0}^{k} \binom{k+1}{j} (-1)^{j+m} = \sum_{j=0}^{k+1} \binom{k+1}{j} (-1)^{j+m}.$$

By the binomial theorem, this last sum is the expansion of the polynomial $(-1)^m (x+1)^{k+1}$ evaluated at $x = -1$, and therefore is equal to zero. $\square$

Note that we are not claiming yet that the Euler characteristic is a topological invariant, because two triangulations of the same compact space are not necessarily combinatorially equivalent. In fact, it *is* a topological invariant. For compact surfaces, this will follow from the classification of surfaces, which we will complete in Chapter 10. For more general simplicial complexes, the proof will require techniques of homology theory, which we will develop in Chapter 13.

Problems

5-1. Suppose X is a topological space, and $G_1, \dots, G_k$ are finitely many closed subspaces of X whose union is X. Show that the topology of X is coherent with these subspaces. Explain what this has to do with the gluing lemma (Lemma 3.8).

5-2. Let v be a vertex of the simplicial complex $\mathcal{K}$, and let $\operatorname{St} v$ (the *open star* of v) be the union of the open simplices $\operatorname{Int}\sigma$ as σ ranges over all simplices that have v as a vertex. Show that $\operatorname{St} v$ is a neighborhood of v in $|\mathcal{K}|$, and the collection of open stars of all the vertices is an open cover of $|\mathcal{K}|$.

5-3. Show that every polyhedron is Hausdorff and locally path connected.

5-4. Let $\mathcal{K}$ be an abstract simplicial complex. Show that $|\mathcal{K}|$ is compact if and only if $\mathcal{K}$ is finite, and locally compact if and only if $\mathcal{K}$ is locally finite.

5-5. Show that an abstract simplicial complex is the vertex scheme of a Euclidean complex if and only if it is finite-dimensional, locally finite, and countable. [Hint: If the complex has dimension n, let the vertices be the points $v_k = (k, k^2, k^3, \dots, k^{2n+1}) \in \mathbb{R}^{2n+1}$. Use the fundamental theorem of algebra to show that no $2n+2$ vertices lie in a proper affine subspace, so any $2n+2$ or fewer vertices are in general position. If two simplices σ, τ with vertices in this set intersect, let σ_0, τ_0 be the smallest face of each containing an intersection point, and consider the set consisting of all the vertices of σ_0 and τ_0. (This proof is from [Sti93].)]

5-6. Define an abstract simplicial complex $\mathcal{K}$ to be the following collection of abstract 2-simplices together with all of their faces:

$$\begin{aligned}\{\{a,b,e\},\{b,e,f\},\{b,c,f\},\{c,f,g\},\{a,c,g\},\{a,e,g\},\\ \{e,f,h\},\{f,h,j\},\{f,g,j\},\{g,j,k\},\{e,g,k\},\{e,h,k\},\\ \{a,h,j\},\{a,b,j\},\{b,j,k\},\{b,c,k\},\{c,h,k\},\{a,c,h\}\}.\end{aligned}$$

Show that the geometric realization of $\mathcal{K}$ is homeomorphic to the torus. [Hint: Look at Figure 5.13.]

5-7. If K is a finite Euclidean simplicial complex, show that its barycentric subdivision SK is in fact a subdivision of K, and that the simplices of SK are those of the form $\langle b_{\sigma_0}, \dots, b_{\sigma_k}\rangle$ in which each σ_j is a simplex in K and σ_j is a face of σ_{j+1} for $j = 0, \dots, k-1$.

5-8. Let $\mathcal{K}$ be a finite complex. Give an explicit algorithm for obtaining the barycentric subdivision of $\mathcal{K}$ as a sequence of elementary subdivisions.

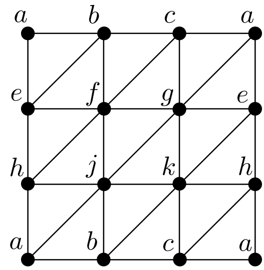


FIGURE 5.13. Triangulation of the torus.

5-9. Let $\mathcal{K}$ be the 1-dimensional abstract complex whose vertices are the nonnegative integers and whose 1-simplices are $\{\langle 0, n\rangle : n \in \mathbb{N}\}$, and let S be the subspace of $\mathbb{R}^2$ obtained by taking the union of all the line segments in Example 5.3. Define a map $F\colon |\mathcal{K}| \to S$ by sending 0 to the origin, sending each $n > 0$ to the point $(1, 1/n)$, and sending each 1-simplex $\langle 0, n\rangle$ linearly onto the corresponding line segment. Show that F is continuous and bijective but not a homeomorphism.

5-10. Suppose $\mathcal{K}$ is any simplicial complex whose geometric realization is a 1-manifold. Show that $\mathcal{K}$ is 1-dimensional.

5-11. Show that every triangulated 1-manifold is orientable.

5-12. Show that orientability is a combinatorial invariant of finite simplicial complexes. [Hint: It suffices to prove that if K is a finite Euclidean complex and K' is a subdivision of K, then an orientation of either K or K' determines an orientation of the other. Show that if σ is an oriented Euclidean n-simplex and σ' is an n-simplex in some subdivision of σ, there is a unique orientation of σ' such that any affine embedding $\sigma \to \mathbb{R}^n$ that determines the given orientation of σ also determines the given orientation of σ'.]

6

Curves and Surfaces

In this chapter we undertake a detailed study of curves (1-manifolds) and surfaces (2-manifolds). These are the manifolds that are most familiar from our everyday experience, and about which the most is known mathematically. They are thus excellent prototypes for the study of manifolds in higher dimensions.

We begin by proving the classification theorem for 1-manifolds, which says that every connected 1-manifold is homeomorphic to $\mathbb{S}^1$ or $\mathbb{R}$. Using the triangulation theorem for 1-manifolds proved in Chapter 5, this is a simple exercise in the combinatorics of graphs.

We then proceed to a general discussion of 2-manifolds and a detailed examination of the basic examples of compact surfaces: the sphere, the torus, and the projective plane. Next we show how to form other compact surfaces by the technique of connected sums, a way of patching together simpler surfaces to form more complicated ones. To unify these results, we introduce the notion of polygonal presentations of surfaces, which generalize simplicial complexes by representing surfaces as a collection of polygons (not necessarily triangles) with edges identified in pairs.

The central part of the chapter presents the main part of the classification theorem for compact surfaces, which says that every compact, connected surface is homeomorphic to a sphere, a connected sum of tori, or a connected sum of projective planes. Again, the triangulation theorem reduces the problem to one of showing that every polygonal presentation can be reduced to a standard presentation of one of the model surfaces.

In the last section we revisit orientations and the Euler characteristic, introduced in the last chapter for simplicial complexes, and reinterpret them in the context of polygonal surface presentations.

Classification of Curves

Our first goal in this chapter is to prove that up to homeomorphism, the only connected 1-manifolds are the line and the circle. (Of course, this classifies the disconnected ones too, because it implies that each component of a disconnected 1-manifold is a line or a circle, so every 1-manifold is homeomorphic to a countable disjoint union of lines and/or circles.) You can think of this classification theorem as a warm-up for the more complicated classification of surfaces to follow; but it is also quite important in its own right.

Theorem 6.1 (Classification of 1-Manifolds). *A connected 1-manifold is homeomorphic to $\mathbb{S}^1$ if it is compact and to $\mathbb{R}$ if it is not.*

Proof. By the triangulation theorem (Theorem 5.10) and Proposition 5.11, M is homeomorphic to a graph in which every vertex lies on exactly two edges. Let $\mathcal{K}$ denote the abstract simplicial complex associated with this graph. We will show that $\mathcal{K}$ is isomorphic either to one of the complexes $\mathcal{K}_m$ of Example 5.7(d) (the vertex scheme of a regular m-gon, whose polyhedron is homeomorphic to $\mathbb{S}^1$) or to the complex $\mathcal{K}_\infty$ of Example 5.7(c) (whose polyhedron is homeomorphic to $\mathbb{R}$). Since isomorphic complexes have homeomorphic polyhedra, this suffices to prove the theorem.

The first step is to show that every vertex in M is contained in a reduced edge path $\{v_n : n \in \mathbb{Z}\}$ that extends indefinitely in both directions. Start with any vertex v_0. By assumption v_0 lies on two edges, so we can label the other vertices of those edges arbitrarily as v_1 and v_{-1}. Now by induction define v_{n+1} for each $n \geq 1$ to be the unique vertex other than v_{n-1} such that v_n and v_{n+1} span an edge. Similarly, we define v_{-n} by induction on n.

Let U be the union of all the edges $\langle v_n, v_{n+1} \rangle \subset M$ for v_n in the edge path. The set U is closed in M (because its intersection with each simplex is closed and M has the coherent topology), and open (because the edges minus their vertices are open, and each vertex has a neighborhood intersecting only two edges, both of which must be in U). Thus $U = M$.

Now we distinguish two cases.

Case I: $v_n \neq v_k$ for any $n \neq k$. In this case, the correspondence sending $n \mapsto v_n$ is easily seen to give an isomorphism between $\mathcal{K}_\infty$ and $\mathcal{K}$, so M is homeomorphic to $\mathbb{R}$.

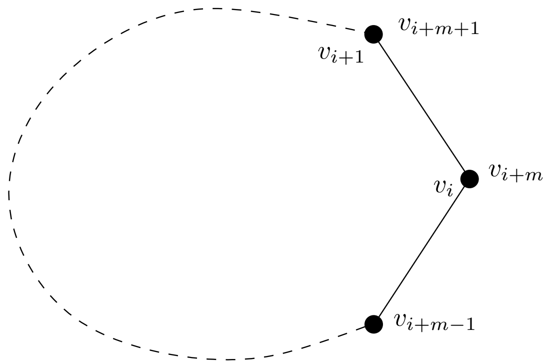


FIGURE 6.1. Periodic edge path.

CASE II: $v_n = v_{n+m}$ for some $n \in \mathbb{Z}$ and some $m > 0$. We may assume that m and n have been chosen so that m is the least positive integer with this property. (Note that $m \geq 3$ because the edge path is reduced.) We will show by induction that the edge path is *periodic*, in the sense that $v_i = v_{i+m}$ for every i. By hypothesis this is true for $i = n$. If it is true for some $i \geq n$, we argue as follows. The two vertices v_{i+m-1} and v_{i+m+1} are the only vertices that are connected to v_{i+m} by edges (Figure 6.1). Since $v_i = v_{i+m}$, the vertex v_{i+1} also is connected by an edge to v_{i+m}, so it must be equal to either v_{i+m-1} or v_{i+m+1}. By minimality of m it cannot be equal to v_{i+m-1}, so $v_{i+1} = v_{i+m+1}$, completing the induction for $i \geq n$. A similar induction takes care of $i \leq n$.

Now let $\mathcal{K}_m$ be the complex of Example 5.7(d), and define a map $f\colon \mathcal{K}_m^{(0)} \to \mathcal{K}^{(0)}$ by $f(n) = v_n$ for $n = 1, \ldots, m$. Again, it is straightforward to check that $\{f(i), f(j)\}$ are the vertices of a simplex of $\mathcal{K}$ if and only if $\{i, j\}$ are the vertices of a simplex of $\mathcal{K}_m$. Thus f extends to a simplicial isomorphism, so M is homeomorphic to $\mathbb{S}^1$. □

Surfaces

The rest of this chapter is devoted to the study of compact surfaces. We have already seen several important examples: the sphere $\mathbb{S}^2$, the torus $\mathbb{T}^2$, and the projective plane $\mathbb{P}^2$ (i.e., the projective space of dimension 2). As we will soon see, these examples are fundamental because every compact surface can be built up from these three.

In order to systematize our knowledge of surfaces, it will be useful to have a uniform way to represent them that is somewhat more general than simplicial complexes. The prototype is the representation of the torus as a

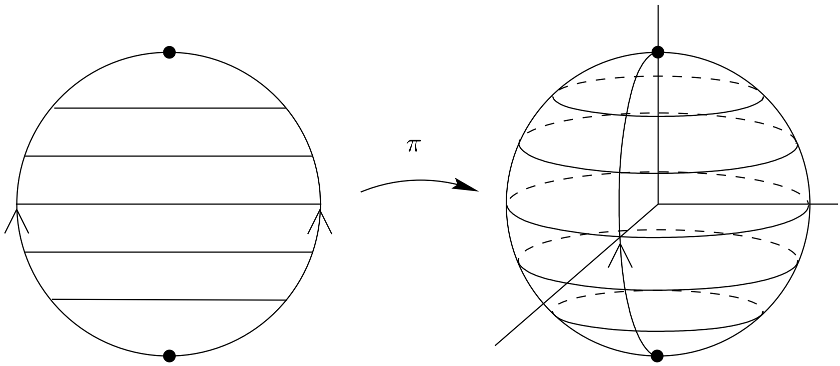


FIGURE 6.2. The sphere as a quotient of the disk.

quotient of the square by identifying the edges in pairs (Example 3.24). It turns out that every compact surface can be represented as a quotient of a polygonal region in the plane by an equivalence relation that identifies its edges in pairs.

Let us begin by seeing how our three basic compact examples can be so represented. We have already shown how to represent the torus as a quotient of a square (Example 3.24), so we focus on the sphere and the projective plane.

Proposition 6.2. *The sphere $\mathbb{S}^2$ is homeomorphic to the following quotient spaces.*

(*a*) *The closed disk $\overline{\mathbb{B}^2} \subset \mathbb{R}^2$ modulo the equivalence relation generated by $(x,y) \sim (-x,y)$ for $x \in \partial\mathbb{B}^2$ (Figure 6.2).*

(*b*) *The square $I \times I$ modulo the equivalence relation generated by $(0,t) \sim (t,0)$ and $(t,1) \sim (1,t)$ for $0 \le t \le 1$ (Figure 6.3).*

Proof. To see that each of these spaces is homeomorphic to the sphere, all we need to do is exhibit a quotient map from the given space to the sphere that makes the same identifications, and then appeal to uniqueness of quotient spaces (Corollary 3.32).

For (a), define a map from the disk to the sphere by wrapping each horizontal line segment around a "latitude circle" (Figure 6.2). Formally, $\pi\colon \overline{\mathbb{B}^2} \to \mathbb{S}^2$ is given by

$$\pi(x,y) = \begin{cases} \left(-\sqrt{1-y^2}\cos\dfrac{\pi x}{\sqrt{1-y^2}}, -\sqrt{1-y^2}\sin\dfrac{\pi x}{\sqrt{1-y^2}}, y \right), & y \neq \pm 1; \\ (0,0,y), & y = \pm 1. \end{cases}$$

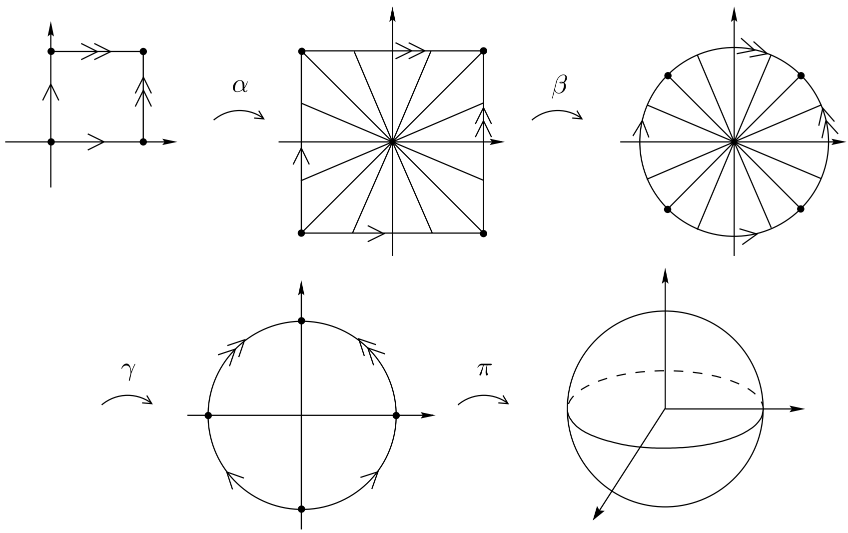


FIGURE 6.3. The sphere as a quotient of the square.

This is a quotient map by the closed map lemma; it is straightforward to check that it makes exactly the same identifications as the given equivalence relation.

To prove (b), we will construct a quotient map from $I \times I$ to the sphere as a composition of several simpler maps (Figure 6.3). First let S denote the square $\{(x,y) : |x|,|y| \le 1\}$, and define a homeomorphism $\alpha\colon I \times I \to S$ by first scaling both coordinates by a factor of 2 and then translating the new center $(1,1)$ back to the origin: $\alpha(x,y) = (2x-1, 2y-1)$. Then let $\beta\colon S \to \overline{\mathbb{B}^2}$ be the homeomorphism whose existence is guaranteed by Proposition 4.26; it sends each radial line segment between the origin and the boundary of S linearly onto the parallel segment between the center of the disk and its boundary. Next let $\gamma\colon \overline{\mathbb{B}^2} \to \overline{\mathbb{B}^2}$ be the counterclockwise rotation through $\pi/4$ radians (obviously a homeomorphism), and consider the composite map $\varphi = \pi \circ \gamma \circ \beta \circ \alpha$:

$$I \times I \xrightarrow{\alpha} S \xrightarrow{\beta} \overline{\mathbb{B}^2} \xrightarrow{\gamma} \overline{\mathbb{B}^2} \xrightarrow{\pi} \mathbb{S}^2,$$

where π is the quotient map of the preceding paragraph. Since this is a composition of quotient maps, it is a quotient map. Threading through the definitions (with help from the pictures!), you will see that it makes the same identifications as the quotient map defined in (b), thus completing the proof. $\square$

Proposition 6.3. *The projective plane $\mathbb{P}^2$ is homeomorphic to each of the following quotient spaces (Figure 6.4).*

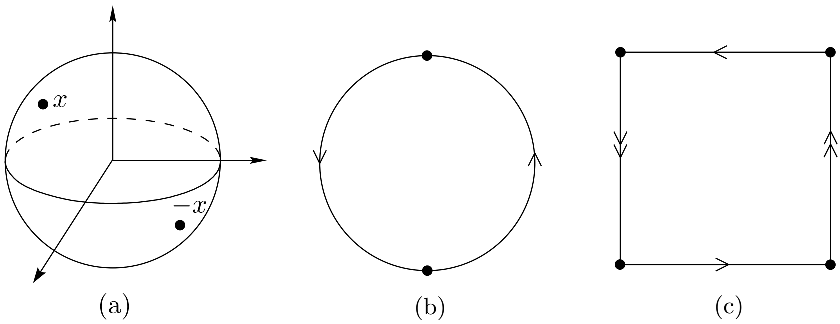


FIGURE 6.4. Representations of $\mathbb{P}^2$ as a quotient space.

(*a*) *The sphere* $\mathbb{S}^2$ *modulo the equivalence relation* $x \sim -x$ *for each* $x \in \mathbb{S}^2$.

(*b*) *The closed disk* $\overline{\mathbb{B}^2}$ *modulo the relation* $(x,y) \sim (-x,-y)$ *for each* $(x,y) \in \partial\mathbb{B}^2$.

(*c*) *The square* $I \times I$ *modulo the relation* $(t,0) \sim (1-t,1)$, $(0,1-t) \sim (1,t)$ *for* $0 \le t \le 1$.

Proof. Let $\mathbb{S}^2/{\sim}$ denote the quotient space of $\mathbb{S}^2$ obtained by identifying each point x with its antipodal point $-x$, and let $p\colon \mathbb{S}^2 \to \mathbb{S}^2/{\sim}$ denote the quotient map. Consider also the composite map

$$\mathbb{S}^2 \xhookrightarrow{\iota} \mathbb{R}^3 \smallsetminus \{0\} \xrightarrow{\pi} \mathbb{P}^2,$$

where ι is inclusion and π is the quotient map defining $\mathbb{P}^2$. Note that $\pi \circ \iota$ is a quotient map by the closed map lemma. It makes exactly the same identifications as p, so by uniqueness of quotient spaces $\mathbb{P}^2$ is homeomorphic to $\mathbb{S}^2/{\sim}$.

If $F\colon \overline{\mathbb{B}^2} \to \mathbb{S}^2$ is the map sending the disk onto the upper hemisphere by $F(x,y) = (x, y, \sqrt{1-x^2-y^2})$, then $p \circ F\colon \overline{\mathbb{B}^2} \to \mathbb{S}^2/{\sim}$ is easily seen to be surjective, and is thus a quotient map by the closed map lemma. It identifies only $(x,y) \in \partial\mathbb{B}^2$ with $(-x,-y) \in \partial\mathbb{B}^2$, so $\mathbb{P}^2$ is homeomorphic to the resulting quotient space.

Part (c) is left as an exercise. □

Exercise 6.1. Prove Proposition 6.3(c).

When doing geometric "cutting and pasting" constructions like the ones in the last two propositions, it is often safe to rely on pictures and a few words to describe the maps and identifications being constructed. So far, we have been careful to give explicit definitions (often with formulas) of all our maps, together with rigorous proofs that they do in fact give the results

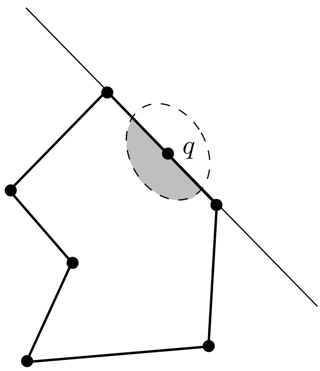


FIGURE 6.5. An edge point.

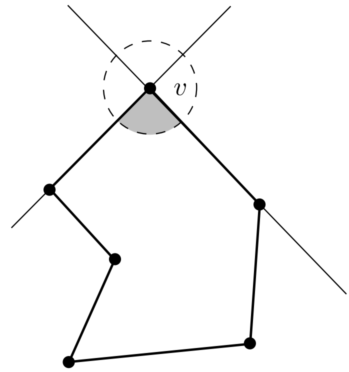


FIGURE 6.6. A vertex.

we claim; but as your sophistication increases and you become adept at carrying out such explicit constructions yourself, you can leave out many of the details. The main thing is that before you skip any such details, you should be absolutely sure that you could quickly write them down and check your claims rigorously; this is the only way to be sure that you are not hiding real difficulties behind "hand-waving." In this book we will begin to leave out some such details in our proofs; for a while, you should fill them in for yourself to be sure that you know how to turn an argument based on pictures into a complete proof.

Now we describe a general method for building surfaces by identifying edges of geometric figures. Let us say that a subset P of the plane is a *polygonal region* if it is a compact subset whose boundary is a finite 1-dimensional Euclidean simplicial complex, satisfying the following conditions:

(i) Each point q of an edge other than a vertex has a neighborhood U in $\mathbb{R}^2$ such that $P \cap U$ is equal to the intersection with U of some closed half-plane $\{(x, y) : ax + by + c \geq 0\}$ (Figure 6.5).

(ii) Each vertex v has a neighborhood V in $\mathbb{R}^2$ such that $P \cap V$ is equal to the intersection of V with two closed half-planes whose boundaries intersect only at v (Figure 6.6).

Any finite collection of disjoint 2-simplices in the plane is easily seen to be a polygonal region, as is a filled-in square, or any compact convex region bounded by finitely many 1-simplices. Below, we will see some more examples of manifolds obtained as quotients of polygonal regions by identifying the edges in pairs. It is a general fact that such a quotient space is always a surface.

Proposition 6.4. *Let P be a polygonal region in the plane with an even number of edges, and suppose we are given an equivalence relation that*

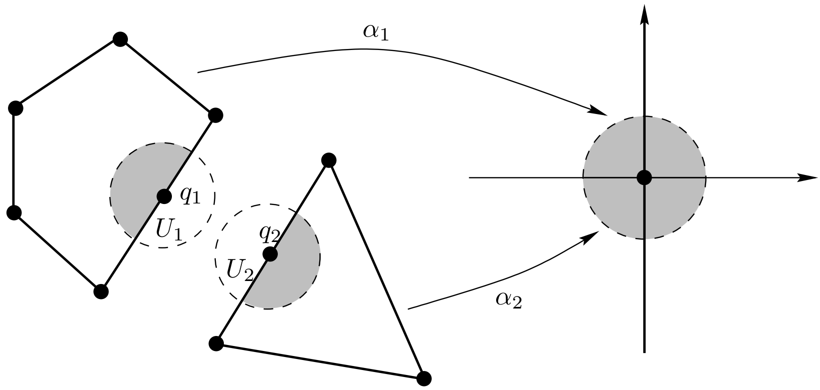


FIGURE 6.7. Euclidean neighborhood of an edge point.

identifies each edge with exactly one other edge by means of a simplicial homeomorphism. The resulting quotient space is a compact 2*-manifold.*

Proof. Let M be the quotient space, and let $\pi\colon P \to M$ denote the quotient map. Clearly, M is compact, because it is the continuous image of the compact space P.

Since the equivalence relation identifies only edges with edges and vertices with vertices, the points of M fall into three disjoint sets: (a) *face points*, whose inverse images in P are in the interior of P; (b) *edge points*, whose inverse images are on edges but not vertices; and (c) *vertex points*, whose inverse images are vertices. To prove that M is locally Euclidean, we consider the three types separately.

Because π is injective on $\operatorname{Int} P$, $\operatorname{Int} P$ is a saturated open subset of P, so the restriction of π to $\operatorname{Int} P$ is a one-to-one quotient map and therefore a homeomorphism. Thus $\pi(\operatorname{Int} P)$ is a Euclidean neighborhood of each face point.

An edge point $\widetilde{q}$ has exactly two inverse images q_1 and q_2, each on a different edge. Using condition (i) in the definition of polygonal region, there exist disjoint neighborhoods U_1 of q_1 and U_2 of q_2 such that $P \cap \overline{U_i}$ is a closed half-disk (Figure 6.7). Let $V_i = P \cap U_i$. It is straightforward to construct affine homeomorphisms α_1 taking V_1 to a half-disk in the upper half-plane and α_2 taking V_2 to a half-disk in the lower half-plane, in such a way that q_1 and q_2 both go to the origin and the boundary identifications are respected. Define $\alpha\colon V_1 \cup V_2 \to \mathbb{R}^2$ by letting $\alpha = \alpha_1$ on V_1 and $\alpha = \alpha_2$ on V_2. Shrinking V_1 and V_2 if necessary, we can ensure that $V_1 \cup V_2$ is a saturated open set in P (this just means that for each point in $V_1 \cap \partial P$, the corresponding boundary point is in V_2, and conversely). Then the restriction of π to $V_1 \cup V_2$ is a quotient map, so α descends to a map $\widetilde{\alpha}\colon \widetilde{V} \to \mathbb{R}^2$, where $\widetilde{V} = \pi(V_1 \cup V_2)$. Its image contains a neighborhood

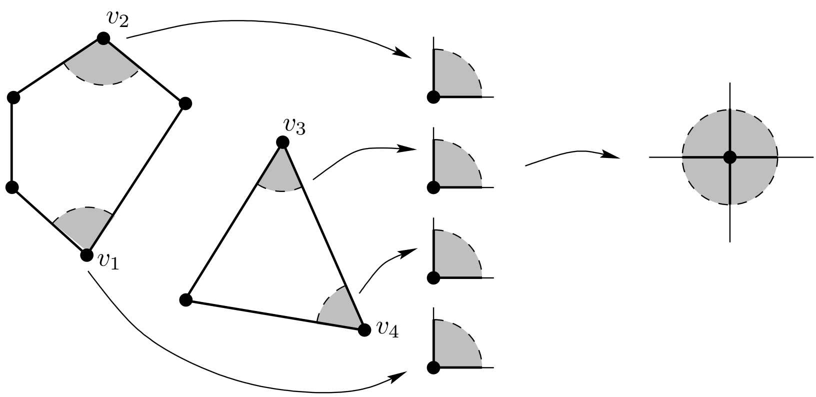


FIGURE 6.8. Euclidean neighborhood of a vertex point.

of the origin by construction, and its domain $\widetilde{V}$ is the image under π of a saturated open set and therefore open. This shows that $\widetilde{q}$ has a locally Euclidean neighborhood.

Similarly, a vertex point $\widetilde{v}$ has as its inverse image a finite set of vertices $\{v_1, \ldots, v_k\} \subset P$. For each i, choose a homeomorphism from a neighborhood of v_i in P to an open subset in a closed "wedge" of angle $2\pi/k$ in the plane, which is a set described in polar coordinates by $\{(r, \theta) : \theta_0 \leq \theta \leq \theta_0 + 2\pi/k\}$. (If we place v_i at the origin, such a homeomorphism is given in polar coordinates by a *fan transformation* of the form $(r, \theta) \mapsto (r, \theta_0 + c\theta)$ for suitable constants θ_0, c.)

Because each edge is paired with exactly one other, the k wedges can be mapped onto a set containing a neighborhood of the origin by rotating and piecing them together (Figure 6.8). However, this may not respect the edge identifications. To correct this, we can subject each wedge to a preliminary transformation that rescales its edges independently. First, by a rotation followed by a fan transformation, take the wedge to the first quadrant so that one edge lies along the positive x-axis and the other along the positive y-axis. Then rescale the two axes by a linear transformation $(x, y) \mapsto (ax, by)$. Finally, use another fan transformation to insert the wedge into its place. (The case $k = 1$ deserves special comment. This case can occur only if the two edges adjacent to the single vertex v_1 are identified with each other; then you can check that our construction maps a neighborhood of v_1 onto a neighborhood of the origin, with both edges going to the same ray.) In each case, we end up with a map defined on a saturated open set in P, which descends to a homeomorphism from a neighborhood of $\widetilde{v}$ to a neighborhood of the origin.

The quotient is second countable by Lemma 3.21. To prove that it is Hausdorff is the same as showing that the fibers of π can be separated by

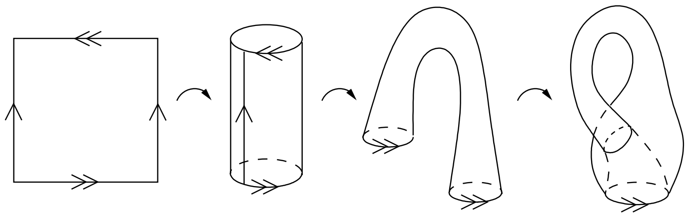


FIGURE 6.9. The Klein bottle.

saturated open sets. It is straightforward to check on a case-by-case basis that the inverse images of sufficiently small Euclidean balls will do. □

Here is another example of a manifold formed as a quotient of a polygonal region.

Example 6.5. The *Klein bottle* is the 2-manifold K obtained by identifying the edges of the square $I \times I$ according to $(0,t) \sim (1,t)$ and $(t,0) \sim (1-t,1)$ for $0 \le t \le 1$. To visualize K, think of gluing the left and right edges together to form a cylinder, and then passing the upper end of the cylinder through the cylinder wall near the lower end, in order to glue the upper circle to the lower one "from the inside" (Figure 6.9). Of course, this cannot be done with a physical model; in fact, it can be shown that the Klein bottle is not homeomorphic to any subspace of $\mathbb{R}^3$. Nonetheless, the preceding proposition shows that it is a 2-manifold.

Connected Sums

To construct other examples of surfaces, we now introduce an important way of producing manifolds by gluing together simpler ones, called "connected sum." Although we will use this primarily for surfaces, it works for manifolds of any dimension, so we will define it in arbitrary dimensions.

Let M_1 and M_2 be connected n-manifolds. Geometrically, the connected sum is obtained by cutting out a small open ball from each of M_1 and M_2 and gluing the resulting spaces together along their boundary spheres. More precisely, let $B_i \subset M_i$ be regular Euclidean balls. Choose a homeomorphism $\sigma\colon \partial B_1 \to \partial B_2$ (such a homeomorphism exists because both boundaries are homeomorphic to $\mathbb{S}^{n-1}$). Let $M_i' = M_i \smallsetminus B_i$, and define a quotient space of $M_1' \amalg M_2'$ by identifying each $q \in \partial B_1$ with $\sigma(q) \in \partial B_2$ (Figure 6.10). The resulting quotient space is called a *connected sum* of M_1 and M_2 and is denoted by $M_1 \# M_2$.

Proposition 6.6. *If M_1 and M_2 are connected n-manifolds, any connected sum $M_1 \# M_2$ is a connected n-manifold.*

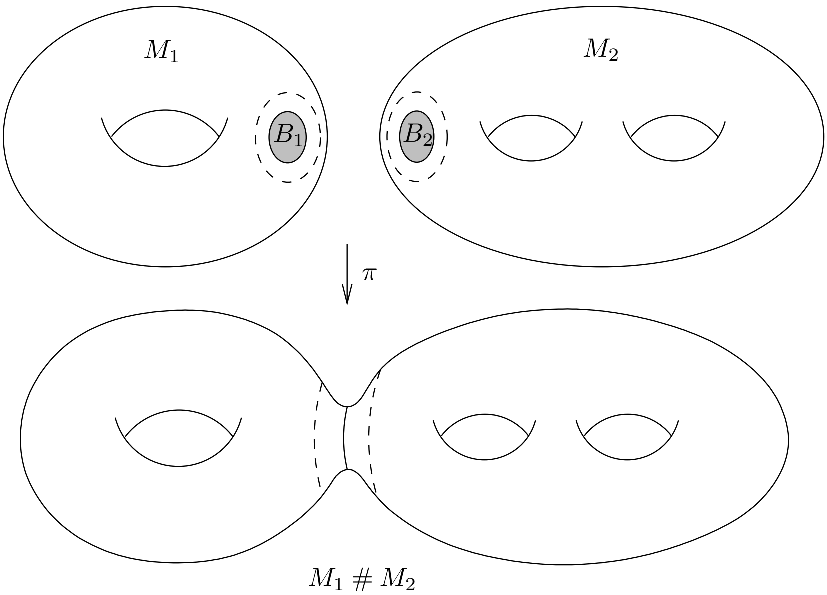


FIGURE 6.10. A connected sum.

Proof. First we show that $M_1 \# M_2$ is locally Euclidean. Let $\pi\colon M_1' \amalg M_2' \to M_1 \# M_2$ denote the quotient map, and let $S = \pi(\partial B_1 \cup \partial B_2)$. Since π is one-to-one and thus a homeomorphism away from S, $M_1 \# M_2 \smallsetminus S$ is a manifold and therefore locally Euclidean. Thus we need only consider points in S.

For any nonempty interval $J \subset [0, \infty)$, let us use the notation A_J to denote the annulus $\{x \in \mathbb{R}^n : |x| \in J\}$. Thus, for example, $A_{[1,2)}$ is equal to $B_2(0) \smallsetminus B_1(0)$. Our regular balls B_i for $i = 1, 2$ come with neighborhoods U_i containing $\overline{B}_i$ and homeomorphisms $\varphi_i\colon U_i \to B_2(0)$ taking $U_i \smallsetminus B_i$ onto $A_{[1,2)}$ (Figure 6.11). (If φ_i takes U_i onto a ball of radius different from 2, we can adjust it by a dilation to make sure that its image is $B_2(0)$.) Notice that φ_i sends ∂B_i to the unit sphere.

The first thing we need to do is to adjust one of these maps to compensate for the fact that $\varphi_2^{-1} \circ \varphi_1$ does not make the same identification between ∂B_1 and ∂B_2 as σ does. To correct this, observe that the composite map $\beta = \varphi_2 \circ \sigma \circ \varphi_1^{-1}$ from $\mathbb{S}^{n-1} \subset \mathbb{R}^n$ to itself is a homeomorphism:

$$\mathbb{S}^{n-1} \xrightarrow{\varphi_1^{-1}} \partial B_1 \xrightarrow{\sigma} \partial B_2 \xrightarrow{\varphi_2} \mathbb{S}^{n-1}.$$

Define a homeomorphism $\widetilde{\beta}$ from $B_2(0)$ to itself by sending the ray through each point $x_0 \in \mathbb{S}^{n-1}$ linearly onto the ray through $\beta(x_0)$; formally, since

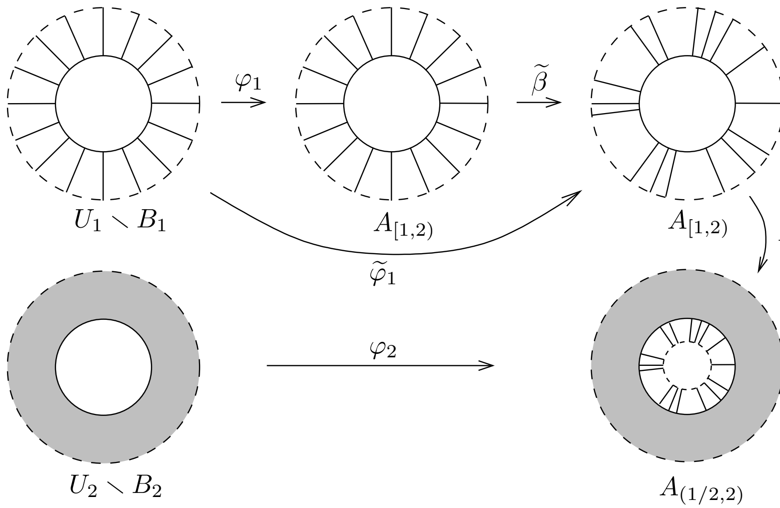


FIGURE 6.11. Euclidean neighborhood of points in S.

$x/|x|$ is the point where the ray through x intersects $\mathbb{S}^{n-1}$,

$$\widetilde{\beta}(x) = |x|\,\beta\left(\frac{x}{|x|}\right).$$

Let $\widetilde{\varphi}_1 = \widetilde{\beta} \circ \varphi_1$. It is immediate from the definitions that

$$\widetilde{\varphi}_1 = \varphi_2 \circ \sigma \quad \text{on } \partial B_1. \tag{6.1}$$

The inversion map $I(x) = x/|x|^2$ maps the annulus $A_{[1,2)}$ homeomorphically onto the annulus $A_{(1/2,1]}$, and restricts to the identity map on $\mathbb{S}^{n-1}$. We define a map Φ from the open set $(U_1 \smallsetminus B_1) \cup (U_2 \smallsetminus B_2) \subset M_1' \amalg M_2'$ to $\mathbb{R}^n$ by

$$\Phi(q) = \begin{cases} I \circ \widetilde{\varphi}_1(q), & q \in U_1 \smallsetminus B_1, \\ \varphi_2(q), & q \in U_2 \smallsetminus B_2. \end{cases}$$

Let us check that Φ respects the identifications made by π. If $q \in \partial B_1$ and $p = \sigma(q) \in \partial B_2$, then

$$\begin{aligned} I \circ \widetilde{\varphi}_1(q) &= \widetilde{\varphi}_1(q) && \text{(because } I \text{ is the identity on } \mathbb{S}^{n-1}) \\ &= \varphi_2 \circ \sigma(q) && \text{(by (6.1))} \\ &= \varphi_2(p). \end{aligned}$$

Thus Φ passes to the quotient and defines a homeomorphism from a neighborhood of S to the open annulus $A_{(1/2,2)}$, showing that $M_1 \# M_2$ is locally

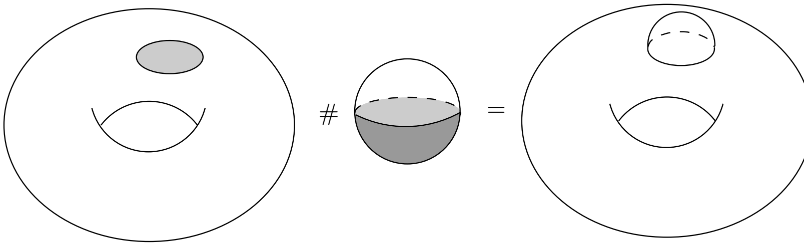


FIGURE 6.12. Connected sum with a sphere.

Euclidean. The proofs that it is second countable and Hausdorff are analogous to those in Proposition 6.4, and are left as an exercise.

To show that $M_1 \# M_2$ is connected, just note $M_1 \# M_2$ is the union of the two sets $\pi(M_1')$ and $\pi(M_2')$, which are both connected and have points of S in common. □

Exercise 6.2. Prove that $M_1 \# M_2$ is Hausdorff and second countable.

Our definition of $M_1 \# M_2$ depends on several choices—the sets B_i and the homeomorphism σ. When M_1 and M_2 are surfaces, it can be shown that different choices yield homeomorphic connected sums. We will not prove this in full generality, but in the case of compact surfaces it will turn out to be a consequence of the classification theorem (see Problem 10-4). Nevertheless, we will sometimes use the imprecise terminology "the connected sum $M_1 \# M_2$" to refer to any connected sum between M_1 and M_2.

Example 6.7. If M is any n-manifold, a connected sum $M \# \mathbb{S}^n$ is homeomorphic to M, at least if we choose the set to cut out of $\mathbb{S}^n$ carefully (Figure 6.12). Let $B_2 \subset \mathbb{S}^n$ be the open lower hemisphere, so $(\mathbb{S}^n)' = \mathbb{S}^n \smallsetminus B_2$ is the closed upper hemisphere, which is homeomorphic to a closed ball. Then $M \# \mathbb{S}^n$ is obtained from M by cutting out the open ball B_1 and gluing back a closed ball along the boundary sphere, so we have not changed anything.

Example 6.8. The n-fold connected sum $\mathbb{T}^2 \# \mathbb{T}^2 \# \cdots \# \mathbb{T}^2$ (Figure 6.13), is called the *n-holed torus*, or the *sphere with n handles*. The latter terminology refers to the fact that this surface is also homeomorphic to $\mathbb{S}^2 \# \mathbb{T}^2 \# \cdots \# \mathbb{T}^2$, and each added torus looks like a "handle" attached to the sphere (Figure 6.14).

Polygonal Presentations of Surfaces

As we mentioned earlier in this chapter, for the classification theorem we need a uniform way to describe surfaces. We will represent all of our surfaces

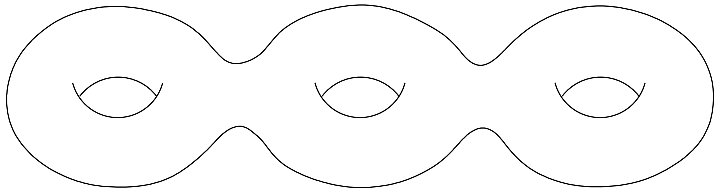


FIGURE 6.13. A connected sum of tori.

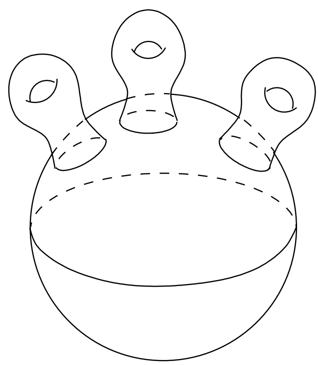


FIGURE 6.14. A sphere with handles.

as quotients of $2n$-sided polygonal regions. Informally, we can describe any edge equivalence relation by labeling the edges with letters $a_1, \dots, a_n$, and giving each edge an arrow pointing toward one of its vertices, in such a way that edges with the same label are to be identified, with the arrows indicating which way the vertices match up. With each such labeling of a polygon we associate a sequence of symbols, obtained by reading off the boundary labels counterclockwise from the top, and for each boundary label a_i, placing a_i in the sequence if the arrow points counterclockwise and a_i^{-1} if it points clockwise. For example, the equivalence relation on $I \times I$ of Example 3.24 that yields the torus might result in the sequence of symbols $aba^{-1}b^{-1}$.

Formally, given a set S, we define a *word* in S to be an ordered k-tuple of symbols, each of the form a or a^{-1} for some $a \in S$. (To be more precise, if you like, you can define a word to be a finite sequence of ordered pairs of the form $(a, 1)$ or $(a, -1)$ for $a \in S$, and then define a and a^{-1} as abbreviations for $(a, 1)$ and $(a, -1)$, respectively.) A *polygonal presentation*, written

$$\mathcal{P} = \langle S \mid W_1, \dots, W_k \rangle,$$

is a finite set S together with finitely many words $W_1, \dots, W_k$ in S of length 3 or more, such that every symbol in S appears in at least one word. As a

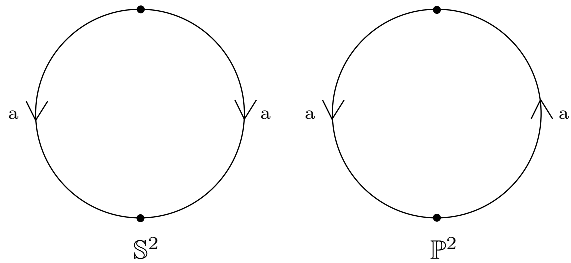


FIGURE 6.15. Presentations of $\mathbb{S}^2$ and $\mathbb{P}^2$.

matter of notation, when the set S is described by listing its elements, we will leave out the braces surrounding the elements of S, and will denote the words W_i by juxtaposition, so for example the presentation with $S = \{a, b\}$ and the single word $W = (a, b, a^{-1}, b^{-1})$ will be written $\langle a, b \mid aba^{-1}b^{-1}\rangle$. We also allow as a special case any presentation in which S has one element and there is a single word of length 2. There are only four such: $\langle a \mid aa\rangle$, $\langle a \mid a^{-1}a^{-1}\rangle$, $\langle a \mid aa^{-1}\rangle$, and $\langle a \mid a^{-1}a\rangle$.

Any polygonal presentation $\mathcal{P}$ determines a topological space $|\mathcal{P}|$, called the *geometric realization* of $\mathcal{P}$, by the following recipe:

1. For each word W_i, let P_i denote the convex k-sided polygonal region in the plane that has its center at the origin, sides of length 1, and one vertex on the positive y-axis. (Here k is the length of the word W_i.)

2. Define a one-to-one correspondence between the symbols of W_i and the edges of P_i in counterclockwise order, starting at the vertex on the y-axis.

3. Let $|\mathcal{P}|$ denote the quotient space of $\coprod_i P_i$ determined by identifying each pair of edges that have the same edge symbol according to the simplicial homeomorphism that matches up the first vertices in counterclockwise order if both edges have the same label a or a^{-1}, and matches the first vertex of one with the second vertex of the other if the edges are labeled a and a^{-1}.

If $\mathcal{P}$ is one of the special presentations with a word of length 2, motivated by Propositions 6.2 and 6.3 we define $|\mathcal{P}|$ to be the sphere if the word is aa^{-1} or $a^{-1}a$, and the projective plane if it is aa or $a^{-1}a^{-1}$ (Figure 6.15).

The interiors, edges, and vertices of the polygonal regions P_i are called the faces, edges, and vertices of the presentation. The number of faces is the same as the number of words, while the number of edges is the total number of symbols that occur in all the words. For an edge labeled a, the *initial vertex* is the first one in counterclockwise order, and the *terminal vertex*

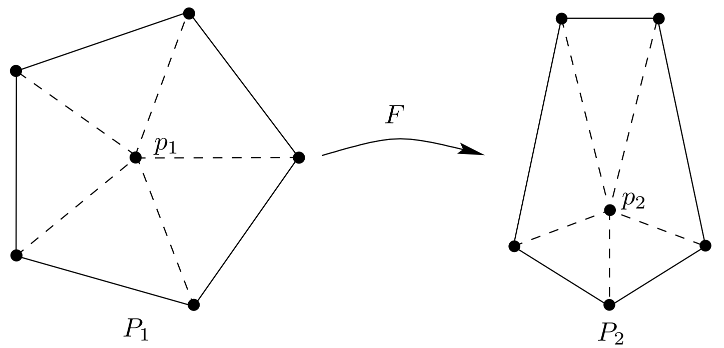


FIGURE 6.16. Simplicial homeomorphism between polygons.

is the other one; for an edge labeled a^{-1}, these definitions are reversed. In terms of our informal description above, if we label each edge with an arrow pointing counterclockwise when the symbol is a and clockwise when it is a^{-1}, the arrow points from the initial vertex to the terminal vertex.

The geometric realization of a presentation with only one face is connected, because it is a quotient of a single connected polygon; with more than one face, it might or might not be connected. Although for definiteness we have defined the geometric realization as a quotient of a disjoint union of a specific collection of polygonal regions, the following lemma shows that we could have used arbitrary disjoint convex polygonal regions in the plane with the same numbers of edges, because between any two polygonal regions with the same sequence of edge labels there is a homeomorphism that respects the edge labels. Because of this, in the arguments that follow we will illustrate our presentations with any convex polygons that are convenient.

Lemma 6.9. *Let P_1, P_2 be convex polygons with the same number of edges, and let $f\colon \partial P_1 \to \partial P_2$ be a simplicial homeomorphism. Then f extends to a homeomorphism $F\colon P_1 \to P_2$.*

Proof. Choose any points $p_i \in \operatorname{Int} P_i$, $i = 1, 2$. By convexity, the line segment from p_i to each vertex of P_i lies entirely in P_i. The edges and vertices of P_i together with p_i, these new line segments, and the triangles they bound form a Euclidean simplicial complex whose polyhedron is P_i (Figure 6.16). The required map $F\colon P_1 \to P_2$ is the simplicial map whose restriction to ∂P_1 is f and that takes p_1 to p_2. □

A polygonal presentation is called a *surface presentation* if each symbol $a \in S$ occurs exactly twice in $W_1, \ldots, W_k$ (counting either a or a^{-1} as one occurrence). By Proposition 6.4, the geometric realization of a surface presentation is a compact surface.

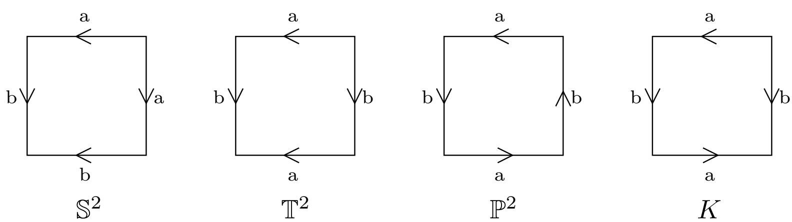


FIGURE 6.17. Standard presentations.

Example 6.10. The following surfaces are determined by the given polygonal presentations, which we call their *standard presentations* (Figures 6.15 and 6.17).

(a) The sphere: $\langle a \mid aa^{-1}\rangle$ or $\langle a, b \mid abb^{-1}a^{-1}\rangle$ (Proposition 6.2).

(b) The torus: $\langle a, b \mid aba^{-1}b^{-1}\rangle$ (Example 3.24).

(c) The projective plane: $\langle a \mid aa\rangle$ or $\langle a, b \mid abab\rangle$ (Proposition 6.3).

(d) The Klein bottle: $\langle a, b \mid abab^{-1}\rangle$ (Example 6.5).

If $\mathcal{K}$ is a 2-dimensional simplicial complex in which every simplex is contained in some 2-simplex, then $\mathcal{K}$ determines in an obvious way a polygonal presentation in which each 2-simplex corresponds to a word of length 3.

For later use in proving the classification theorem, we will develop some general rules for transforming polygonal presentations. If two presentations $\mathcal{P}_1$ and $\mathcal{P}_2$ have homeomorphic geometric realizations, we will say that they are *topologically equivalent* and write $\mathcal{P}_1 \approx \mathcal{P}_2$.

The following operations are called *elementary transformations* of a polygonal presentation. As a matter of notation, in what follows S will denote any sequence of symbols; $a, b, c, a_1, a_2, \ldots$ will denote any symbols from S or their inverses; e will denote any symbol not in S; and $W_1, W_2, \ldots$ will denote any words made from the symbols in S. Given two words W_1, W_2, the notation W_1W_2 denotes the word formed by concatenating W_1 and W_2 together. We adopt the convention that $(a^{-1})^{-1} = a$.

- Relabeling: Changing all occurrences of a symbol a to a new symbol not already in the presentation, interchanging all occurrences of two symbols a and b, or interchanging all occurrences of a and a^{-1} for some $a \in S$.

- Subdividing: Replacing every occurrence of a by ae and every occurrence of a^{-1} by $e^{-1}a^{-1}$, where e is a new symbol not already in the presentation.

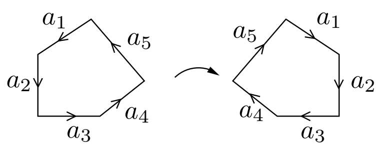


FIGURE 6.18. Reflecting.

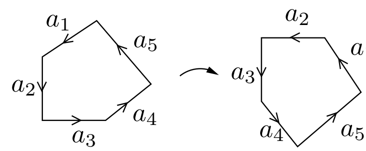


FIGURE 6.19. Rotating.

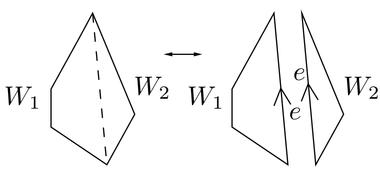


FIGURE 6.20. Cutting/pasting.

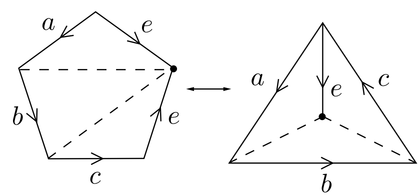


FIGURE 6.21. Folding/unfolding.

- Consolidating: If a and b always occur adjacent to each other either as ab or $b^{-1}a^{-1}$, replacing every occurrence of ab by a and every occurrence of $b^{-1}a^{-1}$ by a^{-1}, provided that the result is one or more words of length at least 3 or a single word of length 2.

- Reflecting (Figure 6.18):

$$\langle S \mid a_1 \dots a_m,\ W_2, \dots, W_k \rangle \mapsto \langle S \mid a_m^{-1} \dots a_1^{-1},\ W_2, \dots, W_k \rangle.$$

- Rotating (Figure 6.19):

$$\langle S \mid a_1 a_2 \dots a_m,\ W_2, \dots, W_k \rangle \mapsto \langle S \mid a_2 \dots a_m a_1,\ W_2, \dots, W_k \rangle.$$

- Cutting (Figure 6.20): If W_1 and W_2 both have length at least 2,

$$\langle S \mid W_1 W_2,\ W_3, \dots, W_k \rangle \mapsto \langle S, e \mid W_1 e,\ e^{-1} W_2,\ W_3, \dots, W_k \rangle.$$

- Pasting (Figure 6.20):

$$\langle S, e \mid W_1 e,\ e^{-1} W_2,\ W_3, \dots, W_k \rangle \mapsto \langle S \mid W_1 W_2,\ W_3, \dots, W_k \rangle.$$

- Folding (Figure 6.21): If W_1 has length at least 3,

$$\langle S, e \mid W_1 e e^{-1},\ W_2, \dots, W_k \rangle \mapsto \langle S \mid W_1,\ W_2, \dots, W_k \rangle.$$

  We also allow W_1 to have length 2, provided that the presentation has only one word.

- UNFOLDING (Figure 6.21):

$$\langle S \mid W_1,\ W_2,\dots,W_k\rangle \mapsto \langle S,e \mid W_1ee^{-1},\ W_2,\dots,W_k\rangle.$$

Proposition 6.11. *Each elementary transformation of a polygonal presentation produces a topologically equivalent presentation.*

Proof. Clearly, subdividing and consolidating are inverses of each other, as are cutting/pasting and folding/unfolding, so by symmetry only one of each pair needs to be proved. We demonstrate the techniques by proving the proposition for cutting and folding, and leave the rest as exercises.

To prove that cutting produces a homeomorphic geometric realization, let P_1 and P_2 be convex polygons labeled W_1e and $e^{-1}W_2$, respectively, and let P' be a convex polygon labeled W_1W_2. For the moment, let us assume that these are the only words in their respective presentations. Let $\pi\colon P_1 \amalg P_2 \to S$ and $\pi'\colon P' \to S'$ denote the respective quotient maps. The line segment going from the terminal vertex of W_1 in P' to its initial vertex lies in P' by convexity; label this segment e. By Lemma 6.9 there is a continuous map $f\colon P_1 \amalg P_2 \to P'$ that takes each edge of P_1 or P_2 to the edge in P' with the corresponding label, and whose restriction to each P_i is a homeomorphism. By the closed map lemma, f is a quotient map. Since f identifies the two edges labeled e and e^{-1} but nothing else, the quotient maps $\pi'\circ f$ and π make precisely the same identifications, so their quotient spaces are homeomorphic. If there are other words $W_3,\dots,W_k$, we just extend f by declaring it to be the identity on their respective polygons and proceed as above.

For folding, as before we can ignore the additional words $W_2,\dots,W_k$. If W_1 has length 2, we can subdivide to lengthen it, then perform the folding operation, and then consolidate, so we assume that W_1 has length at least 3. Assume first that $W_1 = abc$ has length exactly 3. Let P and P' be convex polygons with edge labels $abcee^{-1}$ and abc, respectively, and let $\pi\colon P \to S$, $\pi'\colon P' \to S'$ be the quotient maps. Adding edges as shown in Figure 6.21 turns P and P' into polyhedra of Euclidean simplicial complexes, and there is a unique simplicial map $f\colon P \to P'$ that takes each edge of P to the edge of P' with the same label. As before, $\pi'\circ f$ and π are quotient maps that make the same identifications, so their quotient spaces are homeomorphic.

If W_1 has length 4 or more, we can write $W_1 = Xbc$ for some X of length at least 2. Then we cut along a to obtain

$$\langle S,b,c,e \mid Xbcee^{-1}\rangle \approx \langle S,a,b,c,e \mid Xa^{-1}, abcee^{-1}\rangle,$$

and proceed as before. □

Exercise 6.3. Prove the rest of Proposition 6.11. Note that you will have to consider a word of length 2 as a special case when treating subdividing or consolidating.

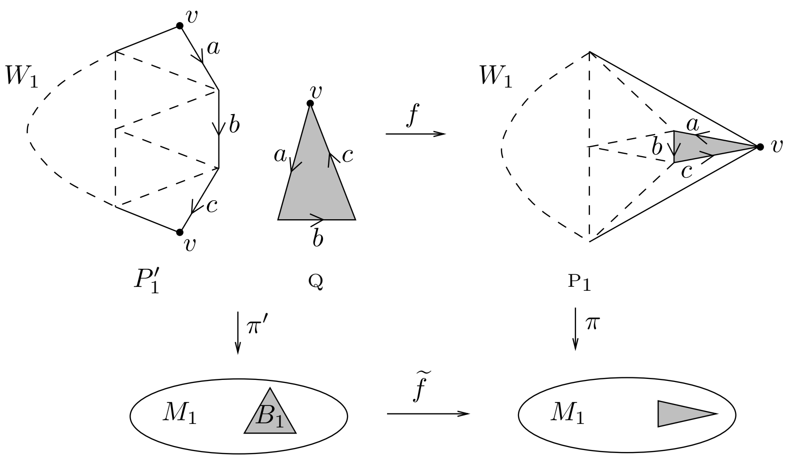


FIGURE 6.22. The presentation $\langle S_1, a, b, c \mid W_1 c^{-1} b^{-1} a^{-1}, abc\rangle$.

Next we need to find standard polygonal presentations for connected sums. The key is the following proposition.

Proposition 6.12. *Let M_1 and M_2 be surfaces determined by presentations $\langle S_1 \mid W_1\rangle$ and $\langle S_2 \mid W_2\rangle$, respectively, in which S_1 and S_2 are disjoint sets and each presentation has a single face. Then $\langle S_1, S_2 \mid W_1 W_2\rangle$ is a presentation of a connected sum $M_1 \# M_2$. (Here $W_1 W_2$ denotes the word formed by concatenating W_1 and W_2 together.)*

Proof. Consider the presentation $\langle S_1, a, b, c \mid W_1 c^{-1} b^{-1} a^{-1}, abc\rangle$ (pictured in the left half of Figure 6.22). Pasting along a and folding twice, we see that this presentation is equivalent to $\langle S_1 \mid W_1\rangle$ and therefore is a presentation of M_1. Let B_1 denote the image in M_1 of the interior of the polygonal region bounded by triangle abc. We will show below that B_1 is a regular Euclidean disk in M_1. Assuming this, it follows immediately that the geometric realization of $\langle S_1, a, b, c \mid W_1 c^{-1} b^{-1} a^{-1}\rangle$ is homeomorphic to $M_1 \smallsetminus B_1$ (which we denote by M_1'), and ∂B_1 is the image of the edges $c^{-1} b^{-1} a^{-1}$. A similar argument shows that $\langle S_2, a, b, c \mid abc W_2\rangle$ is a presentation of M_2 with a Euclidean disk removed (denoted by M_2'). Therefore, $\langle S_1, S_2 \mid W_1 c^{-1} b^{-1} a^{-1}, abc W_2\rangle$ is a presentation of $M_1' \amalg M_2'$ with the boundaries of the respective disks identified, which is $M_1 \# M_2$. Pasting along a and folding twice, we arrive at the presentation $\langle S_1, S_2 \mid W_1 W_2\rangle$.

It remains only to show that B_1 is a regular disk in M_1, i.e., that it has an open disk neighborhood in which $\overline{B}_1$ corresponds to a smaller closed disk. Let P_1, P_1', and Q be convex polygons with edges labeled by the words W_1, $W_1 c^{-1} b^{-1} a^{-1}$, and abc, respectively. Triangulating the polygons as shown in Figure 6.22, we obtain a simplicial map $f\colon P_1' \amalg Q \to P_1$ such that $\pi \circ f$

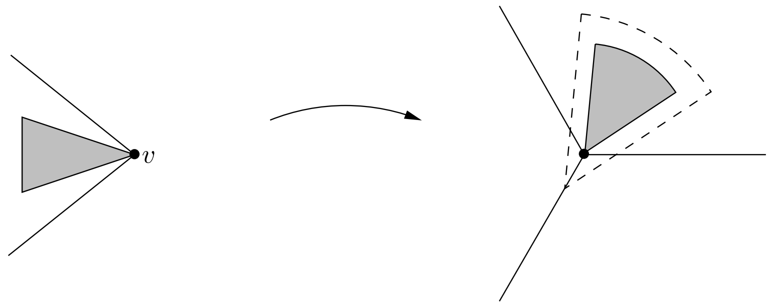


FIGURE 6.23. Showing that B_1 is a regular disk.

makes the same identifications as π', and so descends to a homeomorphism $\widetilde{f}\colon M_1 \to M_1$. The inverse image of $\widetilde{f}(B_1)$ under π is a small triangle in P_1 sharing one vertex v in common with P_1.

Now look back at the proof in Proposition 6.4 that the quotient space of a surface presentation is a manifold. In constructing a Euclidean neighborhood of a vertex point, we assembled "wedges" at the various vertices into a Euclidean disk. Applying that construction to the vertex v, the small triangle is taken to a set that is homeomorphic to a closed disk in the plane (Figure 6.23), and it is an easy matter to extend that homeomorphism to a slightly larger open disk. □

Example 6.13. Using the preceding proposition, we can augment our list of presentations of known surfaces as follows:

- Connected sum of n tori:

$$\langle a_1, b_1, \dots, a_n, b_n \mid a_1 b_1 a_1^{-1} b_1^{-1} \dots a_n b_n a_n^{-1} b_n^{-1} \rangle.$$

- Connected sum of n projective planes:

$$\langle a_1, \dots, a_n \mid a_1 a_1 \dots a_n a_n \rangle.$$

We call these the *standard presentations* of these surfaces.

Classification of Surface Presentations

We are now ready to state the main result in the classification of surfaces. This theorem was first proved in 1907 by Max Dehn and Poul Heegaard [DH07] under the assumption that the surface had some polygonal presentation. Using the triangulation theorem of Chapter 5, we now know that every compact surface has a triangulation, which determines a polygonal presentation.

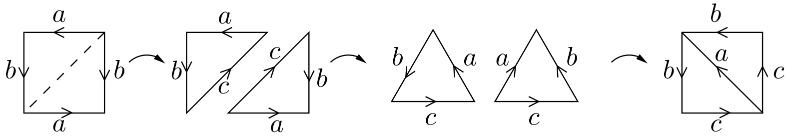


FIGURE 6.24. Transforming the Klein bottle to $\mathbb{P}^2 \# \mathbb{P}^2$.

Theorem 6.14 (Classification of Surface Presentations). *Any surface presentation is equivalent by a sequence of elementary transformations to a presentation of one of the following:*

(*a*) *the sphere* $\mathbb{S}^2$;

(*b*) *a connected sum* $\mathbb{T}^2 \# \cdots \# \mathbb{T}^2$; *or*

(*c*) *a connected sum* $\mathbb{P}^2 \# \cdots \# \mathbb{P}^2$.

Therefore, every compact surface is homeomorphic to one of the surfaces in this list.

Before we prove the theorem, we need to make one important observation. You might have noticed that some surfaces appear to be absent from the list—the Klein bottle, for example, and $\mathbb{T}^2 \# \mathbb{P}^2$, and, for that matter, every connected sum involving both tori and projective planes. These apparent deficiencies are explained by the following two lemmas.

Lemma 6.15. *The Klein bottle is homeomorphic to* $\mathbb{P}^2 \# \mathbb{P}^2$.

Proof. By a sequence of elementary transformations, we find that the Klein bottle has the following presentations (see Figure 6.24):

$$\begin{aligned}
&\langle a, b \mid abab^{-1} \rangle \\
&\qquad \approx \langle a, b, c \mid abc,\ c^{-1}ab^{-1} \rangle && \text{(cut along } c\text{)} \\
&\qquad \approx \langle a, b, c \mid bca, a^{-1}cb \rangle && \text{(rotate and reflect)} \\
&\qquad \approx \langle b, c \mid bbcc \rangle. && \text{(paste along } a \text{ and rotate)}
\end{aligned}$$

This last is a presentation of the connected sum of two projective planes. □

Lemma 6.16. *The connected sum* $\mathbb{T}^2 \# \mathbb{P}^2$ *is homeomorphic to* $\mathbb{P}^2 \# \mathbb{P}^2 \# \mathbb{P}^2$.

Proof. Start with $\langle a, b, c \mid abab^{-1}cc \rangle$ (Figure 6.25), which is a presentation of $K \# \mathbb{P}^2$, and therefore by the preceding lemma of $\mathbb{P}^2 \# \mathbb{P}^2 \# \mathbb{P}^2$. Following Figure 6.25, we cut along d, paste along c, cut along e, and paste along b, rotating and reflecting as necessary, to obtain $\langle a, d, e \mid a^{-1}d^{-1}adee \rangle$, which is a presentation of $\mathbb{T}^2 \# \mathbb{P}^2$. □

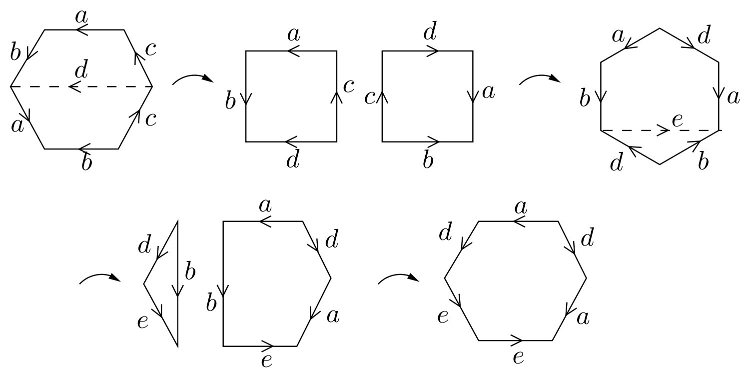


FIGURE 6.25. Transforming $\mathbb{P}^2 \# \mathbb{P}^2 \# \mathbb{P}^2$ to $\mathbb{T}^2 \# \mathbb{P}^2$.

Proof of the classification theorem. By the triangulation theorem we can assume that M comes with a given presentation. We will prove the theorem by transforming this presentation to one of our standard presentations in several steps. Let us say that a pair of edges that are to be identified are *complementary* if they appear in the presentation as both a and a^{-1}, and *twisted* if they appear as $a, \ldots, a$ or as $a^{-1}, \ldots, a^{-1}$. (The terminology reflects the fact that if a polygonal region is cut from a piece of paper, you have to twist the paper to paste together a twisted edge pair, but not for a complementary pair.)

Step 1: *M admits a presentation with only one face.* Since M is connected, if there are two or more faces, some edge in one face must be identified with an edge in a different face; otherwise, M would be the disjoint union of the quotients of its faces, and since each such quotient is open and closed, they would provide a separation of M. Thus by performing successive pasting transformations (together with rotations and reflections as necessary), we can reduce the number of faces in the presentation to one.

Step 2: *Either M is homeomorphic to the sphere, or M admits a presentation in which there are no adjacent complementary pairs.* Each adjacent complementary pair can be eliminated by folding, unless it is the only pair of edges in the presentation; in this case the presentation is equivalent to $\langle a \mid aa^{-1} \rangle$ and M is homeomorphic to the sphere.

From now on, we assume that the presentation is not the standard presentation of the sphere.

Step 3: *M admits a presentation in which all twisted pairs are adjacent.* If a twisted pair is not adjacent, the presentation is described by a word of the form $VaWa$, where neither V nor W is empty. Figure 6.26 shows how to transform the word $VaWa$ into $VW^{-1}bb$ by cutting along b, reflecting, and pasting along a. (Here W^{-1} denotes the word obtained from W by reflecting). In this last presentation, the twisted pair a, a has been replaced

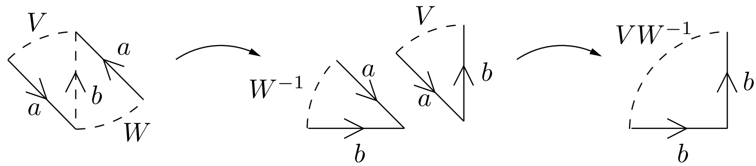


FIGURE 6.26. Making a twisted pair adjacent.

by another twisted pair b, b, which is now adjacent. Moreover, no other adjacent pairs have been separated. We may have created some new twisted pairs when we reflected W, but we decreased the total number of nonadjacent pairs by at least one. Thus, after finitely many such operations, there will be no more nonadjacent twisted pairs. We may also have created some new adjacent complementary pairs. These can be eliminated by repeating Step 2, which does not increase the number of nonadjacent pairs.

STEP 4: *M admits a presentation in which all vertices are identified to a single point.* Choose some equivalence class of vertices, and call it v. If there are vertices that are not identified with v, there must be some edge that goes from a v vertex to a vertex in some other equivalence class; label the edge a and the other vertex class w (Figure 6.27). The other edge that touches a at its v vertex cannot be identified with a: If it were complementary to a, we would have eliminated both edges in Step 3, while if it formed a twisted pair with a, then the quotient map would identify the initial and terminal vertices of a with each other, which we are assuming is not the case. So label this other edge b, and label its other vertex x (this one may be identified with v, w, or neither one).

Somewhere in the polygon is another edge labeled b or b^{-1}. Let us assume for definiteness that it is b^{-1}; the argument for b is similar except for an extra reflection. Thus we can write the word describing the presentation in the form $baXb^{-1}Y$, where X and Y are unknown words, not both empty. Now cut along c and paste along b as in Figure 6.27. In the new presentation, the number of vertices labeled v has decreased, and the number labeled w has increased. We may have introduced a new adjacent complementary pair, so perform Step 2 again to remove it. This may again decrease the number of vertices labeled v (for example, if a v vertex lies between edges labeled aa^{-1} that are eliminated by folding), but it cannot increase their number. So repeating this sequence a finite number of times—decrease the v vertices by one, then eliminate adjacent complementary edges—we eventually eliminate the vertex class v from the presentation altogether. Iterate this procedure for each vertex class until there is only one left.

STEP 5: *If the presentation has any complementary pair a, a^{-1}, then it has another b, b^{-1} that occurs intertwined with the first, as in* $a, \dots, b, \dots, a^{-1}, \dots, b^{-1}$. If this is not the case, then the presentation is of the form $aXa^{-1}Y$, where X contains only matched complementary pairs or

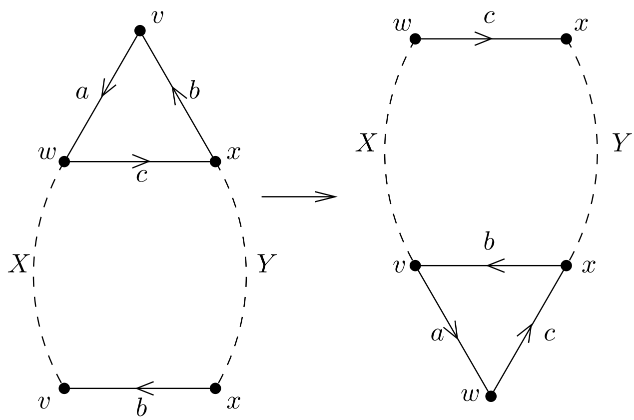


FIGURE 6.27. Reducing the number of vertices equivalent to v.

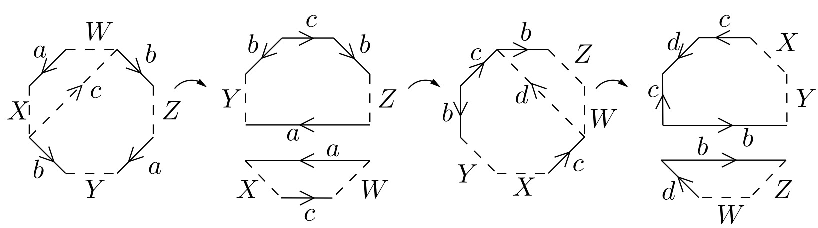


FIGURE 6.28. Bringing intertwined complementary pairs together.

adjacent twisted pairs. Thus each edge in X is identified only with another edge in X, and the same is true of Y. But this means that the terminal vertices of the a and a^{-1} edges, both of which touch only X, can be identified only with vertices in X, while the initial vertices can be identified only with vertices in Y. This is a contradiction, since all vertices are identified together by Step 4.

STEP 6: *M admits a presentation in which all intertwined complementary pairs occur together with no other edges intervening: $aba^{-1}b^{-1}$.* If the presentation is given by the word $WaXbYa^{-1}Zb^{-1}$, perform the elementary transformations indicated in Figure 6.28 (cut along c, paste along a, cut along d, and paste along b) to obtain the new word $cdc^{-1}d^{-1}WZYX$. This replaces the old intertwined set of pairs with a new adjacent set $cdc^{-1}d^{-1}$, without separating any edges that were previously adjacent. Repeat this for each set of intertwined pairs. (Note that this step requires no reflections.)

STEP 7: *M is homeomorphic to either a connected sum of tori or a connected sum of projective planes.* From what we have done so far, all twisted pairs occur adjacent to each other, and all complementary pairs occur in intertwined groups like $aba^{-1}b^{-1}$. This is a presentation of a connected sum of tori (presented by $aba^{-1}b^{-1}$) and projective planes (presented by cc). If there are only tori or only projective planes, we are done.

The only remaining case is that in which the presentation contains both twisted and complementary pairs. In that case, some twisted pair must occur next to a complementary one; thus the presentation is described either by a word of the form $aba^{-1}b^{-1}ccX$ or by one of the form $ccaba^{-1}b^{-1}X$. In either case, this is a connected sum of a torus, a projective plane, and whatever surface is described by the word X. But Lemma 6.16 shows that the standard presentation of $\mathbb{T}^2\#\mathbb{P}^2$ can be transformed to that of $\mathbb{P}^2\#\mathbb{P}^2\#\mathbb{P}^2$. Making this transformation, we eliminate one of the occurrences of $\mathbb{T}^2$ in the connected sum. Iterating this procedure, we eliminate them all, thus completing the proof. $\square$

Combinatorial Invariants

Two of the properties of simplicial complexes introduced in Chapter 5—orientations and the Euler characteristic—generalize easily to polygonal surface presentations, and give us interesting information about surfaces. Let us extend the notion of combinatorial invariance by saying that a property of a polygonal presentation is a combinatorial invariant if it is unchanged by elementary transformations. Of course, any topological invariant of the geometric realization (such as connectedness) is automatically a combinatorial invariant, because elementary transformations yield homeomorphic surfaces. It would not be difficult to show that every polygonal presentation can be triangulated and that the two notions of combinatorial invariants coincide; however, in this case it is easier just to work directly with the polygonal presentations.

The Euler Characteristic

The Euler characteristic can be generalized to polygonal presentations in the following form. If $\mathcal{P}$ is a polygonal presentation, define the *Euler characteristic* of $\mathcal{P}$ to be $V - E + F$, where V is the number of vertices (after identifications), E is the number of edge symbols (which is equal to the number of edges after identifications), and F is the number of faces. If $\mathcal{P}$ is the presentation determined by a 2-dimensional simplicial complex, it is easy to check that this definition agrees with the definition given in Chapter 5.

Theorem 6.17. *The Euler characteristic of a polygonal presentation is unchanged by elementary transformations.*

Proof. It is obvious that relabeling, rotating, and reflecting do not change the Euler characteristic of a presentation, because they leave the numbers of vertices, edges, and faces individually unchanged. For the other transformations, we need only check that the changes to these three numbers cancel out. Subdividing increases both the number of edges and the number of vertices by one, leaving the number of faces unchanged. Cutting increases both the number of edges and the number of faces by one, and leaves the number of vertices unchanged. Unfolding increases the number of edges and the number of vertices by one, and leaves the number of faces unchanged. Finally, consolidating, pasting, and folding leave the Euler characteristic unchanged, since they are the inverses of subdividing, cutting, and unfolding, respectively. □

Proposition 6.18 (Euler Characteristics of Compact Surfaces). *The Euler characteristic of a standard surface presentation is equal to*

(*a*) 2 *for the sphere;*

(*b*) $2 - 2n$ *for the connected sum of* n *tori;*

(*c*) $2 - n$ *for the connected sum of* n *projective planes.*

Proof. Just compute. □

These results allows us to conclude a great deal about a surface from a given presentation, without actually carrying out the reduction to a standard presentation. For example, any presentation with Euler characteristic 2 gives the sphere, and a presentation with Euler characteristic 0 gives either the torus or the Klein bottle $\approx \mathbb{P}^2 \# \mathbb{P}^2$.

At the moment, we do not know that the Euler characteristic is a *topological invariant*, for the simple reason that we still do not know that the standard surfaces on our list are not homeomorphic to each other. (If you do not believe this, just try to prove, for example, that the projective plane is not homeomorphic to the torus using the techniques we have developed so far!) The problem is that we cannot yet rule out the possibility that $\mathbb{P}^2$, say, could have a presentation that is so exotic that it is not related to the standard one by a series of elementary transformations, but somehow manages to reduce to a presentation of the torus after following the algorithm of the classification theorem. We will remedy this deficiency in Chapter 10, when we show that all of our standard compact surfaces are topologically distinct; only then will we be able to complete the classification of compact surfaces.

The Euler characteristic can be used by itself to distinguish presentations that reduce to connected sums of different numbers of tori or connected sums of different numbers of projective planes. However, to distinguish a presentation of the connected sum of n tori from one of the connected sum of $2n$ projective planes (for example, the torus from the Klein bottle), we will need one more property: orientability.

Orientability

In Chapter 5 we introduced the notion of orientation of a simplicial complex. In this section we show how to extend it to polygonal presentations.

Note that a permutation of a set with three elements is even if and only if it is a cyclic permutation (i.e., of the form $i \mapsto j \mapsto k \mapsto i$). Thus an orientation of a 2-simplex is just an equivalence class of vertex orderings modulo cyclic permutations. Suppose $\mathcal{P}$ is a surface presentation arising from a simplicial complex, so that every word has length 3. Each word determines an orientation of the associated 2-simplex, by ordering the vertices in counterclockwise order. It is easy to check that when two simplices are glued together via an edge pairing, their orientations are consistent if and only if the edge pair is complementary. Motivated by this, we make the following definition. A surface presentation $\mathcal{P}$ is said to be *oriented* if it has no twisted edge pairs. Intuitively, this means that you can decide which is the "front" side (or "outside") of $|\mathcal{P}|$ by coloring the top surface of each polygon white and the bottom side gray; the condition on edge pairs ensures that the colors will match up when edges are pasted together.

A compact surface is said to be *orientable* if it admits an oriented presentation. By looking a little more closely at the proof of the classification theorem, we can identify exactly which compact surfaces are orientable.

Proposition 6.19. *A compact surface is orientable if and only if it is homeomorphic to the sphere or a connected sum of one or more tori.*

Proof. The standard presentations of the sphere and the connected sums of tori are oriented, so these surfaces are certainly orientable. To show that an orientable surface is homeomorphic to one of these, let M be any surface that admits at least one orientable presentation. Starting with that presentation, follow the algorithm described in the proof of the classification theorem to transform it to one of the standard presentations. The only elementary transformation that can introduce a twisted pair into an oriented presentation is reflection. The only steps in which reflections are used are Steps 3, 4, and 7, and you can check that none of those steps require any reflections if there were no twisted pairs to begin with. Thus the classification theorem tells us that the presentation can be reduced to one of the standard ones with no twisted pairs, which means that M is homeomorphic to a sphere or a connected sum of tori. □

Because of this result, the connected sum of n tori is also known as the *orientable surface of genus* n, and the connected sum of n projective planes is called the *nonorientable surface of genus* n. By convention, the sphere is the (unique, orientable) surface of genus 0. Technically, this terminology is premature, because we still do not know that a connected sum of projective planes is not homeomorphic to an oriented surface. But for now we will go ahead and use this standard terminology with the caveat that all we know about the "nonorientable surface of genus n" is that its standard presentation is not oriented.

Before moving away from classification theorems, it is worth remarking on the situation with higher-dimensional manifolds. Because of the triangulation theorem for 3-manifolds stated in Chapter 5, one might hope that a similar approach to classifying 3-manifolds might bear fruit. Unfortunately, the combinatorial problem of reducing any given 3-manifold triangulation to some standard form is, so far, unsolved. And this approach cannot get us very far in dimensions higher than 3, because we do not have triangulation theorems. Thus, in order to make any progress in understanding higher-dimensional manifolds, as well as to resolve the question of whether the standard surfaces are distinct, we will need to develop more powerful tools. This we will do in the remainder of the book.

Problems

6-1. For each of the following surface presentations, compute the Euler characteristic, and then apply the algorithm of the classification theorem to determine which of our standard surfaces it is.

(a) $\langle a, b, c \mid abacb^{-1}c^{-1}\rangle$.

(b) $\langle a, b, c \mid abca^{-1}b^{-1}c^{-1}\rangle$.

(c) $\langle a, b, c, d, e, f \mid abc,\ bde,\ c^{-1}df,\ e^{-1}fa\rangle$.

(d) $\langle a, b, c, d, e, f, g, h, i, j, k, l, m, n, o \mid$
$abc,\ bde,\ dfg,\ fhi,\ haj,\ c^{-1}kl,\ e^{-1}mn,$
$g^{-1}ok^{-1},\ i^{-1}l^{-1}m^{-1},\ j^{-1}n^{-1}o^{-1}\rangle$.

6-2. Show that a connected sum of one or more projective planes contains a subspace that is homeomorphic to the Möbius band.

6-3. Show that the projective plane is homeomorphic to the quotient obtained by gluing a Möbius band and a disk together along their common boundary.

6-4. Show that the Klein bottle is homeomorphic to the quotient obtained by gluing two Möbius bands together along their common boundary.

7

Homotopy and the Fundamental Group

The results of the preceding chapter left a serious gap in our attempt to classify compact 2-manifolds up to homeomorphism: Although we have exhibited a list of surfaces and shown that every compact, connected surface is homeomorphic to one on the list, we still have no way of knowing when two surfaces are *not* homeomorphic. For all we know, all the surfaces on our list might be homeomorphic to the sphere! (Think, for example, of the unexpected homeomorphism between $\mathbb{P}^2 \# \mathbb{P}^2 \# \mathbb{P}^2$ and $\mathbb{T}^2 \# \mathbb{P}^2$.)

To distinguish nonhomeomorphic surfaces, we need topological invariants. For some surfaces, the properties we already know suffice. For example, the 2-sphere is not homeomorphic to the plane because one is compact, while the other is not. The plane, the disjoint union of two planes, and the disjoint union of three planes are all topologically distinct, because they have different numbers of components. It follows from Problem 4-1 that the line is not homeomorphic to the plane; the proof involved a rather subtle use of connectedness. But to decide whether, for example, the sphere is homeomorphic to the torus, or the plane is homeomorphic to the punctured plane $\mathbb{R}^2 \smallsetminus \{0\}$, we need to introduce some new invariants.

In this chapter we begin our study of the fundamental group, an algebraic group that measures the number of "holes" in a space, in a certain sense. To set the stage, let us think about the difference between the plane and the punctured plane. Both are connected, noncompact 2-manifolds, so they cannot be distinguished by any of the basic topological properties that we have discussed so far. Yet intuition suggests that they should not be homeomorphic to each other because the punctured plane has a "hole," while the full plane does not.

To see how this distinction might be detected topologically, observe that every closed curve in $\mathbb{R}^2$ can be continuously shrunk to a point (you will prove this rigorously in Exercise 7.3 below); by contrast, it is intuitively clear that a circle drawn around the hole in the space $\mathbb{R}^2 \smallsetminus \{0\}$ can never be continuously shrunk to a point while remaining in the space, and in fact cannot be deformed into any closed path that does not go around the hole.

We will define an equivalence relation on closed paths with a fixed starting and ending point: Two paths are equivalent if one can be continuously deformed into the other while keeping the starting and ending point fixed. The set of equivalence classes is called the fundamental group of the space; the product of two elements of the group is obtained by first following one path and then the other. After making the basic definitions, we will prove that homeomorphic spaces have isomorphic fundamental groups. Then we will prove that the fundamental group satisfies an even stronger invariance property, that of homotopy invariance. As a consequence, we will be able to reduce the computations of fundamental groups of many spaces to those of simpler ones.

Proving that the fundamental group of a space is *not* trivial turns out to be somewhat harder, and we will not do so until the next chapter.

Homotopy

Let X and Y be topological spaces, and let $f, g\colon X \to Y$ be continuous maps. A *homotopy* from f to g is a continuous map $H\colon X \times I \to Y$ (where $I = [0,1]$ is the unit interval) such that for all $x \in X$,

$$H(x,0) = f(x); \qquad H(x,1) = g(x). \tag{7.1}$$

If there exists a homotopy from f to g, we say that f and g are *homotopic*, and write $f \simeq g$ (or $H\colon f \simeq g$ if we want to emphasize the specific homotopy).

A homotopy defines a one-parameter family of maps $H_t(x) = H(x,t)$ for $0 \le t \le 1$ (Figure 7.1), and condition (7.1) says that $H_0 = f$ and $H_1 = g$. We usually think of the parameter t as time, and think of H as giving a "deformation" of f into g as t goes from 0 to 1. The continuity of H guarantees that this deformation proceeds continuously without breaks or jumps.

Lemma 7.1. *For any topological spaces X and Y, homotopy is an equivalence relation on the set of all continuous maps from X to Y.*

Proof. Any map f is homotopic to itself via the trivial homotopy $H(x,t) = f(x)$, so homotopy is reflexive. Similarly, if $H\colon f \simeq g$, then a homotopy from g to f is given by $\widetilde{H}(x,t) = H(x,1-t)$, so homotopy is symmetric. Finally, if $F\colon f \simeq g$ and $G\colon g \simeq h$, define $H\colon X \times I \to Y$ by following F at

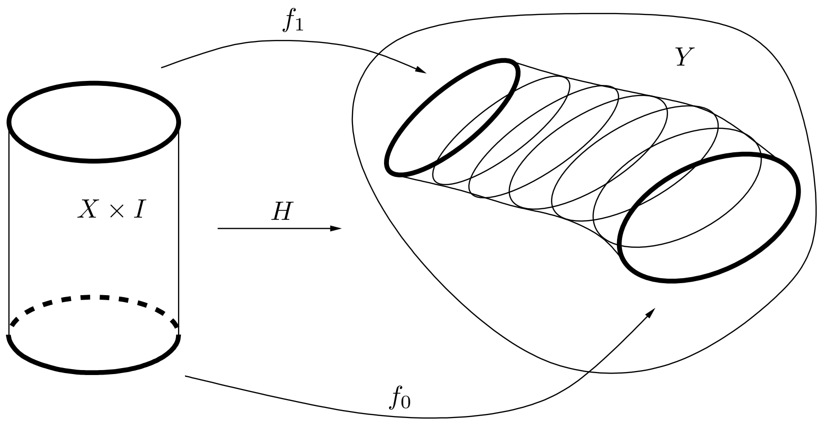


FIGURE 7.1. A homotopy between f_0 and f_1.

double speed for $0 \le t \le \frac{1}{2}$, and then following G at double speed for the remainder of the unit interval. Formally,

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

Since $F(x,1) = g(x) = G(x,0)$, the two definitions of H agree at $t = \frac{1}{2}$, where they overlap. Thus H is continuous by the gluing lemma, and is therefore a homotopy between f and h. This shows that homotopy is transitive. $\square$

Lemma 7.2. *The homotopy relation is preserved by composition: If*

$$f_0, f_1 \colon X \to Y \quad \textit{and} \quad g_0, g_1 \colon Y \to Z$$

are continuous maps with $f_0 \simeq f_1$ *and* $g_0 \simeq g_1$, *then* $g_0 \circ f_0 \simeq g_1 \circ f_1$.

Proof. Suppose $F \colon f_0 \simeq f_1$ and $G \colon g_0 \simeq g_1$ are homotopies. Define $H \colon X \times I \to Z$ by $H(x,t) = G(F(x,t),t)$. At $t = 0$, $H(x,0) = G(f_0(x),0) = g_0(f_0(x))$, and at $t = 1$, $H(x,1) = G(f_1(x),1) = g_1(f_1(x))$. Thus H is a homotopy from $g_0 \circ f_0$ to $g_1 \circ f_1$. $\square$

Example 7.3. Define maps $f \colon \mathbb{R} \to \mathbb{R}^2$ by

$$f(x) = (x, x^2); \qquad g(x) = (x,x).$$

Then the map $H(x,t) = (x, x^2 - tx^2 + tx)$ is a homotopy from f to g.

Example 7.4. Let $B \subset \mathbb{R}^n$ and let X be any topological space. Suppose $f, g \colon X \to B$ are any two continuous maps with the property that for all

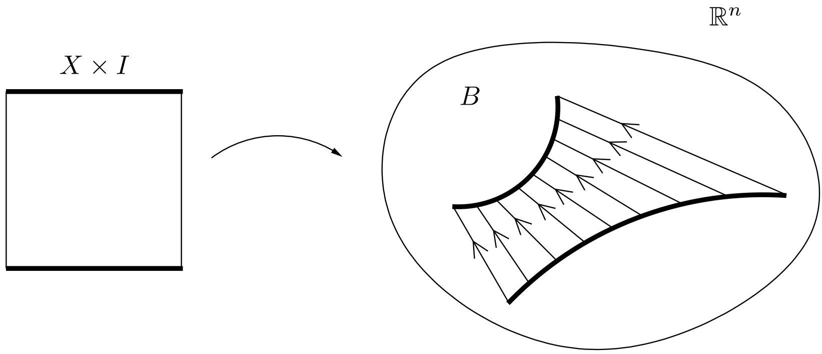


FIGURE 7.2. A straight-line homotopy.

$x \in X$, the line segment from $f(x)$ to $g(x)$ lies in B. This will be the case, for example, if B is convex. Define a homotopy $H\colon f \simeq g$ by letting $H(x,t)$ trace out the line segment from $f(x)$ to $g(x)$ at constant speed as t goes from 0 to 1 (Figure 7.2):

$$H(x,t) = (1-t)f(x) + tg(x).$$

This is called the *straight-line homotopy* between f and g. It shows, in particular, that all maps from a given space into a convex set are homotopic.

The Fundamental Group

Recall that a *path* in a topological space X is a continuous map $f\colon I \to X$. The points $p = f(0)$ and $q = f(1)$ are called the *initial point* and *terminal point* of f, respectively, and we say that f is a path "from p to q." We will use paths to detect "holes" in a space.

Example 7.5. Consider the path $\alpha\colon I \to \mathbb{R}^2 \smallsetminus \{0\}$ defined (in complex notation) by

$$\alpha(s) = e^{2\pi i s}$$

and the map $H\colon I \times I \to \mathbb{R}^2 \smallsetminus \{0\}$ by

$$H(s,t) = e^{2\pi i s t}.$$

At each time t, H_t is a path that follows the circle only as far as angle $2\pi t$, so H_0 is the constant path $c_1(s) \equiv 1$ and $H_1 = \alpha$. Thus H is a homotopy from the constant path to α.

This last example shows that the circular path around the origin is homotopic in $\mathbb{R}^2 \smallsetminus \{0\}$ to a constant path, so that simply asking whether a closed path is homotopic to a constant is not sufficient to detect holes. To remedy this, we will need to consider homotopies of paths throughout which the endpoints stay fixed. More generally, it will be useful to consider homotopies that fix an arbitrary subset of the domain.

Let X and Y be topological spaces, and $A \subset X$ an arbitrary subspace. A homotopy H between maps $f, g\colon X \to Y$ is called a *homotopy relative to A* if

$$H(x,t) = f(x) \quad \text{for all } x \in A,\ t \in I.$$

In other words, for each t, the map H_t agrees with f on A. If there exists such a homotopy, we say that f and g are *homotopic relative to A* and write $f \simeq_A g$. Notice that this implies $g|_A = H_1|_A = f|_A$, so for two maps to be homotopic relative to A they must first of all agree on A.

Now suppose f and g are two paths in X. A *path homotopy* from f to g is a homotopy relative to $\{0, 1\}$, that is, a homotopy that fixes the endpoints for all time. If there exists a path homotopy between f and g, we say they are *path homotopic*, and write $f \sim g$. By the remark above, this is possible only if f and g both have the same initial point and the same terminal point.

Lemma 7.6. *For any points $p, q \in X$, path homotopy is an equivalence relation on the set of all paths from p to q.*

Exercise 7.1. Prove Lemma 7.6.

For any path f in X, we denote the path homotopy equivalence class of f by $[f]$, and call it the *path class* of f. For our purposes, we will be most interested in paths that start and end at the same point. Such a path is called a *loop*. If f is a loop whose initial and terminal point is $q \in X$, we say that f is *based at* q, and we call q the *base point* of f. The set of all loops in X based at q is denoted by $\Omega(X, q)$. The *constant loop* $c_q \in \Omega(X, q)$ is the map $c_q(s) \equiv q$. If a loop is path homotopic to the constant loop, we say that it is *null homotopic*.

One (not very interesting, but sometimes useful) way to get homotopic paths is by the following construction. A *reparametrization* of a path $f\colon I \to X$ is a path of the form $f \circ \varphi$ for some homeomorphism $\varphi\colon I \to I$ fixing 0 and 1.

Lemma 7.7. *Any reparametrization of a path f is path homotopic to f.*

Proof. Suppose $f \circ \varphi$ is a reparametrization of f, and let $H\colon I \times I \to I$ denote the straight-line homotopy from the identity map to φ. Then $f \circ H$ is a homotopy from f to $f \circ \varphi$. □

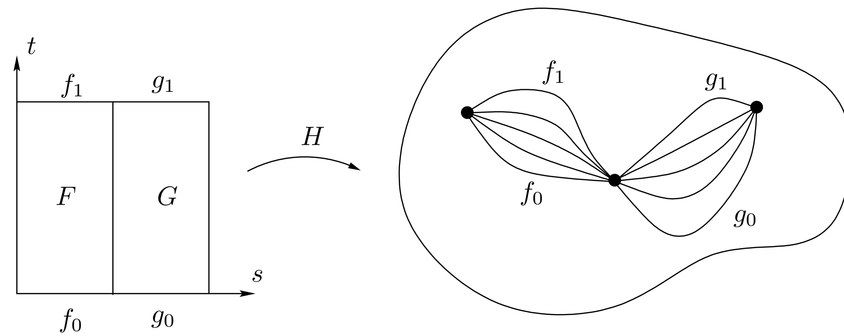


FIGURE 7.3. Homotopy invariance of path multiplication.

Lemma 7.6 says that path homotopy is an equivalence relation on $\Omega(X, q)$. We define the *fundamental group* of X based at q, denoted by $\pi_1(X, q)$, to be the set of path classes of loops based at q.

To make $\pi_1(X, q)$ into a group, we must define a multiplication operation. This is done first on the level of paths: The product of two paths f and g is the path obtained by first following f and then following g, both at double speed. For future use, we will define products of paths in a more general setting—instead of requiring that both paths start and end at the same point, we will require simply that the second one start where the first ends.

Thus let $f, g\colon I \to X$ be paths with $f(1) = g(0)$. We define their *product* $f \cdot g\colon I \to X$ by

$$f \cdot g(s) = \begin{cases} f(2s); & 0 \le s \le \frac{1}{2}; \\ g(2s-1); & \frac{1}{2} \le s \le 1. \end{cases}$$

(Here and throughout the book we will consistently use s as the "space variable" parametrizing individual paths, and reserve t for the "time variable" in homotopies.) The condition $f(1) = g(0)$ guarantees that $f \cdot g$ is continuous by the gluing lemma.

Lemma 7.8 (Homotopy Invariance of Path Multiplication). *Path multiplication is well-defined on path classes. More precisely, if $f_0 \sim f_1$ and $g_0 \sim g_1$, and if $f_0 \cdot g_0$ is defined (that is, if $f_0(1) = g_0(0)$), then $f_1 \cdot g_1$ is also defined and $f_0 \cdot g_0 \sim f_1 \cdot g_1$.*

Proof. Let $F\colon f_0 \sim f_1$ and $G\colon g_0 \sim g_1$ be homotopies (Figure 7.3). The required homotopy $H\colon f_0 \cdot g_0 \sim f_1 \cdot g_1$ is given by

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

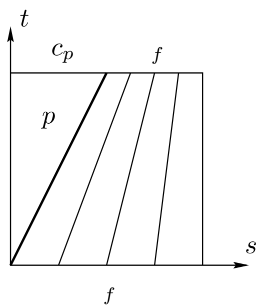


FIGURE 7.4. $f \sim c_p \cdot f$.

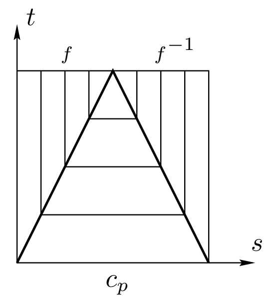


FIGURE 7.5. $c_p \sim f \cdot f^{-1}$.

Again, this is continuous by the gluing lemma. □

With this result, it makes sense to define multiplication of path classes by setting $[f] \cdot [g] = [f \cdot g]$ whenever $f \cdot g$ is defined. In particular, it is always defined for $[f], [g] \in \pi_1(X, q)$. We wish to show that $\pi_1(X, q)$ is a group under this multiplication, which amounts to proving associativity of path class multiplication and the existence of an identity and inverses. Again, it will be useful to prove these properties in a slightly more general setting, for paths that do not necessarily have the same initial and terminal points. For any path f, we define the *reverse path* f^{-1} by $f^{-1}(s) = f(1-s)$; this just retraces f from its terminal point to its initial point. Recall that c_q denotes the constant loop at q.

Theorem 7.9 (Properties of Path Multiplication). *Let f be any path from p to q in a space X, and let g and h be any paths in X. Path multiplication satisfies the following properties:*

(*a*) $[c_p] \cdot [f] = [f] \cdot [c_q] = [f]$.

(*b*) $[f] \cdot [f^{-1}] = [c_p]$; $[f^{-1}] \cdot [f] = [c_q]$.

(*c*) $[f] \cdot ([g] \cdot [h]) = ([f] \cdot [g]) \cdot [h]$ *whenever either side is defined.*

Proof. For (a), let us show that $c_p \cdot f \sim f$; the product the other way works similarly. Define $H\colon I \times I \to X$ (Figure 7.4) by

$$H(s,t) = \begin{cases} p, & t \geq 2s; \\ f\left(\dfrac{2s-t}{2-t}\right), & t \leq 2s. \end{cases}$$

Geometrically, this maps the portion of the square on the left of the line $t = 2s$ to the point p, while it maps the portion on the right along the path f at increasing speeds as t goes from 0 to 1. (The slanted lines in the picture are the level sets of H, i.e., the lines along which H takes the same value.) This map is continuous by the gluing lemma, and you can check that $H(s,0) = f(s)$ and $H(s,1) = c_p \cdot f(s)$. Thus $H\colon f \sim c_p \cdot f$.

For (b), we will just show that $f \cdot f^{-1} \sim c_p$. Since $(f^{-1})^{-1} = f$, the other relation follows by interchanging the roles of f and f^{-1}. Define a homotopy $H\colon c_p \sim f \cdot f^{-1}$ by the following recipe (Figure 7.5): At any time t, the path H_t follows f as far as $f(t)$ at double speed while the parameter s is in the interval $[0, t/2]$; then for $s \in [t/2, 1 - t/2]$ it stays at $f(t)$; then it retraces f at double speed back to q. Formally,

$$H(s,t) = \begin{cases} f(2s), & 0 \le s \le t/2; \\ f(t), & t/2 \le s \le 1 - t/2; \\ f(2-2s) & 1 - t/2 \le s \le 1. \end{cases}$$

It is easy to check that H is a homotopy from c_p to $f \cdot f^{-1}$.

Finally, to prove associativity, we need to show that $(f \cdot g) \cdot h \sim f \cdot (g \cdot h)$. The first path follows f and then g at quadruple speed for $s \in [0, \frac{1}{2}]$, and then follows h at double speed for $s \in [\frac{1}{2}, 1]$, while the second follows f at double speed and then g and h at quadruple speed. The two paths are therefore reparametrizations of each other and thus homotopic. □

Corollary 7.10. *For any space X and any point $q \in X$, $\pi_1(X, q)$ is a group.*

Note that path multiplication is not associative on the level of paths, only on the level of path homotopy classes. For definiteness, let us agree to interpret products of more than two paths as being grouped from left to right if no parentheses are present, so that $f \cdot g \cdot h$ means $(f \cdot g) \cdot h$.

The next question we need to address is how the fundamental group depends on the choice of base point. The first thing to notice is that if X is not path connected, we cannot expect the fundamental groups based at points in different path components to have any relationship to each other; $\pi_1(X, q)$ can give us information only about the path component containing q. Therefore, the fundamental group is usually used only to study path connected spaces. When X is path connected, it turns out that the fundamental groups at different points are all isomorphic; the next theorem gives an explicit isomorphism between them.

Theorem 7.11 (Change of Base Point). *Suppose X is path connected, $p, q \in X$, and g is any path from p to q. The map $\Phi_g\colon \pi_1(X,p) \to \pi_1(X,q)$ defined by*

$$\Phi_g[f] = [g^{-1}] \cdot [f] \cdot [g]$$

is an isomorphism.

Proof. Before we begin, we should verify that Φ_g makes sense (Figure 7.6):

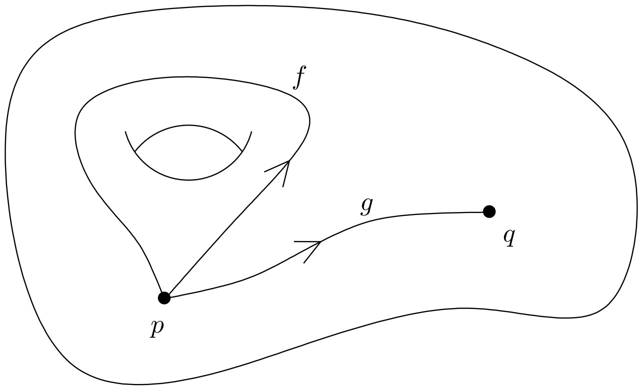


FIGURE 7.6. Change of base point.

Since g goes from p to q and f goes from p to p, paths in the class $[g^{-1}] \cdot [f] \cdot [g]$ go from q to p (by g^{-1}), then from p to p (by f), and then from p back to q (by g), so $\Phi_g(f)$ does indeed define an element of $\pi_1(X, q)$.

To check that Φ_g is a group homomorphism, use Theorem 7.9:

$$\begin{aligned}
&\Phi_g[f_1] \cdot \Phi_g[f_2] \\
&\qquad = [g^{-1}] \cdot [f_1] \cdot [g] \cdot [g^{-1}] \cdot [f_2] \cdot [g] \\
&\qquad = [g^{-1}] \cdot [f_1] \cdot [c_p] \cdot [f_2] \cdot [g] \\
&\qquad = [g^{-1}] \cdot [f_1] \cdot [f_2] \cdot [g] \\
&\qquad = \Phi_g([f_1] \cdot [f_2]).
\end{aligned}$$

(This is one reason why we needed to prove the properties of Theorem 7.9 for paths that start and end at different points.)

Finally, the fact that Φ_g is an isomorphism follows easily from the fact that it has an inverse, given by $\Phi_{g^{-1}} : \pi_1(X, q) \to \pi_1(X, p)$. □

Because of this theorem, when X is path connected we sometimes use the imprecise notation $\pi_1(X)$ to refer to the fundamental group of X with respect to an unspecified base point, if the base point is irrelevant. For example, we might say "$\pi_1(X)$ is trivial" if $\pi_1(X, q) = \{[c_q]\}$ for any $q \in X$; or we might say "$\pi_1(X) \cong \mathbb{Z}$" if there exists an isomorphism $\pi_1(X, q) \to \mathbb{Z}$ for some (hence any) q. However, we cannot dispense with the base point altogether: Since different paths from p to q may give rise to different isomorphisms, when we need to refer to a specific element of the fundamental group, or to a specific homomorphism between fundamental groups, we must be careful to specify all base points.

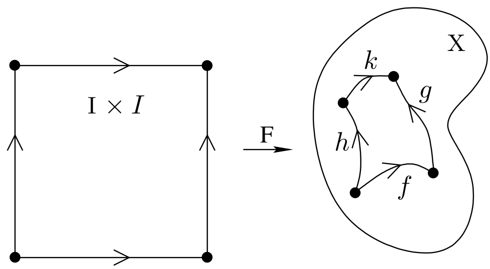


FIGURE 7.7. The map of Lemma 7.12.

If X is path connected and $\pi_1(X)$ is trivial, we say that X is *simply connected*. This means that every loop in X can be continuously shrunk to a constant loop while its endpoints are kept fixed.

Exercise 7.2. Let X be a topological space.

(a) Let $f, g\colon I \to X$ be two paths from p to q. Show that $f \sim g$ if and only if $f \cdot g^{-1} \sim c_p$.

(b) Show that X is simply connected if and only if any two paths in X with the same initial and terminal points are path homotopic.

Exercise 7.3. Show that any convex subset of $\mathbb{R}^n$ is simply connected.

In particular, Exercise 7.3 shows that the plane is simply connected. We will see later that the punctured plane is not, thus proving that the two spaces are not homeomorphic. In fact, we will show that both $\mathbb{R}^2 \smallsetminus \{0\}$ and $\mathbb{S}^1$ have infinite cyclic fundamental groups, generated by the path class of a loop that winds once around the origin. The proof will occupy most of our efforts for the remainder of this chapter and the next.

Lemma 7.12. *Let $F\colon I \times I \to X$ be a continuous map, and let f, g, h, and k be the paths in X defined by*

$$\begin{aligned} f(s) &= F(s,0);\\ g(s) &= F(1,s);\\ h(s) &= F(0,s);\\ k(s) &= F(s,1). \end{aligned}$$

(See Figure 7.7.) Then $f \cdot g \sim h \cdot k$.

Exercise 7.4. Prove Lemma 7.12.

Consider now the loop $\alpha\colon I \to \mathbb{S}^1$ given by $\alpha(s) = e^{2\pi i s}$. This map wraps the interval once around the circle counterclockwise, and maps 0 and 1 to

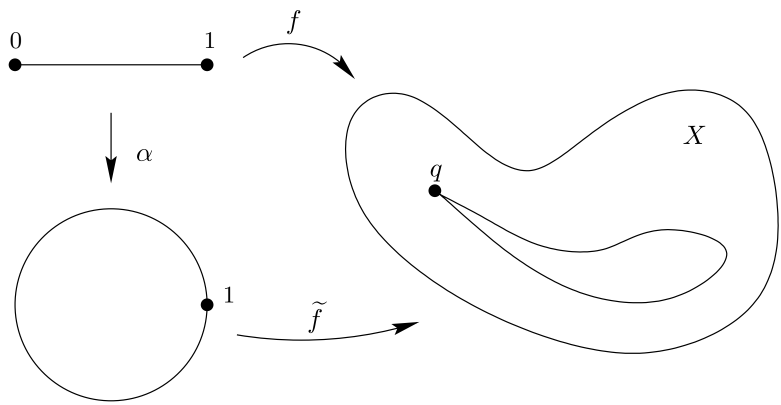


FIGURE 7.8. The circle representative of a loop.

the base point $1 \in \mathbb{S}^1$. By the closed map lemma, it is a quotient map. If $f\colon I \to X$ is any loop in a space X, it passes to the quotient to give a unique map $\widetilde{f}\colon \mathbb{S}^1 \to X$ such that $\widetilde{f} \circ \alpha = f$ (Figure 7.8), which we call the *circle representative* of f. Conversely, any continuous map $\widetilde{f}$ from the circle to X is the circle representative of the map $f = \widetilde{f} \circ \alpha$.

The next proposition gives a convenient criterion for detecting null homotopic loops in terms of their circle representatives.

Proposition 7.13. *A loop in a space X is null homotopic if and only if its circle representative extends to a map from the closed disk into X.*

Proof. Let $f\colon I \to X$ be a loop based at $q \in X$, and $\widetilde{f}\colon \mathbb{S}^1 \to X$ its circle representative. Suppose first that $\widetilde{f}$ extends to a map $F\colon \overline{\mathbb{B}^2} \to X$. Since $\overline{\mathbb{B}^2}$ is convex, the loop α is null homotopic when thought of as a loop in $\overline{\mathbb{B}^2}$. Let $A\colon I \times I \to \overline{\mathbb{B}^2}$ be a path homotopy from α to the constant loop c_1. Because $A(s,t) \in \mathbb{S}^1$ when $t = 0$ or 1, the composite map $F \circ A\colon I \times I \to X$ satisfies

$$F \circ A(s,0) = \widetilde{f} \circ A(s,0) = \widetilde{f} \circ \alpha(s) = f(s),$$

$$F \circ A(s,1) = \widetilde{f} \circ A(s,1) = \widetilde{f} \circ c_1(s) = f(1) = q,$$

and is therefore a homotopy from f to c_q.

Conversely, suppose f is null homotopic, and let $H\colon I \times I \to X$ be a homotopy from c_q to f. Observe that H sends the sides and bottom of the square to the point q. We will show below that there is a quotient map $\pi\colon I \times I \to \overline{\mathbb{B}^2}$ that sends these three sides to $1 \in \mathbb{S}^1$, makes no other identifications, and restricts to α on the top side $I \times \{1\}$. Granting this for the moment, the homotopy H passes to the quotient to give a map

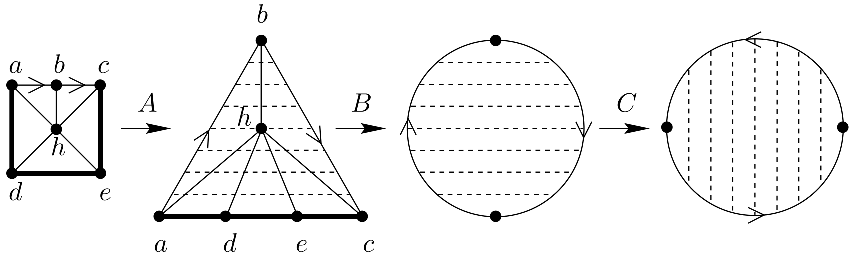


FIGURE 7.9. A quotient map sending three sides of the square to a point.

$\widetilde{H}\colon \overline{\mathbb{B}^2} \to X$ satisfying $\widetilde{H} \circ \pi = H$. The restriction of $\widetilde{H}$ to the circle is clearly the circle representative of f, so $\widetilde{H}$ is the desired extension.

We will construct the claimed quotient map $\pi\colon I \times I \to \overline{\mathbb{B}^2}$ in several steps. Let T be an equilateral triangle of altitude 2 in the plane. Triangulate $I \times I$ and T as shown in Figure 7.9, and let $A\colon I \times I \to T$ be the simplicial homeomorphism determined by the indicated vertex map. Then let $B\colon T \to \overline{\mathbb{B}^2}$ be the (nonsimplicial) map that sends each horizontal line segment in T linearly onto the horizontal segment at the same height in the disk. The resulting composite map $B \circ A$ is a quotient map by the closed map lemma, and it identifies the sides and bottom of the square to a point but makes no other identifications. It is not quite the map we are seeking, because it takes the sides and bottom to $-i$ instead of 1, and maps the top interval around the circle in the wrong direction. A suitable rotation and reflection of the disk, which we denote by C, corrects these problems. Let $\beta\colon I \to \mathbb{S}^1$ denote the restriction of $C \circ B \circ A$ to the top segment $I \times \{1\}$ of the square (identified with I). Although β is still not equal to α, the two maps differ only by a homeomorphism of $\mathbb{S}^1$ (since both α and β are quotient maps that make the same identifications). This homeomorphism extends to a homeomorphism of the closed disk by Problem 4-9, and composing with the inverse of this homeomorphism yields the desired map. □

Homomorphisms Induced by Continuous Maps

In this section we explore the effect of a continuous map on the fundamental groups of its domain and range. The first thing we need to know is that continuous maps preserve the path homotopy relation.

Lemma 7.14. *The path homotopy relation is preserved by composition with continuous maps. That is, if $f_0, f_1\colon I \to X$ are path homotopic and $\varphi\colon X \to Y$ is continuous, then $\varphi \circ f_0$ and $\varphi \circ f_1$ are path homotopic.*

Exercise 7.5. Prove Lemma 7.14

An immediate consequence of this lemma is that any continuous map $\varphi\colon X \to Y$ induces a well-defined map $\varphi_*\colon \pi_1(X,q) \to \pi_1(Y,\varphi(q))$ simply by setting $\varphi_*[f] = [\varphi \circ f]$.

Lemma 7.15. *For any continuous map φ, φ_* is a group homomorphism.*

Proof. Just note that

$$\varphi_*([f] \cdot [g]) = \varphi_*[f \cdot g] = [\varphi \circ (f \cdot g)].$$

Thus it suffices to show that $\varphi \circ (f \cdot g) = (\varphi \circ f) \cdot (\varphi \circ g)$. This is immediate, because expanding both sides using the definition of path multiplication results in identical formulas. □

The homomorphism $\varphi_*\colon \pi_1(X,q) \to \pi_1(Y,\varphi(q))$ is called the *homomorphism induced by φ*. It has the following properties:

Proposition 7.16 (Properties of the Induced Homomorphism).

(a) *Let $\varphi\colon X \to Y$ and $\psi\colon Y \to Z$ be continuous maps. Then $(\psi \circ \varphi)_* = \psi_* \circ \varphi_*$.*

(b) *If $\mathrm{Id}_X\colon X \to X$ denotes the identity map of X, then for any $q \in X$, $(\mathrm{Id}_X)_*$ is the identity map of $\pi_1(X,q)$.*

Proof. Compute:

$$\psi_*(\varphi_*[f]) = \psi_*[\varphi \circ f] = [\psi \circ \varphi \circ f] = (\psi \circ \varphi)_*[f];$$

$$(\mathrm{Id}_X)_*[f] = [\mathrm{Id}_X \circ f] = [f].$$

□

Corollary 7.17 (Topological Invariance of π_1). *Homeomorphic spaces have isomorphic fundamental groups. Specifically, if $\varphi\colon X \to Y$ is a homeomorphism, then $\varphi_*\colon \pi_1(X,q) \to \pi_1(Y,\varphi(q))$ is an isomorphism.*

Proof. If φ is a homeomorphism, then $(\varphi^{-1})_* \circ \varphi_* = (\varphi^{-1} \circ \varphi)_* = (\mathrm{Id}_X)_* = \mathrm{Id}_{\pi_1(X,q)}$, and similarly $\varphi_* \circ (\varphi^{-1})_*$ is the identity on $\pi_1(Y,\varphi(q))$. □

Be warned that injectivity or surjectivity of a continuous map does not necessarily imply that the induced homomorphism has the same property. For example, accepting for the moment the fact that $\pi_1(\mathbb{S}^1)$ is infinite cyclic (we will prove it in the next chapter), the inclusion map $\mathbb{S}^1 \hookrightarrow \mathbb{R}^2$ is injective, but its induced homomorphism is not, while the map $\varphi\colon [0,1) \to \mathbb{S}^1$ of Exercise 2.7 that wraps the interval once around the circle is surjective (in fact bijective), but its induced homeomorphism is the trivial homomorphism because $[0,1)$ is convex and therefore simply connected.

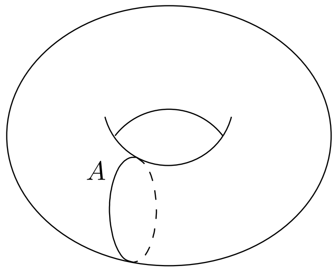


FIGURE 7.10. $\mathbb{S}^1$ is a retract of $\mathbb{T}^2$.

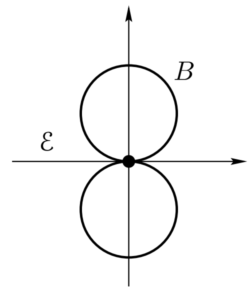


FIGURE 7.11. Figure eight space.

There is, however, one case in which the homomorphism induced by inclusion can be easily shown to be injective. Let X be a space and $A \subset X$ a subspace. A continuous map $r\colon X \to A$ is called a *retraction* if the restriction of r to A is the identity map of A, or equivalently if $r \circ \iota_A = \mathrm{Id}_A$, where $\iota_A\colon A \hookrightarrow X$ is the inclusion map. If there exists a retraction from X to A, we say that A is a *retract* of X.

Proposition 7.18. *Suppose A is a retract of X. If $r\colon X \to A$ is any retraction, then for any $q \in A$, $(\iota_A)*\colon \pi_1(A,q) \to \pi_1(X,q)$ is injective and $r_*\colon \pi_1(X,q) \to \pi_1(A,q)$ is surjective.*

Proof. Since $r \circ \iota_A = \mathrm{Id}_A$, $r_* \circ (\iota_A)*$ is the identity on $\pi_1(A,q)$, from which it follows that $(\iota_A)*$ is injective and r_* is surjective. □

Here are some examples of retractions. For these examples we will use without proof the fact that the fundamental group of the circle is infinite cyclic.

Example 7.19. It is easy to check that the map $r\colon \mathbb{R}^{n+1} \smallsetminus \{0\} \to \mathbb{S}^n$ given by $r(x) = x/|x|$ is a retraction. Thus the homomorphism $\pi_1(\mathbb{S}^n) \to \pi_1(\mathbb{R}^{n+1} \smallsetminus \{0\})$ induced by inclusion is injective. In particular, in the case $n = 1$, this means that $\pi_1(\mathbb{R}^2 \smallsetminus \{0\})$ has an infinite cyclic subgroup.

Example 7.20. The torus $\mathbb{T}^2 = \mathbb{S}^1 \times \mathbb{S}^1$ has a subspace $A = \mathbb{S}^1 \times \{1\}$ homeomorphic to $\mathbb{S}^1$ (Figure 7.10), and the map $r\colon \mathbb{T}^2 \to A$ given by $r(p,q) = (p,1)$ is easily seen to be a retraction. Thus the image of the map $(\iota_A)_*\colon \pi_1(\mathbb{S}^1) \to \pi_1(\mathbb{T}^2)$ is an infinite cyclic subgroup of $\pi_1(\mathbb{T}^2)$.

Example 7.21. Consider the *figure eight space* $\mathcal{E} \subset \mathbb{R}^2$ (Figure 7.11), which is the union of the circles of radius 1 around $(0,1)$ and $(0,-1)$. Let B denote the upper circle. There are at least two different retractions of $\mathcal{E}$ onto B—one that maps the entire lower circle to the origin and is the identity on B, and another that "folds" the lower circle onto the upper one (formally, $(x,y) \mapsto (x,|y|)$). Thus $\pi_1(\mathcal{E})$ has an infinite cyclic subgroup.

Homotopy Equivalence

Although retractions are sometimes useful tools for showing that a certain fundamental group is not trivial, it is much more useful to have a criterion under which a continuous map induces an *isomorphism* of fundamental groups. In this section we explore a very general such criterion.

Let $\varphi\colon X \to Y$ be a continuous map. We say that another continuous map $\psi\colon Y \to X$ is a *homotopy inverse* for φ if $\psi \circ \varphi \simeq \mathrm{Id}_X$ and $\varphi \circ \psi \simeq \mathrm{Id}_Y$. If there exists a homotopy inverse for φ, φ is called a *homotopy equivalence.* In this case, we say that X is *homotopy equivalent* to Y, or X has the same *homotopy type* as Y, and we write $X \simeq Y$.

Lemma 7.22. *Homotopy equivalence is an equivalence relation.*

Exercise 7.6. Prove Lemma 7.22.

One kind of homotopy equivalence is relatively easy to visualize. A subspace $A \subset X$ is said to be a *deformation retract* of X if there exists a retraction $r\colon X \to A$ such that the identity of X is homotopic to $\iota_A \circ r$. The homotopy $H\colon \mathrm{Id}_X \simeq \iota_A \circ r$ is called a *deformation retraction.* Intuitively, this means that X can be continuously deformed into A in such a way that points in A end up where they started. We say that A is a *strong deformation retract* of X if in addition $\mathrm{Id}_X \simeq_A (\iota_A \circ r)$, which means that points in A remain fixed throughout the deformation. In this case, the homotopy H can be called a *strong deformation retraction.*

Example 7.23. In Example 7.19 we showed that $\mathbb{S}^n$ is a retract of $\mathbb{R}^{n+1} \smallsetminus \{0\}$. In fact, it is a strong deformation retract: The deformation retraction is given by $H\colon (\mathbb{R}^{n+1} \smallsetminus \{0\}) \times I \to \mathbb{R}^{n+1} \smallsetminus \{0\}$, where

$$H(x,t) = (1-t)x + t\frac{x}{|x|}.$$

This is just the straight-line homotopy from the identity map to the retraction onto the sphere (Figure 7.12).

If A is a deformation retract of X, since $\iota_A \circ r \simeq \mathrm{Id}_X$ and $r \circ \iota_A = \mathrm{Id}_A$, the inclusion $A \hookrightarrow X$ is a homotopy equivalence.

Our main goal in this section is the following theorem, which is a much stronger invariance property than homeomorphism invariance, and will enable us to compute the fundamental groups of many more spaces.

Theorem 7.24 (Homotopy Invariance of $\boldsymbol{\pi_1}$). *If $\varphi\colon X \to Y$ is a homotopy equivalence, then for any point $q \in X$, $\varphi_*\colon \pi_1(X,q) \to \pi_1(Y,\varphi(q))$ is an isomorphism.*

Before proving the theorem, let us look at several important examples.

Example 7.25. Let X be any space. If the identity map of X is homotopic to a constant map, we say that X is *contractible.* Other equivalent

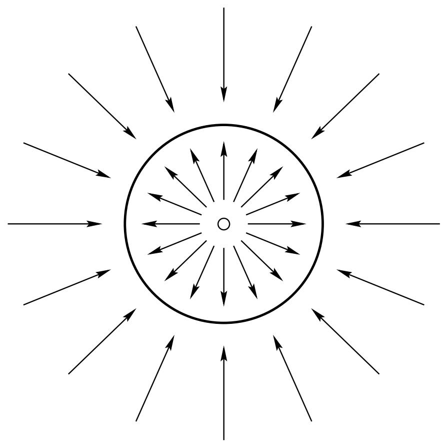


FIGURE 7.12. Strong deformation retraction of $\mathbb{R}^2 \smallsetminus \{0\}$ onto $\mathbb{S}^1$.

definitions are that any point of X is a deformation retract of X, or X is homotopy equivalent to a one-point space (Exercise 7.7). Concretely, contractibility means that there exists a continuous map $H\colon X \times I \to X$ such that

$$H(x,0) = x \text{ for all } x \in X; \qquad H(x,1) = q \text{ for all } x \in X.$$

In other words, the whole space X can be continuously shrunk to a point. Some obvious examples of contractible spaces are convex subsets of $\mathbb{R}^n$, and, more generally, any subset $B \subset \mathbb{R}^n$ that is *star-shaped*, which means that there is some point $q_0 \in B$ such that for every $q \in B$, the line segment from q_0 to q is contained in B. Since a one-point space is simply connected, it follows that every contractible space is simply connected.

Exercise 7.7. Show that the following are equivalent:

(a) X is contractible.

(b) X is homotopy equivalent to a one-point space.

(c) Any point of X is a deformation retract of X.

Example 7.26. Example 7.23 showed that the circle is a strong deformation retract of $\mathbb{R}^2 \smallsetminus \{0\}$. Therefore, inclusion $\mathbb{S}^1 \hookrightarrow \mathbb{R}^2 \smallsetminus \{0\}$ induces an isomorphism of fundamental groups. Once we show that $\pi_1(\mathbb{S}^1)$ is infinite cyclic, this will characterize $\pi_1(\mathbb{R}^2 \smallsetminus \{0\})$ as well.

Example 7.27. The figure eight space $\mathcal{E}$ of Example 7.21 and the *theta space*, defined by

$$\Theta = \{(x,y) \in \mathbb{R}^2 : x^2 + y^2 = 4, \text{ or } y = 0 \text{ and } -2 \le x \le 2\},$$

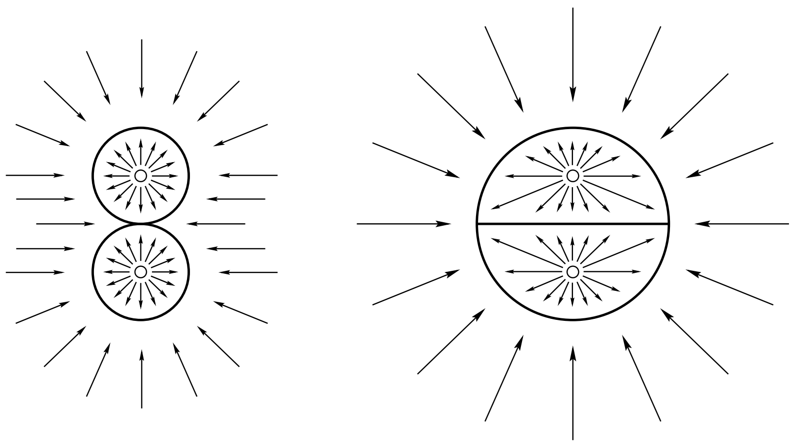


FIGURE 7.13. Deformation retractions onto $\mathcal{E}$ and Θ.

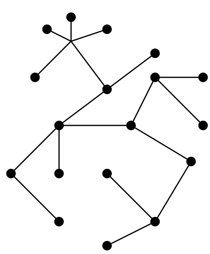


FIGURE 7.14. A tree.

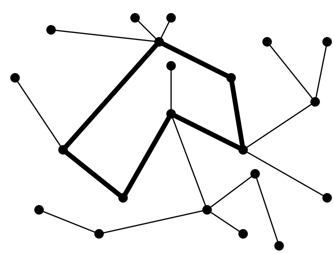


FIGURE 7.15. Not a tree.

are both strong deformation retracts of $\mathbb{R}^2$ with the two points $(0,1)$ and $(0,-1)$ removed. The deformation retractions, indicated schematically in Figure 7.13, are defined by carving the space up into regions in which straight-line homotopies are easily defined; the resulting maps are continuous by the gluing lemma. Therefore, since homotopy equivalence is transitive, $\mathcal{E}$ and Θ are homotopy equivalent to each other.

Example 7.28 (Finite Trees). Let Γ be a graph. A *cycle* in Γ is a closed finite edge path, that is, an edge path $(v_0,\dots,v_n)$ such that $v_0 = v_n$. A *tree* is a connected graph that contains no reduced cycles (see Figures 7.14 and 7.15). We will show that any finite tree T is contractible and thus simply connected.

The proof is by induction on the number of edges in T. If there is only one edge, then T is homeomorphic to a closed interval in $\mathbb{R}$, which is convex and therefore contractible. So suppose every tree with n edges is contractible, and let T be a tree with $n+1$ edges.

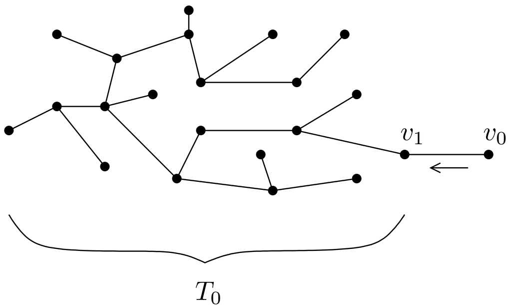


FIGURE 7.16. Proof that a tree is contractible.

If every vertex in T lies on at least two edges, then arguing as in the proof of the classification theorem for 1-manifolds (Theorem 6.1), there must be an infinite reduced edge path $\{v_i : i \in \mathbb{Z}\}$ in T. Because T is finite, there must be some integers n and $n+k > n$ such that $v_n = v_{n+k}$. If n and k are chosen so that k is the minimum positive integer with this property, this means that $(v_n, \ldots, v_{n+k})$ is a reduced cycle, contradicting the assumption that T is a tree. Thus there must be at least one vertex v_0 that lies on only one edge. Since T is connected, v_0 lies on exactly one edge, say $\langle v_0, v_1 \rangle$.

Let T_0 denote the subgraph of T with the vertex v_0 and the edge $\langle v_0, v_1 \rangle$ deleted (Figure 7.16). The straight-line homotopy defines a strong deformation retraction of $\langle v_0, v_1 \rangle$ onto v_1; extending this to be the identity on T_0 yields a strong deformation retraction of T onto T_0, which is continuous because its restriction to each simplex is continuous. Therefore, T is homotopy equivalent to T_0, which is contractible by the induction hypothesis.

Now we turn to the proof of Theorem 7.24. Roughly speaking, we would like to prove the theorem by showing that if ψ is a homotopy inverse for φ, then $\psi \circ \varphi \simeq \mathrm{Id}_X$ implies that $\psi_* \circ \varphi_*$ is the identity map, and similarly for $\varphi_* \circ \psi_*$. This would require us to show that homotopic maps induce the same fundamental group homomorphisms. However, there is an immediate problem with this approach: If two maps F_0 and F_1 are homotopic, we have no guarantee that both maps take the base point $q \in X$ to the same point in Y, so their induced homomorphisms do not even map into the same group!

The following rather complicated-looking lemma is a substitute for the claim that homotopic maps induce the same fundamental group homomorphism. It says, in effect, that homotopic maps induce the same homomorphism up to a canonical change of base point.

Lemma 7.29. *Suppose $\varphi, \psi\colon X \to Y$ are continuous, and $H\colon \varphi \simeq \psi$ is a homotopy. For any $q \in X$, let h be the path in Y from $\varphi(q)$ to $\psi(q)$ defined by $h(t) = H(q,t)$. Let $\Phi_h\colon \pi_1(Y, \varphi(q)) \to \pi_1(Y, \psi(q))$ be the isomorphism*

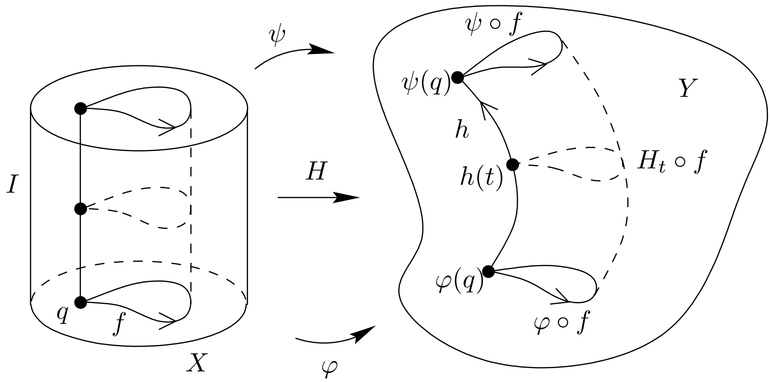


FIGURE 7.17. Induced homomorphisms of homotopic maps.

defined in Theorem 7.11. Then the following diagram commutes:

$$\begin{tikzcd}
& \pi_1(Y, \varphi(q)) \arrow[dd, "\Phi_h"] \\
\pi_1(X, q) \arrow[ur, "\varphi_*"] \arrow[dr, "\psi_*"'] & \\
& \pi_1(Y, \psi(q)).
\end{tikzcd}$$

Proof. Let f be any loop in X based at q. What we need to show is

$$\begin{aligned}
&\psi_*[f] = \Phi_h(\varphi_*[f]) \\
&\qquad \iff \varphi_*[f] = \Phi_{h^{-1}}(\psi_*[f]) \\
&\qquad \iff [\varphi \circ f] = [h] \cdot [\psi \circ f] \cdot [h^{-1}] \\
&\qquad \iff \varphi \circ f \sim h \cdot (\psi \circ f) \cdot h^{-1}.
\end{aligned}$$

At any time t, the map $H_t \circ f$ is a loop based at $h(t)$ (Figure 7.17). Thus we can define a homotopy G from $\varphi \circ f$ to $h \cdot (\psi \circ f) \cdot h^{-1}$ by letting G_t be the "lasso-shaped" loop that first follows h as far as $h(t)$, then follows $H_t \circ f$ at triple speed, then follows h back to $\varphi(q)$. Formally,

$$G(s,t) = \begin{cases} h(3ts), & 0 \le s \le \frac{1}{3}, \\ H(f(3s-1),t), & \frac{1}{3} \le s \le \frac{2}{3}, \\ h(3t(1-s)), & \frac{2}{3} \le s \le 1. \end{cases}$$

It is straightforward to check that G is continuous by the gluing lemma. The path G_0 is a reparametrization of $c_{\varphi(q)} \cdot (\varphi \circ f) \cdot c_{\varphi(q)}$, which is path homotopic to $\varphi \circ f$; and G_1 is a reparametrization of $h \cdot (\psi \circ f) \cdot h^{-1}$. $\square$

Proof of Theorem 7.24. Suppose $\varphi\colon X \to Y$ is a homotopy equivalence, and let $\psi\colon Y \to X$ be a homotopy inverse for it. Consider the sequence of maps

$$\pi_1(X,q) \xrightarrow{\varphi_*} \pi_1(Y,\varphi(q)) \xrightarrow{\psi_*} \pi_1(X,\psi(\varphi(q))) \xrightarrow{\varphi_*} \pi_1(Y,\varphi(\psi(\varphi(q)))). \quad (7.2)$$

We need to prove that the first φ_* above is bijective. As we mentioned earlier, ψ_* is not an inverse for it, because it does not map into the right space.

Since $\psi \circ \varphi \simeq \mathrm{Id}_X$, Lemma 7.29 shows that there is a path h in X such that the following diagram commutes:

$$\begin{tikzcd} & \pi_1(X,q) \arrow[dd, "\Phi_h"] \\ \pi_1(X,q) \arrow[ur, "\mathrm{Id}"] \arrow[dr, "\psi_* \circ \varphi_*"'] & \\ & \pi_1(X,\psi(\varphi(q))). \end{tikzcd} \quad (7.3)$$

Thus $\psi_* \circ \varphi_* = \Phi_h$, which is an isomorphism. In particular, this means that the first φ_* in (7.2) is injective and ψ_* is surjective.

Similarly, the homotopy $\varphi \circ \psi \simeq \mathrm{Id}_Y$ leads to the diagram

$$\begin{tikzcd} & \pi_1(Y,\varphi(q)) \arrow[dd, "\Phi_k"] \\ \pi_1(Y,\varphi(q)) \arrow[ur, "\mathrm{Id}"] \arrow[dr, "\varphi_* \circ \psi_*"'] & \\ & \pi_1(Y,\varphi(\psi(\varphi(q)))), \end{tikzcd}$$

from which it follows that $\varphi_* \circ \psi_*\colon \pi_1(Y,\varphi(q)) \to \pi_1(Y,\varphi(\psi(\varphi(q))))$ is an isomorphism. This means in particular that ψ_* is injective; since we already showed that it is surjective, it is an isomorphism. Therefore, going back to (7.3), we conclude that $\varphi_* = (\psi_*)^{-1} \circ \Phi_h\colon \pi_1(X,q) \to \pi_1(Y,\varphi(q))$ is also an isomorphism. $\square$

Homotopy Equivalence and Deformation Retraction

In Example 7.27 we showed that the figure eight and theta spaces are homotopy equivalent by showing that they are both deformation retracts

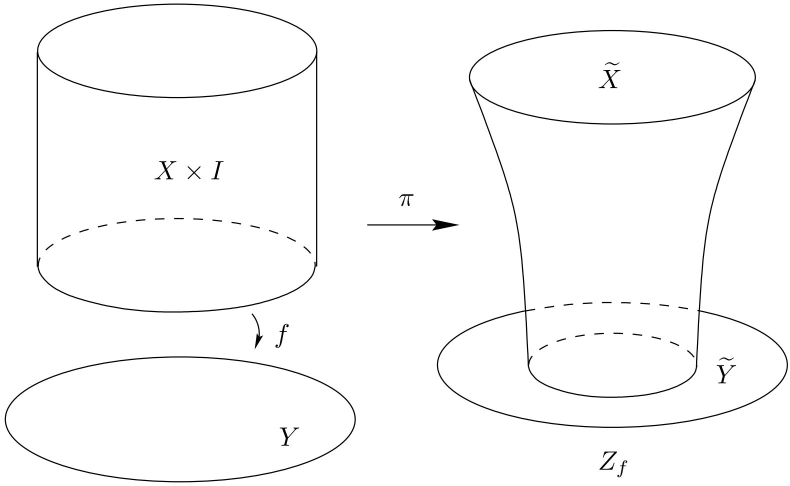


FIGURE 7.18. The mapping cylinder.

of a single larger space. This example is not as special as it might seem. As the next proposition shows, two spaces are homotopy equivalent if and only if both are homeomorphic to deformation retracts of a single larger space. This gives a rather concrete way to think about homotopy equivalence.

Let X and Y be topological spaces, and let $f\colon X \to Y$ be a continuous map. Define the *mapping cylinder* Z_f of f to be the quotient space of $(X \times I) \amalg Y$ by the equivalence relation generated by $(x,0) \sim f(x)$ for all $x \in X$. Let π denote the quotient map. The space Z_f can be visualized as a "top hat" (Figure 7.18) formed by gluing the "cylinder" $X \times I$ to Y (the "brim") by attaching each point $(x,0)$ on the bottom of the cylinder to its image $f(x)$ in Y.

The subspace $X \times \{1\} \subset (X \times I) \amalg Y$ is a saturated closed subset homeomorphic to X. The restriction of π to this subset is thus a one-to-one quotient map, so its image $\widetilde{X}$ is also homeomorphic to X. Similarly, $\widetilde{Y} = \pi(Y)$ is homeomorphic to Y.

Proposition 7.30. *With notation as above, if f is a homotopy equivalence, then $\widetilde{Y}$ and $\widetilde{X}$ are deformation retracts of Z_f. Thus two spaces are homotopy equivalent if and only if they are both homeomorphic to deformation retracts of a single space.*

Proof. For any $(x,s) \in X \times I$, let $[x,s] = \pi(x,s)$ denote its equivalence class in Z_f; similarly, $[y] = \pi(y)$ is the equivalence class of $y \in Y$.

First we show that $\widetilde{Y}$ is a strong deformation retract of Z_f, assuming only that f is continuous. We define a retraction $A\colon Z_f \to Z_f$, which collapses

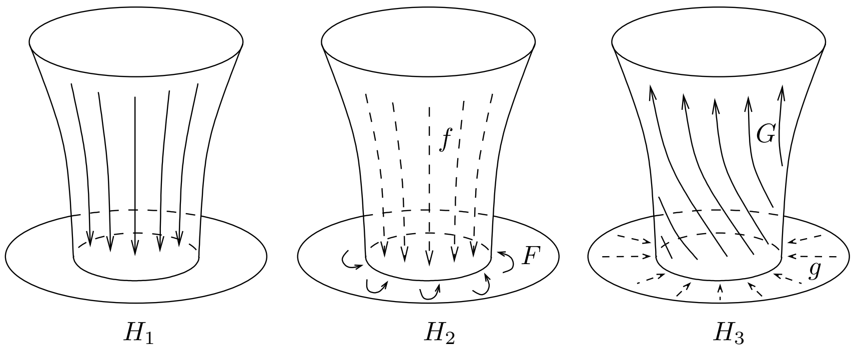


FIGURE 7.19. Homotopies of the mapping cylinder.

Z_f down onto $\widetilde{Y}$, by

$$A[x, s] = [x, 0];$$

$$A[y] = [y].$$

To be a bit more precise, we should define a map $\widetilde{A}\colon (X \times I) \amalg Y \to Z_f$ by $\widetilde{A}(x, s) = [x, 0]$ and $\widetilde{A}(y) = [y]$. This map is evidently continuous because its restrictions to $X \times I$ and Y are compositions of continuous maps. Because $\widetilde{A}(x, 0) = [f(x)] = \widetilde{A}(f(x))$, $\widetilde{A}$ respects the identifications made by π, so it passes to the quotient to yield the continuous map A defined above. This kind of standard argument will be used repeatedly to show that a map from Z_f is continuous; we will generally abbreviate it by saying something like "A is well-defined and continuous because $A[x, 0] = [f(x)] = A[f(x)]$."

Define a homotopy $H_1\colon Z_f \times I \to Z_f$ (Figure 7.19) by

$$H_1([x, s], t) = [x, s(1 - t)];$$

$$H_1([y], t) = [y].$$

Because $H_1([x, 0], t) = [x, 0] = [f(x)] = H_1([f(x)], t)$, H_1 is well-defined. To check that it is continuous, we need only observe that it respects the identifications made by the map $\pi \times \mathrm{Id}\colon ((X \times I) \amalg Y) \times I \to Z_f \times I$, which is a quotient map by Lemma 4.35. Since $H_1(\zeta, 0) = \zeta$ and $H_1(\zeta, 1) = A(\zeta)$ for any $\zeta \in Z_f$, H_1 is a homotopy between the identity map of Z_f and A. Since, moreover, $H_1([y], t) = [y]$ for all $y \in Y$, it is in fact a strong deformation retraction.

Now suppose f is a homotopy equivalence, and let $g\colon Y \to X$ be a homotopy inverse for it. Thus there exist homotopies $F\colon Y \times I \to Y$ and $G\colon X \times I \to X$ such that $F\colon f \circ g \simeq \mathrm{Id}_Y$ and $G\colon g \circ f \simeq \mathrm{Id}_X$. Define two more homotopies H_2 and H_3 by

$$H_2([x,s],t) = [F(f(x),1-t)];$$

$$H_2([y],t) = [F(y,1-t)];$$

$$H_3([x,s],t) = [G(x,st),t];$$

$$H_3([y],t) = [g(y),t].$$

The straightforward verification that H_2 and H_3 are well-defined and continuous is left to the reader. Geometrically, H_2 deforms all of Z_f into the image of f in $\widetilde{Y}$ along the homotopy F, and then H_3 collapses Z_f onto $\widetilde{X}$ by deforming each point along the homotopy G (Figure 7.19).

Inserting $t = 0$ and $t = 1$ into the definitions of H_2 and H_3, we find that $H_2\colon A \simeq B$ and $H_3\colon B \simeq C$, where

$$B[x,s] = [g(f(x)),0];$$

$$B[y] = [g(y),0];$$

$$C[x,s] = [G(x,s),1];$$

$$C[y] = [g(y),1].$$

Because homotopy is transitive, the three homotopies H_1, H_2, H_3 yield $\mathrm{Id}_{Z_f} \simeq A \simeq B \simeq C$. Since $G(x,1) = x$, we find that $C[x,1] = [x,1]$, so C is a retraction onto $\widetilde{X}$, which shows that $\widetilde{X}$ is a deformation retract of Z_f. □

Higher Homotopy Groups

You might have wondered what the subscript 1 stands for in $\pi_1(X)$. As the notation suggests, the fundamental group is just one in a series of groups associated with a topological space, all of which measure "holes" of various dimensions. In this section we introduce the basic definitions without much detail, just so that you will recognize this construction when you see it again. We will not use this material anywhere else in the book.

The definition of the higher homotopy groups is motivated by the fact that by identifying loops with their circle representatives as described earlier in this chapter, we can regard the fundamental group $\pi_1(X,q)$ as the set of equivalence classes of maps from $\mathbb{S}^1$ into X taking 1 to q, modulo homotopy relative to the base point 1. Generalizing this, for any nonnegative integer n, we define $\pi_n(X,q)$ to be the set of equivalence classes of maps from $\mathbb{S}^n$ into X taking $(1, 0, \ldots, 0)$ to q, modulo homotopy relative to the base point. Just as in the case of the fundamental group, it can be shown that $\pi_n(X, q)$ is a topological invariant.

The simplest case is $n = 0$. Since $\mathbb{S}^0 = \{\pm 1\}$, a map from $\mathbb{S}^0$ to X sending the base point 1 to q is determined by where it sends -1. Two such maps are homotopic if and only if the two images of -1 lie in the same path component of X. Therefore, $\pi_0(X, q)$ can be identified with the set of path components of X. There is no canonical group structure on $\pi_0(X, q)$; it is merely a set with a distinguished element (the component containing q). It is conventional to define $\pi_0(X)$ to be the set of path components without any distinguished element.

For $n > 1$, $\pi_n(X, q)$ has a multiplication operator (which we will not describe here) under which it turns out to be an abelian group, called the *nth homotopy group* of X based at q. These groups measure the inequivalent ways of mapping $\mathbb{S}^n$ into X, and tell us, in a sense, about the n-dimensional "holes" in X. For example, we will see later that $\pi_1(\mathbb{R}^3 \smallsetminus \{0\})$ is trivial; but it can be shown that the inclusion $\mathbb{S}^2 \hookrightarrow \mathbb{R}^3 \smallsetminus \{0\}$ represents a nontrivial element of $\pi_2(\mathbb{R}^3 \smallsetminus \{0\})$.

The higher homotopy groups are notoriously hard to compute. In fact, only a limited amount is known about $\pi_k(\mathbb{S}^n)$ for k much larger than n. The structure and computation of these groups form the embarkation point for a vast branch of topology known as *homotopy theory*. See [Whi78] for an excellent introduction to the subject.

Categories and Functors

In this section we digress a bit to give a brief introduction to category theory, a powerful idea that unifies many of the concepts we have seen so far, and indeed much of mathematics. We will only touch on these ideas from time to time in this book, but you will use them extensively if you do more advanced work in algebraic topology, so it is important to familiarize yourself with the basic concepts.

A *category* C consists of the following:

- a class (not necessarily a set) of *objects*;
- for each pair of objects X, Y a set $\operatorname{Hom}_{\mathsf{C}}(X, Y)$ of *morphisms*; and
- for each triple X, Y, Z of objects a function called *composition*: $\operatorname{Hom}_{\mathsf{C}}(X, Y) \times \operatorname{Hom}_{\mathsf{C}}(Y, Z) \to \operatorname{Hom}_{\mathsf{C}}(X, Z)$, written $(f, g) \mapsto g \circ f$;

such that the following axioms are satisfied:

(i) Composition is associative: $(f \circ g) \circ h = f \circ (g \circ h)$.

(ii) For each object X there exists an *identity morphism* $\mathrm{Id}_X \in \mathrm{Hom}_{\mathsf{C}}(X,X)$ such that for any morphism $f \in \mathrm{Hom}_{\mathsf{C}}(X,Y)$ we have $\mathrm{Id}_Y \circ f = f = f \circ \mathrm{Id}_X$.

There are many alternative notations in use. The set $\mathrm{Hom}_{\mathsf{C}}(X,Y)$ is also sometimes denoted by $\mathsf{C}(X,Y)$ or even just $\mathrm{Hom}(X,Y)$ if the category in question is understood. An element $f \in \mathrm{Hom}_{\mathsf{C}}(X,Y)$ is often written $f\colon X \to Y$. For the most part you can think of the objects in a category as sets with some special structure and the morphisms as maps that preserve the structure, although the definitions do not require this, and we will see below that there are natural examples that are not of this type.

Here are some familiar examples of categories, which we describe by specifying their objects and morphisms; the composition laws and identity morphisms are the obvious ones.

- SET: Sets and functions.
- GROUP: Groups and group homomorphisms.
- AB: Abelian groups and group homomorphisms.
- RING: Rings and ring homomorphisms.
- CRING: Commutative rings and ring homomorphisms.
- $\mathsf{VECT}_{\mathbb{R}}$: Real vector spaces and $\mathbb{R}$-linear maps.
- $\mathsf{VECT}_{\mathbb{C}}$: Complex vector spaces and $\mathbb{C}$-linear maps.
- TOP: Topological spaces and continuous maps.
- TOP_*: *Pointed topological spaces*—topological spaces together with a choice of base point in each—and base-point-preserving continuous maps.
- SIMP: Abstract simplicial complexes and simplicial maps.

In each case, the verification of the axioms of a category is straightforward. The main point is to show that a composition of the appropriate structure-preserving maps again preserves the structure. Associativity is automatic because it holds for composition of maps.

In any category C, a morphism $f \in \mathrm{Hom}_{\mathsf{C}}(X,Y)$ is called an *isomorphism* if there exists a morphism $g \in \mathrm{Hom}_{\mathsf{C}}(Y,X)$ such that $f \circ g = \mathrm{Id}_Y$ and $g \circ f = \mathrm{Id}_X$. For example, in SET, the isomorphisms are just the bijections; in GROUP they are the group isomorphisms; and in TOP they are the homeomorphisms.

A *subcategory* of C is a category D whose objects are (some of the) objects of C and whose sets of morphisms are subsets of the morphisms in C, with the composition law and identities inherited from C. A *full subcategory* is one in which $\operatorname{Hom}_{\mathsf{D}}(X,Y) = \operatorname{Hom}_{\mathsf{C}}(X,Y)$ whenever X, Y are objects of D. For example, AB is a full subcategory of GROUP.

The real power of category theory becomes apparent when we consider relations between categories. Suppose C and D are categories. A *covariant functor* $\mathcal{F}$ from C to D is a rule that assigns to each object X in C an object $\mathcal{F}(X)$ in D, and to each morphism $f \in \operatorname{Hom}_{\mathsf{C}}(X,Y)$ an *induced morphism* $\mathcal{F}(f) \in \operatorname{Hom}_{\mathsf{D}}(\mathcal{F}(X), \mathcal{F}(Y))$, in such a way that composition and identities are preserved:

$$\mathcal{F}(g \circ h) = \mathcal{F}(g) \circ \mathcal{F}(h); \qquad \mathcal{F}(\mathrm{Id}_X) = \mathrm{Id}_{\mathcal{F}(X)}\,.$$

In many cases, if the functor is understood, it is traditional to write the induced morphism $\mathcal{F}(g)$ as g_*.

It is also frequently useful to consider *contravariant functors*, which are defined in exactly the same way as covariant functors, except that the induced morphisms go in the reverse direction: If $g\colon X \to Y$, then $\mathcal{F}(g)\colon \mathcal{F}(Y) \to \mathcal{F}(X)$; and the composition law becomes

$$\mathcal{F}(g \circ h) = \mathcal{F}(h) \circ \mathcal{F}(g).$$

It is common for the morphism $\mathcal{F}(f)$ induced by a contravariant functor $\mathcal{F}$ to be written f^* if the functor is understood. (Note the upper star: The use of a lower star to denote a covariant induced morphism and an upper star to denote a contravariant one is universal.)

Here are some important examples of functors.

Example 7.31 (Covariant Functors).

- The *fundamental group functor* $\pi_1\colon \mathsf{TOP}_* \to \mathsf{GROUP}$ assigns to each pointed topological space (X, q) its fundamental group based at q, and to each base-point-preserving continuous map its induced homomorphism. The fact that it is a covariant functor is the content of Proposition 7.16.

- The functor $\pi_0\colon \mathsf{TOP} \to \mathsf{SET}$ assigns to each topological space its set of path components. For any continuous map $f\colon X \to Y$, the induced map $f_*\colon \pi_0(X) \to \pi_0(Y)$ just takes a path component X_0 of X to the path component of Y containing $f(X_0)$.

- The *forgetful functor* $\mathcal{F}\colon \mathsf{TOP} \to \mathsf{SET}$ just assigns to each topological space its underlying set, and to each continuous map its underlying set map. In fact, such a functor exists for any category whose objects are sets with some extra structure and whose morphisms are structure-preserving maps.

- The *geometric realization functor* from SIMP to TOP takes an abstract complex $\mathcal{K}$ to its geometric realization, and a simplicial map

$f\colon \mathcal{K} \to \mathcal{L}$ to the continuous map $|f|\colon |\mathcal{K}| \to |\mathcal{L}|$. It is a functor because of Lemma 5.5.

Example 7.32 (Contravariant Functors).

- The *dual space functor* from $\mathsf{VECT}_{\mathbb{R}}$ to itself assigns to each vector space V its *dual space* V^* (the vector space of linear maps $V \to \mathbb{R}$), and to each linear map $\varphi\colon V \to W$ the *dual map* or *transpose* $\varphi^*\colon W^* \to V^*$ defined by $\varphi^*(f)x = f(\varphi x)$. The verification of the functorial properties can be found in most linear algebra texts.

- The functor $\mathcal{C}\colon \mathsf{TOP} \to \mathsf{CRING}$ assigns to each topological space X its commutative ring $\mathcal{C}(X)$ of continuous real-valued functions $f\colon X \to \mathbb{R}$. For any continuous map $\varphi\colon X \to Y$, the induced map $\varphi^*\colon \mathcal{C}(Y) \to \mathcal{C}(X)$ is given by $\varphi^*(f) = f \circ \varphi$.

- If X and Z are abelian groups, the set $\operatorname{Hom}(X, Z)$ of group homomorphisms is also an abelian group under pointwise addition. We get a contravariant functor from AB to itself by sending each group X to the group $\operatorname{Hom}(X, Z)$, and each homomorphism $f\colon X \to Y$ to the dual homomorphism $f^*\colon \operatorname{Hom}(Y, Z) \to \operatorname{Hom}(X, Z)$ defined by $f^*(g) = g \circ f$.

An immediate consequence of the definitions is that any (covariant or contravariant) functor from C to D takes isomorphisms in C to isomorphisms in D: The proof is exactly the same as the proof for the fundamental group functor (Corollary 7.17).

The examples considered so far are all categories whose objects are sets with some structure and whose morphisms are structure-preserving maps. Here are some examples that are not of this type.

Example 7.33 (Homotopy Categories).

- The *homotopy category* HTOP is the category whose objects are topological spaces as in TOP, but whose morphisms are *homotopy classes* of continuous maps. Since composition preserves the homotopy relation, this is indeed a category. The isomorphisms in this category are the (homotopy classes of) homotopy equivalences.

- A closely related category is the *pointed homotopy category* HTOP_*, which has the same objects as TOP_* but whose morphisms are the equivalence classes of continuous maps modulo homotopy relative to the base point. One consequence of the homotopy invariance of the fundamental group is that π_1 defines a functor from HTOP_* to GROUP.

Example 7.34 (Groups as Categories). Suppose C is a category with one object, in which every morphism is an isomorphism. If we call the object X, the entire structure of the category is contained in the set $\operatorname{Hom}_{\mathsf{C}}(X, X)$ of morphisms and its composition law. The axioms for a category say that any two morphisms can be composed to obtain a third morphism, that composition is associative, and that there is an identity morphism. The additional assumption that every morphism is an isomorphism means that each morphism has an inverse. In other words, $\operatorname{Hom}_{\mathsf{C}}(X, X)$ is a group! Functors between such categories are just group homomorphisms. In fact, every group can be identified with such a category. One way to see this is to identify a group G with the subcategory of SET consisting of the one object G and the maps $L_g\colon G \to G$ given by left translation.

Another ubiquitous and useful technique in category theory goes by the name of "universal mapping properties." These give a unified way to define common constructions that arise in many categories, such as products and sums.

Let $\{X_\alpha : \alpha \in A\}$ be any indexed collection of objects in a category C. An object P together with a set of morphisms $\pi_\alpha\colon P \to X_\alpha$ called *projections* is said to be a *product* of the objects $\{X_\alpha\}$ if given any object W in C and morphisms $f_\alpha\colon W \to X_\alpha$, there exists a unique morphism $f\colon W \to P$ such that each of the following diagrams commutes:

$$\begin{tikzcd}
& P \arrow[d, "\pi_\alpha"] \\
W \arrow[ur, "f"] \arrow[r, "f_\alpha"'] & X_\alpha.
\end{tikzcd}$$

Lemma 7.35. *If a product exists in any category, it is unique up to an isomorphism that respects the projections.*

Proof. If $(P, \{\pi_\alpha\})$ and $(P', \{\pi'_\alpha\})$ are both products of objects $\{X_\alpha\}$, the defining property guarantees the existence of maps $f\colon P \to P'$ and $f'\colon P' \to P$ satisfying $\pi'_\alpha \circ f = \pi_\alpha$ and $\pi_\alpha \circ f' = \pi'_\alpha$. Arguing exactly as in the proof that the product topology on $X_1 \times \cdots \times X_n$ is the unique one satisfying the characteristic property (Theorem 3.10), we conclude that $f \circ f'$ and $f' \circ f$ are both identity maps. $\square$

In any particular category, products may or may not exist. Here are some examples of familiar categories in which products always exist.

Example 7.36 (Categorical Products).

(a) The product of a collection of sets in the category SET is just their Cartesian product.

(b) In the category TOP of topological spaces and continuous maps, the product of finitely many spaces $X_1, \ldots, X_n$ is the space $X_1 \times \cdots \times X_n$ with the product topology. The characteristic property of the product topology guarantees that the product space satisfies the defining condition for a product. (The categorical definition of product, by the way, determines the correct definition of the product topology on a product of infinitely many spaces; see Problem 7-12.)

(c) The product of groups $\{G_\alpha : \alpha \in A\}$ in GROUP is their direct product group $\prod_\alpha G_\alpha$, with the group structure obtained by multiplying elements componentwise.

Exercise 7.8. Prove that each of the above constructions satisfies the defining property of a product in its category.

If we reverse all the morphisms in the definition of a product, we get a dual concept called a sum. A *sum* of objects $\{X_\alpha\}$ in a category C is an object S together with morphisms $\iota_\alpha \colon X_\alpha \to S$ called *injections* such that given any object W in C and morphisms $f_\alpha \colon X_\alpha \to W$, there exists a unique morphism $f \colon S \to W$ such that each of the following diagrams commutes:

$$\begin{tikzcd} S \arrow[dr, "f"] & \\ X_\alpha \arrow[u, "\iota_\alpha"] \arrow[r, "f_\alpha"'] & W. \end{tikzcd}$$

Lemma 7.37. *If a sum exists in a category, it is unique up to an isomorphism that respects the injections.*

Exercise 7.9. Prove Lemma 7.37.

Some examples of categorical sums are given in the problems.

Problems

7-1. Let $B \subset \mathbb{R}^n$ be any convex set, X any topological space, and A any subset of X. Show that any two continuous maps $f, g\colon X \to B$ that agree on A are homotopic relative to A.

7-2. Prove the following facts about retracts.

(a) A retract of a Hausdorff space is closed.

(b) A retract of a connected space is connected.

(c) A retract of a compact space is compact.

(d) A retract of a simply connected space is simply connected.

(e) A retract of a retract is a retract, i.e., if $A \subset B \subset X$, A is a retract of B, and B is a retract of X, then A is a retract of X.

7-3. Show that the Möbius band (see Chapter 5) is homotopy equivalent to $\mathbb{S}^1$.

7-4. Let X be a path connected space and $p, q \in X$. Determine an algebraic necessary and sufficient condition on the fundamental group of X under which all path classes from p to q give the same isomorphism of $\pi_1(X, p)$ with $\pi_1(X, q)$.

7-5. For any compact surface S, show that S with one point removed is homotopy equivalent to a bouquet of finitely many circles.

7-6. Let $\mathcal{K}$ be a simplicial complex and $\sigma \in \mathcal{K}$ any simplex. Show that $|\sigma|$ has a neighborhood $U \subset |\mathcal{K}|$ such that $|\sigma|$ is a strong deformation retract of U. [Hint: Let $U = |\mathcal{K}| \smallsetminus \bigcup\{|\tau| : \tau \cap \sigma = \varnothing\}$. For any k-simplex $\tau = \langle v_0, \dots, v_k\rangle$ whose intersection with σ is nonempty, define $\pi\colon \tau \cap U \to \tau \cap U$ by

$$\pi\left(\sum_{i=0}^{k} t_i v_i\right) = \left(\sum_{i=0}^{l} t_i\right)^{-1}\left(\sum_{i=0}^{l} t_i v_i\right),$$

where $v_0, \dots, v_l$ are the vertices of $\tau \cap \sigma$, and let $H\colon (\tau \cap U) \times I \to \tau \cap U$ be the straight-line homotopy between Id and π. Show that H defines a strong deformation retraction of U onto $|\sigma|$.]

7-7. Give another proof of Lemma 7.29 by considering the map $F\colon I \times I \to Y$ defined by $F(s,t) = H(f(s), t)$ and applying Lemma 7.12.

7-8. Show that any two vertices in a tree are joined by a unique reduced edge path.

7-9. Given any collection $\{X_\alpha : \alpha \in A\}$ of topological spaces, show that their disjoint union $\coprod_\alpha X_\alpha$, together with the disjoint union topology and the natural inclusions $\iota_\alpha \colon X_\alpha \hookrightarrow \coprod_\alpha X_\alpha$, is their sum in the category TOP.

7-10. Given any collection of abelian groups $\{G_\alpha : \alpha \in A\}$, let $\bigoplus_\alpha G_\alpha$ be the subgroup of the direct product $\prod_\alpha G_\alpha$ consisting of all those elements $\{g_\alpha\}_{\alpha \in A}$ such that g_α is the identity element in G_α for all but finitely many α; and for each α let $\iota_\alpha \colon G_\alpha \hookrightarrow \bigoplus_\alpha G_\alpha$ be the obvious injection. Show that this group, called the *direct sum* of the groups G_α, is the sum of the G_α's in the category AB.

7-11. Show that the construction described in Problem 7-10 does not yield the sum in the category GROUP as follows: Take $G_1 = G_2 = \mathbb{Z}$, and find homomorphisms f_1 and f_2 from $\mathbb{Z}$ to some (necessarily non-abelian) group H such that no homomorphism $f \colon \mathbb{Z} \oplus \mathbb{Z} \to H$ makes the following diagram commute for $i = 1, 2$:

$$\begin{tikzcd}
\mathbb{Z} \oplus \mathbb{Z} \arrow[dr, "f"] & \\
\mathbb{Z} \arrow[u, "\iota_i"] \arrow[r, "f_i"'] & H.
\end{tikzcd}$$

(We will see how to construct the sum in GROUP in Chapter 9.)

7-12. Given any collection $\{X_\alpha : \alpha \in A\}$ of topological spaces, define a basis in the Cartesian product $\prod_\alpha X_\alpha$ consisting of product sets of the form $\prod_\alpha U_\alpha$, where U_α is open in X_α and $U_\alpha = X_\alpha$ for all but finitely many α. Show that this is a basis, and that $\prod_\alpha X_\alpha$ with this topology is the product of the spaces X_α in the category TOP.

7-13. Show that any vertex in a connected finite tree is a strong deformation retract of the tree.

8

Circles and Spheres

So far, we have not actually computed any nontrivial fundamental groups. The main goal of this chapter is to remedy this by computing the fundamental group of the circle. We will show, as promised, that $\pi_1(\mathbb{S}^1, 1)$ is an infinite cyclic group generated by the path α that goes once around the circle counterclockwise at constant speed. Thus each element of $\pi_1(\mathbb{S}^1, 1)$ is uniquely determined by an integer, called its "winding number," which counts the net number of times and in which direction the path winds around the circle.

Here is the essence of the plan. We would like to show that any loop in the circle is in the path class $[\alpha]^n$ for a unique integer n. The idea is to represent a path by giving its angle $\theta(s)$ as a function of the parameter, and then the winding number should be essentially $1/(2\pi)$ times the total change in angle, $\theta(1) - \theta(0)$.

Since the angle θ is not a well-defined continuous function on the circle, in order to make rigorous sense of this, we need to undertake a detailed study of the exponential quotient map $\varepsilon\colon \mathbb{R} \to \mathbb{S}^1$ defined at the end of Chapter 3. An angle function for a loop f is just (up to a constant multiple) a "lift" of f to a path in $\mathbb{R}$. Because $\mathbb{R}$ is simply connected, we can always construct a homotopy between two lifts that have the same total change in angle.

There are three key lifting lemmas that make this all work. In the beginning of the chapter we state those lifting lemmas, and then we use them to prove that the fundamental group of the circle is infinite cyclic. In the next section we prove the lifting lemmas. These three properties will make another very important appearance later in the book, when we discuss covering spaces.

At the end of the chapter we compute the fundamental groups of the higher-dimensional spheres and product spaces, and show that the fundamental group of any manifold is countable.

The Fundamental Group of the Circle

Throughout this chapter we will think of the circle as lying in the complex plane, and we will always use the base point $1 \in \mathbb{C}$, which corresponds to $(1,0) \in \mathbb{R}^2$.

Let $\alpha\colon I \to \mathbb{S}^1$ denote the loop $\alpha(s) = e^{2\pi i s}$. The complete structure of the fundamental group of the circle is described by the following theorem.

Theorem 8.1. *The group $\pi_1(\mathbb{S}^1, 1)$ is infinite cyclic, with generator $[\alpha]$.*

To prove this theorem we will use a concrete representative of the path class $[\alpha]^n$, defined as follows. For any integer n, let $\alpha_n\colon I \to \mathbb{S}^1$ be the loop $\alpha_n(s) = e^{2\pi i n s}$. It is easy to see that α_n is a reparametrization of $\alpha_{n-1} \cdot \alpha$, so by induction $[\alpha_n] = [\alpha]^n$.

As mentioned in the introduction to this chapter, the proof of the theorem is based on a close examination of the quotient map $\varepsilon\colon \mathbb{R} \to \mathbb{S}^1$ defined by $\varepsilon(x) = e^{2\pi i x}$. If $\varphi\colon B \to \mathbb{S}^1$ is any continuous map, a *lift* of φ is a continuous map $\widetilde{\varphi}\colon B \to \mathbb{R}$ such that the following diagram commutes:

$$\begin{tikzcd} & \mathbb{R} \arrow[d, "\varepsilon"] \\ B \arrow[ur, "\widetilde{\varphi}"] \arrow[r, "\varphi"'] & \mathbb{S}^1. \end{tikzcd}$$

For example, if $f\colon I \to \mathbb{S}^1$ is a path in $\mathbb{S}^1$, then a lift $\widetilde{f}$ of f can be interpreted geometrically by observing that $\theta(s) = 2\pi \widetilde{f}(s)$ is a continuous choice of angle function such that $f(s) = e^{i\theta(s)}$.

It is important to be aware that some maps may have no lifts at all. For example, suppose $\sigma\colon \mathbb{S}^1 \to \mathbb{R}$ were a lift of the identity map of $\mathbb{S}^1$:

$$\begin{tikzcd} & \mathbb{R} \arrow[d, "\varepsilon"] \\ \mathbb{S}^1 \arrow[ur, "\sigma"] \arrow[r, "\mathrm{Id}_{\mathbb{S}^1}"'] & \mathbb{S}^1. \end{tikzcd}$$

Then the equation $\varepsilon \circ \sigma = \mathrm{Id}_{\mathbb{S}^1}$ means that $2\pi\sigma$ is a continuous choice of angle function on the circle. It is intuitively evident that this cannot exist, because any choice of angle function would have to change by 2π as one goes once around the circle, and thus could not be continuous on the whole circle. Assuming for the moment the result of Theorem 8.1, we can prove rigorously that σ cannot exist just by noting that if there were such a map, the induced homomorphism $\varepsilon_* \circ \sigma_*$ would be the identity on $\pi_1(\mathbb{S}^1, 1)$, which would mean in particular that $\varepsilon_* \colon \pi_1(\mathbb{R}, 0) \to \pi_1(\mathbb{S}^1, 1)$ was surjective. Since $\pi_1(\mathbb{R}, 0)$ is the trivial group and $\pi_1(\mathbb{S}^1, 1)$ is not, this is impossible.

The first important fact about lifts is the following uniqueness lemma. This is not actually used directly in the computation of $\pi_1(\mathbb{S}^1)$, but it is necessary for proving the other two lifting properties.

Lemma 8.2 (Unique Lifting Property of the Circle). *Suppose B is connected, $\varphi\colon B \to \mathbb{S}^1$ is continuous, and $\widetilde{\varphi}_1, \widetilde{\varphi}_2 \colon B \to \mathbb{R}$ are lifts of φ that agree at some point of B. Then $\widetilde{\varphi}_1 \equiv \widetilde{\varphi}_2$.*

The next lifting lemma shows that paths in the circle, at least, always have lifts.

Lemma 8.3 (Path Lifting Property of the Circle). *Suppose $f\colon I \to \mathbb{S}^1$ is any path, and $r_0 \in \mathbb{R}$ is any point in the fiber of ε over $f(0)$. Then there exists a unique lift $\widetilde{f}\colon I \to \mathbb{R}$ of f such that $\widetilde{f}(0) = r_0$.*

Our third lifting lemma concerns lifts of homotopies: It says that lifts of path homotopic paths are path homotopic, as long as they both start at the same point.

Lemma 8.4 (Homotopy Lifting Property of the Circle). *Suppose $f_0, f_1 \colon I \to \mathbb{S}^1$ are path homotopic, and $\widetilde{f}_0, \widetilde{f}_1 \colon I \to \mathbb{R}$ are lifts of f_0 and f_1 with the same initial points. Then $\widetilde{f}_0 \sim \widetilde{f}_1$.*

Assuming these lifting lemmas, let us now carry out the proof of our theorem about the fundamental group of the circle.

Proof of Theorem 8.1. Define a map $j\colon \mathbb{Z} \to \pi_1(\mathbb{S}^1, 1)$ by $j(n) = [\alpha]^n$. It suffices to show that j is an isomorphism (considering $\mathbb{Z}$ as an additive group). Because $[\alpha]^{n+m} = [\alpha]^n[\alpha]^m$, j is a homomorphism. We will show that it is injective and surjective.

To prove surjectivity, let $[f]$ be any element of $\pi_1(\mathbb{S}^1, 1)$. By the path lifting property of the circle, f has a lift $\widetilde{f}\colon I \to \mathbb{R}$ such that $\widetilde{f}(0) = 0$ (Figure 8.1). Now, $e^{2\pi i \widetilde{f}(1)} = \varepsilon \circ \widetilde{f}(1) = f(1) = 1$, so $\widetilde{f}(1)$ is an integer n. We will show that $[f] = [\alpha_n] = j(n)$.

If we let $b_n \colon I \to \mathbb{R}$ be the path $b_n(s) = ns$, the two paths $\widetilde{f}$ and b_n both start at 0 and end at n. Because $\mathbb{R}$ is simply connected, $\widetilde{f} \sim b_n$. Since continuous maps preserve path homotopy, this implies $f = \varepsilon \circ \widetilde{f} \sim \varepsilon \circ b_n = \alpha_n$, thus proving that j is surjective.

To prove injectivity, suppose some $n \in \mathbb{Z}$ is mapped by j to the identity element $[c_1] \in \pi_1(\mathbb{S}^1, 1)$, or in other words, $[\alpha]^n = [c_1]$. Representing the

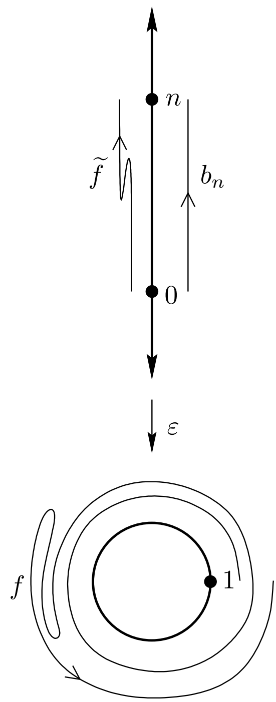


FIGURE 8.1. Proof that every path in $\mathbb{S}^1$ is homotopic to α_n.

path class $[\alpha]^n$ by α_n as above, the assumption is $\alpha_n \sim c_1$. If $\widetilde{\alpha}_n$ and $\widetilde{c}_1$ are lifts of α_n and c_1 starting at 0, the homotopy lifting property of the circle guarantees that $\widetilde{\alpha}_n \sim \widetilde{c}_1$ in $\mathbb{R}$. In particular, they both have the same terminal point. Now the lift of α_n starting at $0 \in \mathbb{R}$ is easily seen to be b_n, and the lift of c_1 is the constant loop c_0. Thus $n = b_n(1) = c_0(1) = 0$, so j is injective. □

A close examination of this proof shows that we have actually constructed an explicit inverse for j, which is of interest in its own right. Define a map $N\colon \pi_1(\mathbb{S}^1, 1) \to \mathbb{Z}$ as follows: For any $[f] \in \pi_1(\mathbb{S}^1, 1)$, let $N[f] = \widetilde{f}(1)$, where $\widetilde{f}$ is the lift of f starting at $0 \in \mathbb{R}$. Such a lift exists by the path lifting property, and $\widetilde{f}(1)$ is independent of the choice of f by the homotopy lifting property. The proof above shows that $N = j^{-1}$.

If we think of $2\pi\widetilde{f}$ as a continuous choice of angle function for f, then $2\pi N[f]$ represents, intuitively, the total change in the angle of $f(s)$ as s goes from 0 to 1, and $N[f]$ represents the number of times f winds around the circle. For this reason, $N[f]$ is called the *winding number* of the path f.

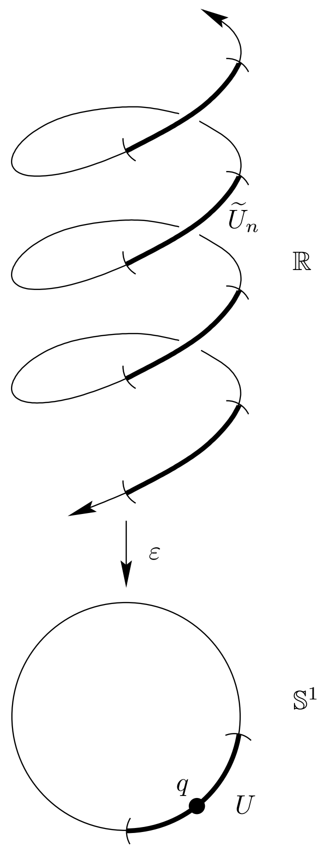


FIGURE 8.2. Evenly covered neighborhood in $\mathbb{S}^1$.

Proofs of the Lifting Lemmas

Before proving the three lifting lemmas, we need some preliminary results. The first is a precise description of the behavior of the quotient map ε.

Lemma 8.5. *Each point $q \in \mathbb{S}^1$ has a neighborhood U such that $\varepsilon^{-1}(U)$ is a disjoint union of countably many open intervals $\widetilde{U}_n$, on each of which the restriction of ε is a homeomorphism from $\widetilde{U}_n$ onto U (Figure 8.2).*

Proof. This is just a straightforward computation from the definition of ε. We can take, for example, $U = \mathbb{S}^1 \smallsetminus \{-q\}$; then the sets $\widetilde{U}_n$ are the open intervals of the form $(r+n, r+n+1)$, where r is a fixed number such that $\varepsilon(r) = -q$ and n ranges over the integers. For each n, $\varepsilon\colon \widetilde{U}_n \to U$ is a bijective open map and therefore a homeomorphism. $\square$

An open set $U \subset \mathbb{S}^1$ that has the properties described in this lemma is said to be *evenly covered.* The most important property of an evenly covered open set is that it admits local right inverses for ε, as we now describe. First, a bit of terminology. If $p\colon X \to Y$ is any surjective continuous map, a *section* of p is a continuous map $\sigma\colon Y \to X$ such that $p \circ \sigma = \mathrm{Id}_Y$ (i.e., a right inverse for p):

$$\begin{tikzcd} X \arrow[d, "p"'] \\ Y. \arrow[u, bend right, "\sigma"'] \end{tikzcd}$$

If $U \subset Y$ is an open set, a *local section* of p over U is a continuous map $\sigma\colon U \to X$ such that $p \circ \sigma = \mathrm{Id}_U$.

Lemma 8.6 (Local Section Property of the Circle). *Let $U \subset \mathbb{S}^1$ be an evenly covered open set. For any $q \in U$ and any r in the fiber of ε over q, there is a local section σ of ε over U such that $\sigma(q) = r$.*

Proof. By definition of an evenly covered open set, r is contained in some open set $\widetilde{U}_n \subset \mathbb{R}$ such that $\varepsilon\colon \widetilde{U}_n \to U$ is a homeomorphism. Thus $\sigma = (\varepsilon|_{\widetilde{U}_n})^{-1}$ is the desired local section. □

Proof of the Unique Lifting Property. Let $\mathcal{S} = \{b \in B : \widetilde{\varphi}_1(b) = \widetilde{\varphi}_2(b)\}$. By hypothesis $\mathcal{S}$ is not empty. Since B is connected, if we can show that $\mathcal{S}$ is open and closed in B, it must be all of B.

To show that $\mathcal{S}$ is open, let $b \in \mathcal{S}$. Write $r = \widetilde{\varphi}_1(b) = \widetilde{\varphi}_2(b)$ and $q = \varepsilon(r) = \varphi(b)$. Let $U \subset \mathbb{S}^1$ be an evenly covered neighborhood of q, and let $\widetilde{U}$ be the component of $\varepsilon^{-1}(U)$ containing r (Figure 8.3). If we set $V = \widetilde{\varphi}_1^{-1}(\widetilde{U}) \cap \widetilde{\varphi}_2^{-1}(\widetilde{U})$, then V is a neighborhood of b on which both $\widetilde{\varphi}_1$ and $\widetilde{\varphi}_2$ take their values in $\widetilde{U}$. Now, the fact that $\widetilde{\varphi}_1$ and $\widetilde{\varphi}_2$ are lifts of φ translates to $\varphi = \varepsilon \circ \widetilde{\varphi}_1 = \varepsilon \circ \widetilde{\varphi}_2$. Since ε is injective on $\widetilde{U}$, we conclude that $\widetilde{\varphi}_1$ and $\widetilde{\varphi}_2$ agree on V, which is to say that $V \subset \mathcal{S}$, so $\mathcal{S}$ is open.

To show that it is closed, we will show that its complement is open. Let $b \notin \mathcal{S}$, and set $r_1 = \widetilde{\varphi}_1(b)$ and $r_2 = \widetilde{\varphi}_2(b)$, so that $r_1 \neq r_2$. As above, let $q = \varepsilon(r_1) = \varepsilon(r_2) = \varphi(b)$, and let U be an evenly covered neighborhood of q. Then there are disjoint neighborhoods $\widetilde{U}_1$ of r_1 and $\widetilde{U}_2$ of r_2 such that ε is a homeomorphism from $\widetilde{U}_1$ to U and from $\widetilde{U}_2$ to U. Letting $V = \widetilde{\varphi}_1^{-1}(\widetilde{U}_1) \cap \widetilde{\varphi}_2^{-1}(\widetilde{U}_2)$, we conclude that $\widetilde{\varphi}_1(V) \subset \widetilde{U}_1$ and $\widetilde{\varphi}_2(V) \subset \widetilde{U}_2$, so $\widetilde{\varphi}_1 \neq \widetilde{\varphi}_2$ on V, which is to say that $V \cap \mathcal{S} = \varnothing$. Thus $\mathcal{S}$ is closed, and the proof is complete. □

Proof of the path lifting property. Let $f\colon I \to \mathbb{S}^1$ be a path, and $r_0 \in \varepsilon^{-1}(f(0))$ as in the statement of the lemma. If $\mathcal{U}$ is an open cover of the circle by evenly covered open sets, the collection $\{f^{-1}(U) : U \in \mathcal{U}\}$ is an open cover of I. Let δ be a Lebesgue number for this cover. Choosing an

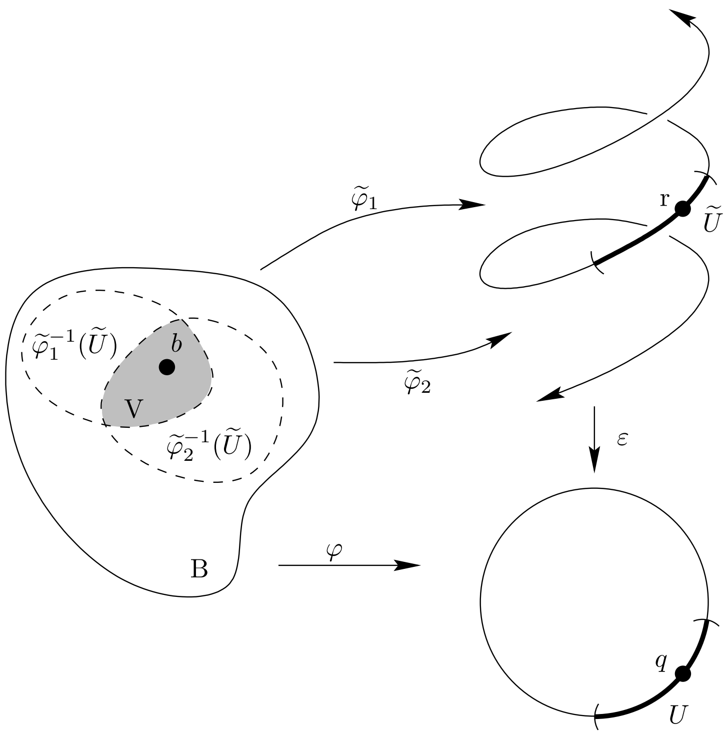


FIGURE 8.3. Proof of the unique lifting property.

integer n large enough that $1/n < \delta$, the Lebesgue number lemma says that each interval $[k/n, (k+1)/n]$ of length $1/n$ is contained in one of the sets $f^{-1}(U)$, which is to say that it is mapped by f into an evenly covered open set.

We define the lift $\widetilde{f}\colon I \to \mathbb{R}$ inductively as follows. First choose an evenly covered open set U_0 such that $f[0, 1/n] \subset U_0$. Letting $\sigma_0\colon U_0 \to \mathbb{R}$ denote the local section such that $\sigma_0(f(0)) = r_0$, we set

$$\widetilde{f} = \sigma_0 \circ f \quad \text{on } [0, 1/n].$$

It follows immediately that $\widetilde{f}$ is continuous and satisfies $\widetilde{f}(0) = r_0$ and $\varepsilon \circ \widetilde{f} = f$.

Proceeding by induction, suppose we have defined a continuous lift of f on an interval of the form $[0, k/n]$ for some integer k. As before, $f[k/n, (k+1)/n]$ is contained in an evenly covered set U_k. Letting $\sigma_k\colon U_k \to \mathbb{R}$ be the local section such that $\sigma_k(f(k/n)) = \widetilde{f}(k/n)$, we set

$$\widetilde{f} = \sigma_k \circ f \quad \text{on } [k/n, (k+1)/n].$$

The resulting map is continuous by the gluing lemma. By induction we obtain a lift $\widetilde{f}$ defined on all of I. It is unique by the unique lifting property. $\square$

Proof of the homotopy lifting property. Now suppose f_0, f_1 are path homotopic paths in the circle, and $\widetilde{f}_0, \widetilde{f}_1$ are any lifts of them starting at the same point $r_0 \in \mathbb{R}$. Let $H\colon f_0 \sim f_1$ be a path homotopy. This means that $H\colon I \times I \to \mathbb{S}^1$ satisfies

$$\begin{aligned} H(s,0) &= f_0(s); \\ H(s,1) &= f_1(s); \\ H(0,t) &= f_0(0) = f_1(0); \\ H(1,t) &= f_0(1) = f_1(1). \end{aligned}$$

We will show below that there exists a lift of H to a map $\widetilde{H}\colon I \times I \to \mathbb{R}$ such that $\widetilde{H}(0,0) = r_0$. Assuming this, we argue as follows. First, note that $\varepsilon \circ \widetilde{H}(0,t) = H(0,t) = f_0(0)$. Therefore, $t \mapsto \widetilde{H}(0,t)$ is a lift of the constant loop at $f_0(0)$ to a path starting at r_0; by the unique lifting property, it must be the constant loop at r_0, so $\widetilde{H}(0,t) = r_0$ for all t. The same argument shows that $\widetilde{H}(1,t)$ is constant for all t, so $\widetilde{H}$ is a path homotopy. Moreover, $\widetilde{H}_0(s) = \widetilde{H}(s,0)$ is a lift of f_0 starting at r_0, and similarly, $\widetilde{H}_1$ is a lift of f_1 starting at r_0. By the unique lifting property, these must be equal to the given lifts $\widetilde{f}_0$ and $\widetilde{f}_1$, respectively, and $\widetilde{H}$ provides a path homotopy between them.

All that remains is to prove the existence of the lift $\widetilde{H}$. As in the proof of the path lifting property, there exists $\delta > 0$ such that any subset of $I \times I$ whose diameter is less than δ is mapped by H into an evenly covered subset of $\mathbb{S}^1$. Choose n large enough that each square of side $1/n$ has diameter less than δ. (Any $n > \sqrt{2}/\delta$ will do.)

For any integers i, j such that $0 \le i, j \le n-1$, let S_{ij} denote the square $[i/n, (i+1)/n] \times [j/n, (j+1)/n]$ (Figure 8.4). For any point $x \in \mathbb{R}$ in the fiber of ε over $H(i/n, j/n)$, there exists a unique lift $\widetilde{H}_{ij}$ of H over S_{ij} satisfying $\widetilde{H}_{ij}(i/n, j/n) = x$, given by $\widetilde{H}_{ij} = \sigma \circ H_{ij}$, where σ is the local section of ε such that $\sigma(H(i/n, j/n)) = x$.

We define $\widetilde{H}$ inductively on $I \times I$ as follows. On S_{00}, let $\widetilde{H}_{00}$ be the lift of H such that $\widetilde{H}(0,0) = r_0$. On the next square to the right, S_{10}, let $\widetilde{H}_{10}$ be the lift such that $\widetilde{H}_{10}(1/n, 0) = \widetilde{H}_{00}(1/n, 0)$. We now have two lifts defined on the line segment $\{(1/n, t) : 0 \le t \le 1/n\}$ where the two squares overlap. But on this line segment, the paths $t \mapsto \widetilde{H}_{00}(1/n, t)$ and $t \mapsto \widetilde{H}_{10}(1/n, t)$ are both lifts of the path $t \mapsto H(1/n, t)$ starting at the same point; thus by the unique lifting property they are equal.

Continuing in this way, we define lifts on each of the squares S_{i0}, $i = 0, \dots, n-1$, and then on the squares of the second row, and so on. Suppose by induction that we have defined lifts $\widetilde{H}_{i'j'}$ on all squares $S_{i'j'}$ for $j' < j$,

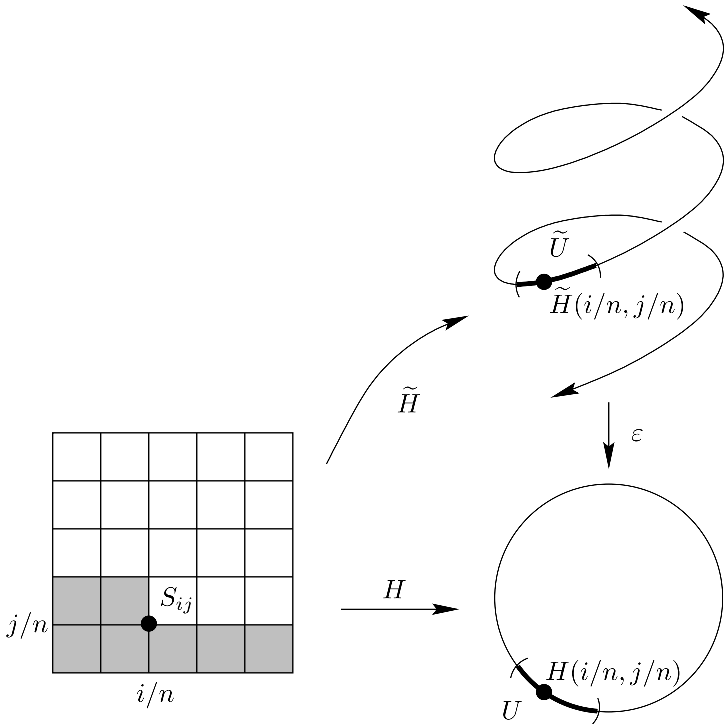


FIGURE 8.4. Proof of the homotopy lifting property (inductive step).

and for $j' = j$ and $i' < i$, and all such lifts agree where they overlap. We let $\widetilde{H}_{ij}$ be the unique lift of H on S_{ij} that agrees with any (hence all) previous lifts at the lower left corner $(i/n, j/n)$. At a typical such square, we have to check that the new lift agrees with two different old ones: one coming from the square $S_{i-1,j}$ to the left, and one coming from the square $S_{i,j-1}$ below. Just as in the preceding paragraph, the unique lifting property guarantees that the old and new lifts agree on both of these line segments.

In the end we obtain the desired lift $\widetilde{H}$ by letting $\widetilde{H} = \widetilde{H}_{ij}$ on S_{ij}; it is continuous by the gluing lemma. □

Fundamental Groups of Spheres

The situation is much simpler for the higher-dimensional spheres. The sphere minus the north pole is homeomorphic to $\mathbb{R}^n$ by stereographic projection (see Example 3.6). In fact, composing stereographic projection with a suitable rotation of the sphere, it is easy to see that the sphere minus *any* point is homeomorphic to $\mathbb{R}^n$. Therefore, if we knew that any loop in

$\mathbb{S}^n$ omitted at least one point in the sphere, we could consider it as a loop in $\mathbb{R}^n$; since it is null homotopic there, it is null homotopic in $\mathbb{S}^n$.

Unfortunately, an arbitrary loop might not omit any points. For example, there is a continuous surjective map $f\colon I \to I \times I$ (a "space-filling curve"—see, e.g., [Rud76]). Composing this with a surjective map $I \times I \to \mathbb{S}^2$ such as the one constructed in Proposition 6.2(b) yields a path whose image is all of $\mathbb{S}^2$. But as the proof of the next proposition shows, we can modify any curve by a homotopy so that it does miss a point.

Theorem 8.7. *For $n \geq 2$, $\mathbb{S}^n$ is simply connected.*

Proof. Let $N = (0, \ldots, 0, 1)$ denote the north pole, and $S = -N$ the south pole. Both the open sets $U = \mathbb{S}^n \smallsetminus \{N\}$ and $V = \mathbb{S}^n \smallsetminus \{S\}$ are homeomorphic to $\mathbb{R}^n$. If $f\colon I \to S^n$ is any path, by the Lebesgue number lemma there is an integer n such that on each subinterval $[k/n, (k+1)/n]$, f takes its values either in U or in V. Now, $V \smallsetminus \{N\}$ is homeomorphic to $\mathbb{R}^n \smallsetminus \{0\}$, which is connected. (Here is where the dimensional restriction comes in—when $n = 1$, $\mathbb{R}^n \smallsetminus \{0\}$ is disconnected.) Thus, for each such segment that lies in V, there is another path in V with the same endpoints that misses N; since V is simply connected, these two paths are path homotopic in V and thus in $\mathbb{S}^n$. Of course, each segment that lies in U already misses N. This shows that f is homotopic to a path in $\mathbb{S}^n \smallsetminus \{N\} \approx \mathbb{R}^n$, so f is null homotopic. □

Corollary 8.8. *For $n \geq 3$, $\mathbb{R}^n \smallsetminus \{0\}$ is simply connected.*

Proof. The map $F\colon \mathbb{R}^n \smallsetminus \{0\} \to \mathbb{S}^{n-1}$ given by $F(x) = x/|x|$ is a strong deformation retraction (by a straight-line homotopy). □

Fundamental Groups of Product Spaces

In this section we will show how to compute the fundamental group of an arbitrary (finite) product of topological spaces in terms of the fundamental groups of the factors. As an application, we will compute the fundamental groups of tori.

Let $X_1, \ldots, X_n$ be topological spaces, and let $p_i\colon X_1 \times \cdots \times X_n \to X_i$ denote projection on the ith factor. (We are avoiding our usual notation π_i for the projections here so as not to create confusion with the notation π_1 for the fundamental group.) Choosing base points $q_i \in X_i$, we get maps

$$p_{i*}\colon \pi_1(X_1 \times \cdots \times X_n, (q_1, \ldots, q_n)) \to \pi_1(X_i, q_i).$$

Putting these together, we define a map

$$\widetilde{p}\colon \pi_1(X_1 \times \cdots \times X_n, (q_1, \ldots, q_n)) \to \pi_1(X_1, q_1) \times \cdots \times \pi_1(X_n, q_n)$$

by

$$\widetilde{p}[f] = (p_{1*}[f], \ldots, p_{n*}[f]). \tag{8.1}$$

Proposition 8.9 (Fundamental Group of a Product). *If X_1, ..., X_n are any topological spaces, the map $\widetilde{p}\colon \pi_1(X_1 \times \cdots \times X_n, (q_1, \ldots, q_n)) \to \pi_1(X_1, q_1) \times \cdots \times \pi_1(X_n, q_n)$ defined by* (8.1) *is an isomorphism.*

Proof. First we will show that $\widetilde{p}$ is surjective. Let $[f_i] \in \pi_1(X_i, q_i)$ be arbitrary for $i = 1, \ldots, n$. Define a loop f in the product space by $f(s) = (f_1(s), \ldots, f_n(s))$. Since the component functions of f satisfy $f_i = p_i \circ f$, we compute $\widetilde{p}[f] = (p_{1*}[f], \ldots, p_{n*}[f]) = ([p_1 \circ f], \ldots, [p_n \circ f]) = ([f_1], \ldots, [f_n])$.

To show injectivity, suppose f is a loop in the product space, and $\widetilde{p}[f]$ is the identity element of $\pi_1(X_1, q_1) \times \cdots \times \pi_1(X_n, q_n)$. Writing f in terms of its component functions as $f(s) = (f_1(s), \ldots, f_n(s))$, the hypothesis means that $[c_{q_i}] = p_{i*}[f] = [p_i \circ f] = [f_i]$ for each i. If we choose homotopies $H_i\colon f_i \sim c_{q_i}$, it follows easily that the map $H\colon X_1 \times \cdots \times X_n \times I \to X_1 \times \cdots \times X_n$ given by

$$H(x_1, \ldots, x_n, t) = (H_1(x_1, t), \ldots, H_n(x_n, t))$$

is a homotopy from f to the constant loop $c_{(q_1, \ldots, q_n)}$. □

Corollary 8.10 (Fundamental Groups of Tori). *Let $\mathbb{T}^n = \mathbb{S}^1 \times \cdots \times \mathbb{S}^1$ be the n-dimensional torus, and let α_i denote the standard loop in the ith copy of $\mathbb{S}^1$:*

$$\alpha_i(s) = (1, \ldots, 1, e^{2\pi i s}, 1, \ldots, 1).$$

Using $q = (1, \ldots, 1)$ as base point, the map $\varphi\colon \mathbb{Z}^n \to \pi_1(\mathbb{T}^n, q)$ given by $\varphi(k_1, \ldots, k_n) = [\alpha_1]^{k_1} \cdots [\alpha_n]^{k_n}$ is an isomorphism.

Fundamental Groups of Manifolds

We conclude this chapter by proving an important theorem about fundamental groups of manifolds. This does not have to do with circles or spheres per se, but it does use techniques similar to those used in the other proofs in this chapter, so this is a convenient time to insert it.

Theorem 8.11. *The fundamental group of a manifold is countable.*

Proof. Let M be a manifold, and let $\mathcal{U}$ be a countable cover of M by Euclidean balls. For each $U, U' \in \mathcal{U}$ the intersection $U \cap U'$ has at most countably many components; choose a point in each such component and

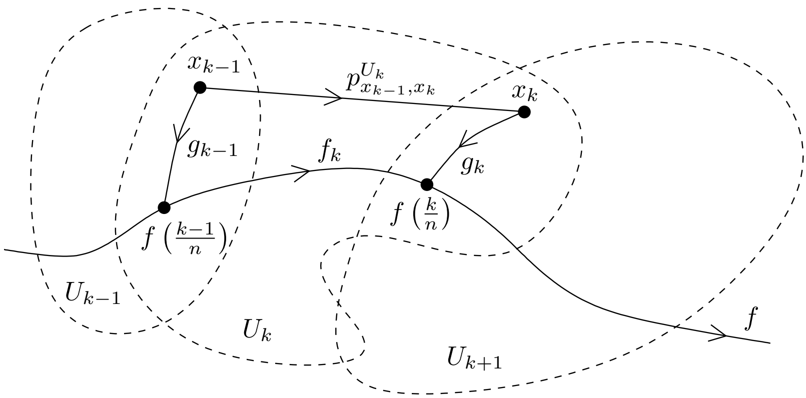


FIGURE 8.5. Proof that a manifold has countable fundamental group.

let $\mathcal{X}$ denote the (countable) set consisting of all the chosen points as U, U' range over all the sets in $\mathcal{U}$. For each $U \in \mathcal{U}$ and $x, x' \in \mathcal{X}$ such that $x, x' \in U$, choose a definite path $p^U_{x,x'}$ from x to x' in U.

Now choose any point $q \in \mathcal{X}$ as base point. Let us say that a loop based at q is *special* if it is a finite product of paths of the form $p^U_{x,x'}$. Because both $\mathcal{U}$ and $\mathcal{X}$ are countable sets, there are only countably many special loops. Each special loop determines an element of $\pi_1(M, q)$. If we can show that every element of $\pi_1(M, q)$ is obtained in this way, we will be done, because we will have exhibited a surjective map from a countable set onto $\pi_1(M, q)$.

So suppose f is any loop based at q. By the Lebesgue number lemma there is an integer n such that f maps each subinterval $[(k-1)/n, k/n]$ into one of the balls in $\mathcal{U}$; call this ball U_k. Let $f_k = f|_{[(k-1)/n,k/n]}$, reparametrized on the unit interval, so that $[f] = [f_1] \cdot \cdots \cdot [f_n]$.

Since for each $k = 1, \dots, n-1$, $f(k/n) \in U_k \cap U_{k+1}$, there is some $x_k \in \mathcal{X}$ that lies in the same component of $U_k \cap U_{k+1}$ as $f(k/n)$. Choose a path g_k in $U_k \cap U_{k+1}$ from x_k to $f(k/n)$ (Figure 8.5), and set $\widetilde{f}_k = g_{k-1} \cdot f_k \cdot g_k^{-1}$ (taking $x_k = q$ and g_k to be the constant path c_q when $k = 0$ or n). It is immediate that $[f] = [\widetilde{f}_1] \cdot \cdots \cdot [\widetilde{f}_n]$, because all the g_k's cancel out. But for each k, $\widetilde{f}_k$ is a path in U_k from x_{k-1} to x_k, and since U_k is simply connected, $\widetilde{f}_k$ is path homotopic to $p^{U_k}_{x_{k-1}x_k}$. This shows that f is path homotopic to a special loop and completes the proof. $\square$

Problems

8-1. Prove that the circle is not a retract of the closed disk.

8-2. Identifying the circle with the subspace $\mathbb{S}^1 \times \{1\} \subset \mathbb{T}^2$, prove that the circle is not a deformation retract of the torus.

8-3. Prove that the fundamental group of any topological group is abelian. [Hint: If f and g are loops based at $1 \in G$, consider the map F from $I \times I$ into G given by $F(s,t) = f(s)g(t)$ and use Lemma 7.12.]

8-4. Suppose $U \subset \mathbb{R}^2$ is an open set and $x \in U$. Show that $U \smallsetminus \{x\}$ is not simply connected. [Hint: Let S be a small circle around x, and consider the sequence of inclusions $S \hookrightarrow U \smallsetminus \{x\} \hookrightarrow \mathbb{R}^2 \smallsetminus \{x\}$.]

8-5. Show that a topological space cannot be simultaneously a 2-manifold and an n-manifold for some $n > 2$. [Hint: If $n > 2$, any n-manifold has a basis of open sets in which the complement of any point is simply connected.]

8-6. Let M be a 2-dimensional manifold with boundary. Show that the set of boundary points of M is disjoint from the set of interior points. Conclude that a 2-manifold with boundary is a manifold if and only if its boundary is empty.

8-7. Let $\varphi\colon \mathbb{S}^1 \to \mathbb{S}^1$ be a continuous map such that $\varphi(1) = 1$. Because $\pi_1(\mathbb{S}^1, 1)$ is infinite cyclic, there is an integer n, called the *degree* of φ and denoted by $\deg\varphi$, such that $\varphi(\gamma) = \gamma^n$ for all $\gamma \in \pi_1(\mathbb{S}^1, 1)$. If $\varphi\colon \mathbb{S}^1 \to \mathbb{S}^1$ is an arbitrary continuous map, we define the degree of φ to be the degree of $\rho \circ \varphi$, where $\rho\colon \mathbb{S}^1 \to \mathbb{S}^1$ is the rotation $\rho(z) = z/\varphi(1)$ (in complex notation), which takes $\varphi(1)$ to 1.

(a) Show that two maps $\varphi, \psi\colon \mathbb{S}^1 \to \mathbb{S}^1$ are homotopic if and only if they have the same degree.

(b) Show that $\deg(\varphi \circ \psi) = \deg\varphi \deg\psi$ for any two continuous maps $\varphi, \psi\colon \mathbb{S}^1 \to \mathbb{S}^1$.

(c) For each $n \in \mathbb{Z}$, compute the degrees of the nth power map $p_n(z) = z^n$ and its conjugate $\overline{p}_n(z) = \overline{z}^n$.

(d) Show that $\varphi\colon \mathbb{S}^1 \to \mathbb{S}^1$ has an extension to a continuous map $\Phi\colon \overline{\mathbb{B}^2} \to \mathbb{S}^1$ if and only if it has degree zero.

8-8. Prove the *fundamental theorem of algebra:* Every complex polynomial of positive degree has a zero. [Hint: If $p(z) = z^n + a_{n-1}z^{n-1} + \cdots + a_0$, write $p_\varepsilon(z) = \varepsilon^n p(z/\varepsilon)$ and show that there exists $\varepsilon > 0$ such that $|p_\varepsilon(z) - z^n| < 1$ when $z \in \mathbb{S}^1$. If p has no zeros, prove that $p_\varepsilon|_{\mathbb{S}^1}$ is homotopic to $p_n(z) = z^n$, and use the results of Problem 8-7 to derive a contradiction.]

8-9. The *Brouwer fixed point theorem* says that any continuous map $f\colon \overline{\mathbb{B}^n} \to \overline{\mathbb{B}^n}$ has a fixed point (i.e., a point x such that $f(x) = x$). Prove this in the case $n = 2$ as follows: If $f\colon \overline{\mathbb{B}^2} \to \overline{\mathbb{B}^2}$ has no fixed point, define $\varphi\colon \overline{\mathbb{B}^2} \to \mathbb{S}^1$ by $\varphi(x) = (f(x) - x)/|f(x) - x|$. Derive a contradiction by showing that the restriction of φ to $\mathbb{S}^1$ is homotopic to the identity. [If you crumple a map of the country that you are in and drop it on the ground, this theorem guarantees that some point on the map will lie exactly over the point it represents.]

8-10. Let V be a *vector field* on $\mathbb{R}^2$, i.e., a continuous map $V\colon \mathbb{R}^2 \to \mathbb{R}^2$. A point $q \in \mathbb{R}^2$ is called a *singular point* of V if $V(q) = 0$, and a *regular point* if $V(q) \neq 0$. A singular point is *isolated* if it has a neighborhood containing no other singular points. Let $\mathcal{R}_V \subset \mathbb{R}^2$ denote the set of regular points of V. For any loop $f\colon I \to \mathcal{R}_V$, define the *index* of V with respect to f, denoted by $\operatorname{Ind}(V, f)$, to be the winding number of the loop $\widetilde{f}\colon I \to \mathbb{S}^1$, given by

$$\widetilde{f}(s) = \frac{V(f(s))}{|V(f(s))|}.$$

(a) Show that $\operatorname{Ind}(V, f)$ depends only on the path class of f.

(b) If q is an isolated singular point of V, show that $\operatorname{Ind}(V, f_\varepsilon)$ is independent of ε for ε sufficiently small, where $f_\varepsilon(s) = q + \varepsilon\alpha(s)$, and α is the standard counterclockwise loop around the unit circle. This integer is called the index of V at q, and is denoted by $\operatorname{Ind}(V, q)$.

(c) If V has finitely many singular points in the closed unit disk, all in the interior, show that the index of V with respect to the loop α around the unit circle is equal to the sum of the indices of V at the interior singular points.

(d) Compute the index of each of the following vector fields at the origin:

$$V_1(x, y) = (x, y);$$

$$V_2(x, y) = (-x, -y);$$

$$V_3(x, y) = (x + y, x - y).$$

9

Some Group Theory

In this chapter we depart from topology for a while to discuss group theory. Our goal, of course, is to use the group theory to solve topological problems, and in the next chapter we will compute the fundamental groups of all compact surfaces, and use them to show, among other things, that the different surfaces listed in the classification theorem of Chapter 6 are not homeomorphic to each other.

Before we do so, however, we need to develop some tools for constructing and describing groups. We will discuss four such tools in this chapter: free products of groups, free groups, presentations of groups by generators and relations, and free abelian groups. These will all play central roles in our computations of fundamental groups in the next chapter, and the material on free abelian groups will also be used in the discussion of homology in Chapter 13.

This chapter assumes that you are familiar with the basic facts of group theory as summarized in the Appendix. If your group theory is rusty, this would be a good time to pull out an algebra text and refresh your memory.

Free Products

There is a familiar way to create a group as a product of two or more other groups: The direct product of groups $G_1, \dots, G_n$ (see the Appendix) is the Cartesian product set $G_1 \times \cdots \times G_n$ with the group structure obtained by multiplying the entries in two n-tuples component by component.

For each i, the direct product $G_1 \times \cdots \times G_n$ has a subgroup $\{1\} \times \cdots \times \{1\} \times G_i \times \{1\} \times \cdots \times \{1\}$ isomorphic to G_i, and it is easy to verify that elements of two distinct such subgroups commute with each other. As we mentioned in Chapter 7, this construction yields the product in the category of groups.

In our study of fundamental groups, we will need to build another kind of product, in which the elements of different groups are not assumed to commute. This situation arises, for example, in computing the fundamental group of the wedge $X \vee Y$ of two spaces X and Y, defined in Example 3.25. As we will see in the next chapter, the fundamental group of $X \vee Y$ contains subgroups isomorphic to $\pi_1(X)$ and $\pi_1(Y)$, and any loop in $X \vee Y$ is equivalent to a product of loops lying in one space or the other. But in general, path classes of loops in X do not commute with those in Y.

In this section we will introduce a more complicated product of groups $G_1, \ldots, G_n$ that includes each G_i as a subgroup, but in which elements of the different subgroups do not commute with each other. It is called the "free product," and roughly speaking, it is just the set of expressions you can get by formally multiplying together elements of the different groups, with no relations assumed other than those that come from the multiplication in each group G_i. It turns out (despite its name) to be the sum in the category of groups.

Because terms such as "expressions you can get" and "multiplying elements of different groups" are too vague to use in mathematical arguments, the actual construction of the free product is rather involved. We begin with some preliminary terminology.

Let $\{G_\alpha\}_{\alpha \in \mathcal{A}}$ be an indexed collection of groups. The index set $\mathcal{A}$ can be finite or infinite; for our applications we will need only the finite case, so you are free to think of finite collections throughout this chapter. We will usually omit mention of $\mathcal{A}$ and denote the collection simply by $\{G_\alpha\}$, with Greek letters understood to range over all elements of some implicitly understood index set.

A *word* in $\{G_\alpha\}$ is a finite sequence of length $m \geq 0$ of elements of the disjoint union $\coprod_\alpha G_\alpha$. In other words, a word is an ordered m-tuple of the form $(g_1, \ldots, g_m)$, where each g_i is an element of some G_α. (Recall that formally, an element of the disjoint union is a pair (g, α), where α is a "tag" to distinguish which group g came from. We will suppress the tag in our notation, but remember that elements of groups corresponding to different indices have to be considered distinct, even if the groups are the same.) The sequence of length zero, called the *empty word*, is denoted by $(\ )$. Let $\mathcal{W}$ denote the set of all words in $\{G_\alpha\}$. We denote the identity element of G_α by 1_α.

Define a multiplication operation in $\mathcal{W}$ by concatenation:

$$(g_1, \ldots, g_m)(h_1, \ldots, h_l) = (g_1, \ldots, g_m, h_1, \ldots, h_l).$$

Clearly, this multiplication is associative, and has the empty word as a two-sided identity element. However, there are two problems with this structure as it stands: First, $\mathcal{W}$ is not a group under this operation because there are no inverses; and second, the group structures of the various groups G_α have not played a role in the definition so far.

To solve both of these problems, we define an equivalence relation on the set of words as follows. An *elementary reduction* is an operation of one of the following forms:

$$(g_1, \dots, g_i, g_{i+1}, \dots, g_m) \mapsto (g_1, \dots, g_i g_{i+1}, \dots, g_m) \text{ if } g_i, g_{i+1} \in \text{some } G_\alpha;$$
$$(g_1, \dots, g_{i-1}, 1_\alpha, g_{i+1}, \dots, g_m) \mapsto (g_1, \dots, g_{i-1}, g_{i+1}, \dots, g_m).$$

The first operation just multiplies together two consecutive entries, provided that they are elements of the same group, and the second deletes any identity element that appears in a word. We say that two words are *equivalent*, written $W \sim W'$, if one can be transformed into the other by a finite sequence of elementary reductions or their inverses; this is obviously an equivalence relation. The set of equivalence classes is called the *free product* of the groups $\{G_\alpha\}$, and is denoted by $\ast_\alpha G_\alpha$. In the case of a finite set of groups, we just write $G_1 * \cdots * G_n$.

Lemma 9.1. *Given any collection of groups $\{G_\alpha\}$, their free product is a group under the multiplication operation induced by multiplication of words.*

Proof. First we need to check that multiplication of words respects the equivalence relation. If V' is obtained from V by an elementary reduction, then it is easy to see that $V'W$ is similarly obtained from VW, as is WV' from WV. If $V \sim V'$ and $W \sim W'$, it follows by induction on the number of elementary reductions that $VW \sim V'W'$. Thus multiplication is well-defined on equivalence classes.

The equivalence class of the empty word ( ) is obviously an identity element, and multiplication is associative on equivalence classes because it already is on words. Finally, for any word $(g_1, \dots, g_m)$, it is easy to check that

$$(g_1, \dots, g_m)(g_m^{-1}, \dots, g_1^{-1}) \sim (\ ) \sim (g_m^{-1}, \dots, g_1^{-1})(g_1, \dots, g_m),$$

so the equivalence class of $(g_m^{-1}, \dots, g_1^{-1})$ is an inverse for that of $(g_1, \dots, g_m)$. □

Henceforth, we will denote the identity element of the free product (the equivalence class of the empty word) by 1.

For many purposes it is important to have a *unique* representative of each equivalence class in the free product. We say that a word $(g_1, \dots, g_m)$ is *reduced* if it cannot be shortened by an elementary reduction. Specifically, this means that no element g_i is the identity of its group, and no two consecutive elements g_i, g_{i+1} come from the same group. It is easy to see that any word is equivalent to a reduced word: Just perform elementary reductions until it is impossible to perform any more. What is not so easy to see is that the reduced word representing any given equivalence class is unique.

Proposition 9.2. *Every element of $*_\alpha G_\alpha$ is represented by a unique reduced word.*

Proof. We showed above that every equivalence class contains a reduced word, so we need only check that two reduced words representing the same equivalence class must be equal. This amounts to constructing a "canonical reduction algorithm."

Let $\mathcal{R}$ denote the set of reduced words. We begin by constructing a map

$$\mathcal{W} \times \mathcal{R} \to \mathcal{R},$$

which sends $(g_1, \dots, g_m) \in \mathcal{W}$ and $(h_1, \dots, h_l) \in \mathcal{R}$ to a reduced word that we denote by $(g_1, \dots, g_m) \cdot (h_1, \dots, h_l) \in \mathcal{R}$. We will define the map by induction on m, the length of the word in $\mathcal{W}$. For $m = 0$, define

$$(\,) \cdot (h_1, \dots, h_l) = (h_1, \dots, h_l).$$

For $m = 1$ and $g \in G_\alpha$, set

$$(g) \cdot (h_1, \dots, h_l) = \begin{cases} (h_2, \dots, h_l), & h_1 \in G_\alpha \text{ and } gh_1 = 1_\alpha; \\ (gh_1, \dots, h_l), & h_1 \in G_\alpha \text{ and } gh_1 \neq 1_\alpha; \\ (h_1, \dots, h_l), & h_1 \notin G_\alpha \text{ and } g = 1_\alpha; \\ (g, h_1, \dots, h_l), & h_1 \notin G_\alpha \text{ and } g \neq 1_\alpha. \end{cases}$$

(The idea is just to multiply the two words and reduce them in the obvious way; what is important about this definition is that there are no arbitrary choices involved.) For $m > 1$, define the map recursively:

$$\begin{aligned} (g_1, \dots, g_m) \cdot (h_1, \dots, h_l) &= (g_1) \cdot \big((g_2, \dots, g_m) \cdot (h_1, \dots, h_l)\big) \\ &= (g_1) \cdot (g_2) \cdot \dots \cdot (g_m) \cdot (h_1, \dots, h_l), \end{aligned}$$

where we understand the dot operation to be performed from right to left: $U \cdot V \cdot W = U \cdot (V \cdot W)$.

The key feature of this operation is that it takes equivalent words to the same reduced word: If $W \sim W'$, then $W \cdot V = W' \cdot V$ for all reduced words V. To prove this, it suffices to assume that W' is obtained from W by an elementary reduction. There are two cases, corresponding to the two types of elementary reduction. Suppose first that $W = (g_1, \dots, g_i, g_{i+1}, \dots, g_m)$, and $W' = (g_1, \ldots, g_i g_{i+1}, \ldots, g_m)$ is obtained by multiplying together two consecutive elements g_i, g_{i+1} from the same group G_α. Then

$$W \cdot V = (g_1) \cdot \cdots \cdot (g_{i-1}) \cdot (g_i) \cdot (g_{i+1}) \cdot (g_{i+2}) \cdot \cdots \cdot (g_m) \cdot V.$$

Writing $(g_{i+2}) \cdot \cdots \cdot (g_m) \cdot V = (h_1, \ldots, h_l)$, it suffices to show that

$$(g_i) \cdot (g_{i+1}) \cdot (h_1, \ldots, h_l) = (g_i g_{i+1}) \cdot (h_1, \ldots, h_l).$$

Applying the definition of the dot operator twice and keeping careful track of the various cases, you can compute

$$\begin{aligned}
&(g_i) \cdot (g_{i+1}) \cdot (h_1, \ldots, h_l) \\
&\qquad = \begin{cases} (h_2, \ldots, h_l), & h_1 \in G_\alpha,\ g_i g_{i+1} h_1 = 1_\alpha; \\ (g_i g_{i+1} h_1, \ldots, h_l), & h_1 \in G_\alpha,\ g_i g_{i+1} h_1 \neq 1_\alpha; \\ (h_1, \ldots, h_l), & h_1 \notin G_\alpha,\ g_i g_{i+1} = 1_\alpha; \\ (g_i g_{i+1}, h_1, \ldots, h_l), & h_1 \notin G_\alpha,\ g_i g_{i+1} \neq 1_\alpha. \end{cases}
\end{aligned}$$

On the other hand, $(g_i g_{i+1}) \cdot (h_1, \ldots, h_l)$ is equal to the same value by definition. The second case, in which W contains an identity element that is deleted to obtain W', follows in a similar way from the fact that $(1_\alpha) \cdot V = V$.

Now we define our canonical reduction operator $r\colon \mathcal{W} \to \mathcal{R}$ by $r(W) = W \cdot (\ )$. Clearly, if W is already reduced, then $r(W) = W$. Moreover, by the argument above, if $W \sim W'$, then $r(W) = r(W')$. Thus, if $W \sim W'$ and both are reduced words, we have $W = r(W) = r(W') = W'$. This proves the proposition. □

For each group G_α, there is a canonical map $\iota_\alpha\colon G_\alpha \to *_\alpha G_\alpha$, defined by sending $g \in G_\alpha$ to the equivalence class of the word (g). Each of these maps is a homomorphism, since $(g_1 g_2) \sim (g_1)(g_2)$ for $g_1, g_2 \in G_\alpha$. Each map is also injective: If $g \neq 1_\alpha$, both the words $\iota_\alpha(g) = (g)$ and $\iota_\alpha(1_\alpha) = (\ )$ are reduced, and therefore cannot represent the same equivalence class because of the preceding proposition. We usually identify G_α with its image under the injection ι_α, and write the equivalence class of the word (g) simply as g. Therefore, the equivalence class of a word $(g_1, g_2, \ldots, g_m)$ can be written $g_1 g_2 \cdots g_m$; by a slight abuse of terminology, we will also call such a product a word, and say that it is reduced if the word $(g_1, g_2, \ldots, g_m)$ is reduced. Multiplication in the free product is indicated by juxtaposition of such words. Thus we have finally succeeded in making mathematical sense of products of elements in different groups.

Example 9.3. Let $\mathbb{Z}/\langle 2 \rangle$ denote the group of integers modulo 2. The free product $\mathbb{Z}/\langle 2 \rangle * \mathbb{Z}/\langle 2 \rangle$ can be described as follows. If we let β and γ denote the nontrivial elements of the first and second copies of $\mathbb{Z}/\langle 2\rangle$, respectively, each element of $\mathbb{Z}/\langle 2\rangle * \mathbb{Z}/\langle 2\rangle$ other than the identity has a unique representation as a string of alternating β's and γ's. Multiplication is performed by concatenating the strings and deleting all consecutive pairs of β's or γ's. For example,

$$(\beta\gamma\beta\gamma\beta)(\gamma\beta\gamma\beta) = \beta\gamma\beta\gamma\beta\gamma\beta\gamma\beta;$$

$$(\gamma\beta\gamma\beta)(\beta\gamma\beta\gamma\beta) = \beta.$$

This group is not abelian.

Example 9.4. Later we will need to consider the free product $\pi_1(\mathbb{S}^1, 1) * \pi_1(\mathbb{S}^1, 1)$. Letting $\alpha(s) = e^{2\pi i s}$ as in the preceding chapter, and letting β, γ denote the path classes of α in the first and second copies of $\pi_1(\mathbb{S}^1, 1)$, respectively, each element of $\pi_1(\mathbb{S}^1, 1) * \pi_1(\mathbb{S}^1, 1)$ other than the identity has a unique expression of the form $\beta^{i_1}\gamma^{j_1}\cdots\beta^{i_m}\gamma^{j_m}$, where i_1 or j_m may be zero, but none of the other exponents is zero.

The free product of groups has an important characteristic property.

Theorem 9.5 (Characteristic Property of the Free Product). *For any group H and any collection of homomorphisms $\varphi_\alpha\colon G_\alpha \to H$, there exists a unique homomorphism $\Phi\colon \mathop{\ast}_\alpha G_\alpha \to H$ such that for each α the following diagram commutes:*

$$\begin{tikzcd} \mathop{\ast}_\alpha G_\alpha \arrow[dr, "\Phi"] & \\ G_\alpha \arrow[u, "\iota_\alpha"] \arrow[r, "\varphi_\alpha"'] & H. \end{tikzcd} \tag{9.1}$$

Proof. Suppose we are given a collection of homomorphisms $\varphi_\alpha\colon G_\alpha \to H$. The requirement that $\Phi \circ \iota_\alpha = \varphi_\alpha$ implies that the desired homomorphism Φ must satisfy

$$\Phi(g) = \varphi_\alpha(g) \quad \text{if } g \in G_\alpha, \tag{9.2}$$

where, as usual, we identify G_α with its image under ι_α. Since Φ is supposed to be a homomorphism, it must satisfy

$$\Phi(g_1\cdots g_m) = \Phi(g_1)\cdots\Phi(g_m). \tag{9.3}$$

Therefore, if Φ and $\widetilde{\Phi}$ both satisfy the conclusion, they must be equal because both must satisfy (9.2) and (9.3). This proves that Φ is unique if it exists.

To prove existence of Φ, we use (9.2) and (9.3) to *define* it. This is clearly a homomorphism that satisfies the required properties, provided that it is well-defined. To verify that it is well-defined, we need to check that it gives the same result when applied to equivalent words. As usual, we need only check elementary reductions. If $g_i, g_{i+1} \in G_\alpha$, we have

$$\Phi(g_i g_{i+1}) = \varphi_\alpha(g_i g_{i+1}) = \varphi_\alpha(g_i)\varphi_\alpha(g_{i+1}) = \Phi(g_i)\Phi(g_{i+1}),$$

from which it follows that the definition of Φ is unchanged by multiplying together successive elements of the same group. Similarly, $\Phi(1_\alpha) = \varphi_\alpha(1_\alpha) = 1 \in H$, which shows that Φ is unchanged by deleting an identity element. This completes the proof. □

Corollary 9.6. *The free product is the sum in the category of groups.*

Proof. The characteristic property is exactly the defining property of the sum in a category. □

Corollary 9.7. *The free product is the unique group* (*up to isomorphism*) *satisfying the characteristic property.*

Proof. Lemma 7.37 shows that sums in any category are unique up to isomorphism. □

In some texts, a free product is *defined* as any group satisfying the characteristic property, or as the sum in the category of groups. One must then prove the existence of such a group by some construction such as the one we have given before one is entitled to talk about "the" free product. Once existence is proved, uniqueness follows automatically from category theory. The nice thing about this uniqueness result is that no matter what specific construction is used to define the free product (and you will find many in the literature), they are all the same up to isomorphism.

Free Groups

In this section we will use the free product construction to create a new class of groups called "free groups," consisting of all possible products of a set of "generators," with no relations imposed at all. We begin with a few more definitions.

Let G be a group. A subset $S \subset G$ is said to *generate* G, and the elements of S are called *generators* for G, if every element of G can be written as a product of elements of S. Of course, any group has a set of generators, since we can take S to be the whole group G. But it is more interesting to find a small set of generators when possible.

For example, a cyclic group is a group with one generator (see the Appendix). The cyclic groups are all isomorphic either to $\mathbb{Z}$ or to $\mathbb{Z}/\langle n\rangle$ for some n (Exercise A.26).

In this section we will be concerned mostly with infinite cyclic groups. Given any object α, we can form an infinite cyclic group generated by α, denoted by $\langle\alpha\rangle$, as follows: $\langle\alpha\rangle$ is the set $\{\alpha\}\times\mathbb{Z}$ with multiplication $(\alpha,m)(\alpha,k)=(\alpha,k+m)$. We identify α with the element $(\alpha,1)$; thus we can abbreviate (α,m) by α^m, and think of $\langle\alpha\rangle$ as the group of all integral powers of α with the obvious multiplication.

Now suppose we are given any set S. We define the *free group on S*, denoted by $\langle S\rangle$, to be the free product of all the infinite cyclic groups generated by elements of S:

$$\langle S\rangle = \mathop{\ast}_{\alpha\in S}\langle\alpha\rangle.$$

There is a natural injection $\iota\colon S\hookrightarrow\langle S\rangle$, defined by sending each $\alpha\in S$ to the word $\alpha\in\langle S\rangle$. Thus we can consider S as a subset of $\langle S\rangle$, and each element of $\langle S\rangle$ can be expressed as a word $\alpha_1^{n_1}\alpha_2^{n_2}\cdots\alpha_m^{n_m}$, where each α_i is some element of S and each n_i is an integer. Multiplication is performed by juxtaposition and combining consecutive powers of the same α_i by the rule $\alpha_i^n\alpha_i^k=\alpha_i^{n+k}$. In case $S=\{\alpha_1,\dots,\alpha_n\}$ is a finite set, we denote the free group on S by $\langle\alpha_1,\dots,\alpha_n\rangle$ instead of the more accurate but cumbersome notation $\langle\{\alpha_1,\dots,\alpha_n\}\rangle$. (We will rely on the context and typographical differences to make clear the distinction between the free group $\langle S\rangle$ on the elements of the set S and the infinite cyclic group $\langle\alpha\rangle$, which is also equal to the free group on the singleton $\{\alpha\}$.)

Example 9.8. The free group on the empty set, denoted by $\langle\ \rangle$, is by convention just the trivial group $\{1\}$. The free group on a singleton $\{\alpha\}$ is the infinite cyclic group $\langle\alpha\rangle$. The free group on the two-element set $\{\beta,\gamma\}$ is $\langle\beta,\gamma\rangle=\langle\beta\rangle*\langle\gamma\rangle$, which is the same as the group described in Example 9.4.

Theorem 9.9 (Characteristic Property of the Free Group). *Let S be a set. For any group H and any map $\varphi\colon S\to H$, there exists a unique homomorphism $\widetilde{\varphi}\colon\langle S\rangle\to H$ extending φ:*

$$\begin{tikzcd} \langle S\rangle \arrow[dr, "\widetilde{\varphi}"] & \\ S \arrow[u, "\iota"] \arrow[r, "\varphi"'] & H. \end{tikzcd} \tag{9.4}$$

Proof. This can be proved directly as in the proof of Theorem 9.5. Alternatively, recalling that the free group is defined as a free product, we can proceed as follows. There is a one-to-one correspondence between set functions $\varphi\colon S\to H$ and collections of homomorphisms $\varphi_\alpha\colon\langle\alpha\rangle\to H$ for all

$\alpha \in S$, by the equation

$$\varphi_\alpha(\alpha^n) = \varphi(\alpha)^n.$$

Translating the characteristic property of the free product to this special case and using this correspondence yields the result. The details are left as an exercise. □

Exercise 9.1. Carry out the details of the proof of Theorem 9.9.

Exercise 9.2. Prove that the free group on S is the unique group (up to isomorphism) satisfying the characteristic property.

Presentations of Groups

It is often convenient to describe a group by giving a set of generators for it, and listing a few rules, or "relations," that describe how to multiply the generators together. For example, the cyclic group of order n generated by γ might be described as the group generated by γ with the single relation $\gamma^n = 1$; all other relations in the group, such as $\gamma^{3n} = 1$ or $\gamma^{k-n} = \gamma^k$, follow from this one. The direct product group $\mathbb{Z} \times \mathbb{Z}$ might be described as the group with two generators β, γ satisfying the relation $\beta\gamma = \gamma\beta$. The free group $\langle \beta, \gamma \rangle$ can be described as the group generated by β, γ with no relations.

So far, this is mathematically very vague. What does it mean to say that "all other relations follow from a given one"? In this section we develop a way to make these notions precise.

We define a *group presentation* to be an ordered pair, denoted by $\langle S|R \rangle$, where S is an arbitrary set and R is a set of elements of the free group $\langle S \rangle$. The elements of S and R are called the *generators* and *relators*, respectively, of the presentation. A group presentation defines a group, also denoted by $\langle S|R \rangle$, as the following quotient:

$$\langle S|R \rangle = \langle S \rangle / \overline{R},$$

where $\overline{R}$ is the *normal closure* of R in $\langle S \rangle$, which is the intersection of all normal subgroups of $\langle S \rangle$ containing R; thus $\overline{R}$ is the "smallest" normal subgroup containing R.

Since the quotient of a group by a normal subgroup is again a group (see the Appendix), $\langle S|R \rangle$ is indeed a group. Each of the generators $s \in S$ determines an element in $\langle S|R \rangle$ (its coset in the quotient group), which we usually write also as s. Each of the relators $r \in R$ represents a particular product of generators and their inverses that is equal to 1 in the quotient.

Here is the intuition behind this construction. If G is any group generated by S, there is a surjective homomorphism $\Phi\colon \langle S\rangle \to G$, whose existence is guaranteed by the characteristic property of the free group. If all the words of R are to be equal to the identity in G, then the kernel of Φ must at least contain R, and since it is normal, it must contain $\overline{R}$; thus by the first isomorphism theorem (Theorem A.13 in the Appendix), G is isomorphic to a quotient of $\langle S\rangle$ by a normal subgroup containing $\overline{R}$. By dividing out exactly $\overline{R}$, we ensure that the only relations that hold in $\langle S|R\rangle$ are those that are forced by the relators in R. Thus, in a certain sense, $\langle S|R\rangle$ is the "largest" group generated by S in which all the products represented by elements of R are equal to 1.

If G is a group and there exists an isomorphism $\langle S|R\rangle \cong G$, we say that $\langle S|R\rangle$ is a *presentation of* G. At this point, the question naturally arises whether every group has a presentation. In fact, the answer is yes, but the result is not as satisfying as we might have hoped. Given a group G, the set of all elements of G certainly generates G. By the characteristic property of the free group, the identity map of G to itself has a unique extension to a homomorphism $\Phi\colon \langle G\rangle \to G$. If we set $R = \operatorname{Ker}\Phi$, then the first isomorphism theorem says that $G \cong \langle G\rangle/R$. Since R is normal, it is equal to its normal closure, and therefore G has the presentation $\langle G|R\rangle$. This is highly inefficient, of course, since both $\langle G\rangle$ and R are vastly larger than G itself.

If G admits a presentation $\langle S|R\rangle$ in which both S and R are finite sets, we say that G is *finitely presented.* In this case, we usually write the presentation as $\langle \alpha_1,\dots,\alpha_n \mid r_1,\dots,r_m\rangle$. Since the r_i actually all become equal to the identity in the group defined by the presentation, it is also often convenient to replace the relators by the equations obtained by setting them equal to the identity, called *relations* of the presentation, as in

$$\langle \alpha_1,\dots,\alpha_n \mid r_1 = 1,\dots,r_m = 1\rangle$$

or even

$$\langle \alpha_1,\dots,\alpha_n \mid r_1 = q_1,\dots,r_m = q_m\rangle.$$

We take this to be an alternative notation for $\langle \alpha_1,\dots,\alpha_n \mid r_1q_1^{-1},\dots, r_mq_m^{-1}\rangle$.

We conclude this section by describing one important example in detail.

Proposition 9.10. *The group $\mathbb{Z}\times\mathbb{Z}$ has the presentation $\langle\beta,\gamma \mid \beta\gamma = \gamma\beta\rangle$.*

Proof. For brevity, write $G = \langle\beta,\gamma \mid \beta\gamma = \gamma\beta\rangle = \langle\beta,\gamma \mid \beta\gamma\beta^{-1}\gamma^{-1}\rangle$. As usual, we will use the symbols β and γ to denote either the generators of the free group $\langle\beta,\gamma\rangle$ or their images in the quotient group G. We begin by noting that G is abelian: The equation $\beta\gamma\beta^{-1}\gamma^{-1} = 1$, which holds in G by definition, immediately implies $\beta\gamma = \gamma\beta$, and then a simple induction shows that any products of powers of β and γ commute with each other. Since β and γ generate G, this suffices.

We will prove the proposition by defining homomorphisms $\Phi\colon G \to \mathbb{Z}\times\mathbb{Z}$ and $\Psi\colon \mathbb{Z}\times\mathbb{Z} \to G$ and showing that they are inverses of each other. To define Φ, we first define $\widetilde{\Phi}\colon \langle\beta,\gamma\rangle \to \mathbb{Z}\times\mathbb{Z}$ by setting $\widetilde{\Phi}(\beta) = (1,0)$ and $\widetilde{\Phi}(\gamma) = (0,1)$; this uniquely determines $\widetilde{\Phi}$ by the characteristic property of the free group. Explicitly, $\widetilde{\Phi}$ is given by

$$\widetilde{\Phi}(\beta^{i_1}\gamma^{j_1}\cdots\beta^{i_m}\gamma^{j_m}) = (i_1+\cdots+i_m, j_1+\cdots+j_m). \tag{9.5}$$

Because $\beta\gamma\beta^{-1}\gamma^{-1} \in \operatorname{Ker}\widetilde{\Phi}$ by direct computation, $\widetilde{\Phi}$ descends to a map $\Phi\colon G \to \mathbb{Z}\times\mathbb{Z}$ still given by (9.5).

In the other direction, we define $\Psi\colon \mathbb{Z}\times\mathbb{Z} \to G$ by

$$\Psi(m,n) = \beta^m\gamma^n.$$

It follows from the fact that G is abelian that Ψ is a homomorphism. A simple computation shows that $\Psi\circ\Phi(\beta) = \beta$, $\Psi\circ\Phi(\gamma) = \gamma$, and $\Phi\circ\Psi(m,n) = (m,n)$. Thus Φ and Ψ are inverses, so $G \cong \mathbb{Z}\times\mathbb{Z}$. □

In some ways, a presentation gives a very simple and concrete way to understand the properties of a group, and we will describe the fundamental groups of surfaces in the next chapter by giving presentations. However, you should be aware that even with a finite presentation in hand, some very basic questions about a group may still be difficult or impossible to answer. For example, two of the most basic problems concerning group presentations were first posed around 1910 by topologists Heinrich Tietze and Max Dehn, shortly after the invention of the fundamental group: The *isomorphism problem* for groups is to decide, given two finite presentations, whether the resulting groups are isomorphic; and the *word problem* is to decide, given a finite presentation $\langle S|R\rangle$ and a specific word formed from elements of S, whether that word represents the identity element of the group $\langle S|R\rangle$. It was shown in the 1950s that there is no algorithm for solving either of these problems that is guaranteed to yield an answer for every presentation in a finite amount of time! (See [Sti82] for references and historical background.) These ideas form the basis for the subject called *combinatorial group theory*, which is a lively research field at the intersection of algebra, topology, and geometry.

Free Abelian Groups

There is an analogue of free groups in the category of abelian groups. In this section, since all our groups will be abelian, we will always write the group operation additively, and denote the identity element by 0 and the inverse of x by $-x$. If G is an abelian group, $g \in G$, and $n \in \mathbb{Z}$, the notation ng means the n-fold sum $g + \cdots + g$, and nG is the subgroup $\{ng : g \in G\}$.

Given a nonempty set S, let $\mathbb{Z}\langle S\rangle$ denote the set of all functions $k\colon S \to \mathbb{Z}$ such that $k(s) = 0$ for all but finitely many $s \in S$. This is easily seen to be an abelian group under addition, called the *free abelian group* on S. Just as we did for the free vector space defined in Chapter 5, we can identify each $s \in S$ with the element of $\mathbb{Z}\langle S\rangle$ that takes the value 1 on s and zero on every other element of S, so we consider S as a subset of $\mathbb{Z}\langle S\rangle$, and each element of $\mathbb{Z}\langle S\rangle$ can be written uniquely as a finite sum of the form

$$\sum_{i=1}^{n} k_i s_i$$

where s_i are elements of S and k_i are integers. When $S = \{s_1, \dots, s_n\}$ is a finite set, we will usually write the free abelian group on S as $\mathbb{Z}\langle s_1, \dots, s_n\rangle$. By convention, the free abelian group on the empty set is the trivial group $\{0\}$ (we consider a "linear combination of no elements" to sum to 0).

Lemma 9.11 (Properties of Free Abelian Groups). *Let S be a nonempty set.*

(*a*) Characteristic Property: *Given any abelian group H and any set map $f\colon S \to H$, there exists a unique homomorphism $\widetilde{f}\colon \mathbb{Z}\langle S\rangle \to H$ extending f.*

(*b*) *$\mathbb{Z}\langle S\rangle$ is isomorphic to the direct sum $\bigoplus_{s\in S}\langle s\rangle$ of all the infinite cyclic groups generated by elements of S.*

(*c*) *If $S = \{s_1, \dots, s_n\}$ is finite, then $\mathbb{Z}\langle s_1, \dots, s_n\rangle$ is isomorphic to $\mathbb{Z}^n$ via the map $(k_1, \dots, k_n) \mapsto k_1 s_1 + \cdots + k_n s_n$.*

Exercise 9.3. Prove Lemma 9.11.

Let G be an abelian group. By analogy with vector spaces, a finite sum of elements of G with integer coefficients is called a *linear combination* of elements of G. A nonempty subset $S \subset G$ is said to be *linearly independent* if the only linear combination of elements of S that equals zero is the one for which all the coefficients are zero. A *basis* for G is a linearly independent subset that generates G. Just as in the case of vector spaces, if S is a basis for G, every element of G can be written uniquely as a linear combination of elements of S. For example, S is a basis for the free abelian group $\mathbb{Z}\langle S\rangle$. The set of elements $e_i = (0, \dots, 1, \dots, 0)$ (with a 1 in the ith place) for $i = 1, \dots, n$ is a basis for $\mathbb{Z}^n$, which we call the *standard basis*.

If a group G is isomorphic to $\mathbb{Z}\langle S\rangle$ for some set S, G is also said to be free abelian.

Exercise 9.4.

(a) Show that an abelian group is free abelian if and only if it has a basis.
(b) Show that any two free abelian groups whose bases have the same cardinality are isomorphic.

Lemma 9.12. *If G has a finite basis, then every finite basis has the same number of elements.*

Proof. Suppose G has a basis with n elements. Then $G \cong \mathbb{Z}^n$ by Lemma 9.11(c), and the quotient group $G/2G$ is easily seen to be isomorphic to $(\mathbb{Z}/\langle 2\rangle)^n$, which has exactly 2^n elements. Since the order of $G/2G$ is independent of the choice of basis, every finite basis must have n elements. $\square$

In view of this lemma, if G is a free abelian group with a finite basis, we define the *rank* of G to be the number of elements in any finite basis. (In fact, in that case every basis is finite—see Problem 9-6.) If G has no finite basis, we say it has infinite rank.

Proposition 9.13. *Suppose G is a free abelian group of finite rank. Any subgroup of G is free abelian of rank less than or equal to that of G.*

Proof. We may assume without loss of generality that $G = \mathbb{Z}^n$. We will prove the proposition by induction on n. For $n = 1$, it follows from the fact that any subgroup of a cyclic group is cyclic.

Suppose the result is true for subgroups of $\mathbb{Z}^{n-1}$, and let H be any subgroup of $\mathbb{Z}^n$. Identifying $\mathbb{Z}^{n-1}$ with the subgroup $\{(k_1, \dots, k_{n-1}, 0)\}$ of $\mathbb{Z}^n$, the inductive hypothesis guarantees that $H \cap \mathbb{Z}^{n-1}$ is free abelian of rank $m - 1 \le n - 1$, so has a basis $\{h_1, \dots, h_{m-1}\}$. If $H \subset \mathbb{Z}^{n-1}$, we are done. Otherwise, the image of H under the projection $\pi_n \colon \mathbb{Z}^n \to \mathbb{Z}$ onto the nth factor is a nontrivial cyclic subgroup of $\mathbb{Z}$. Let $c \in \mathbb{Z}$ be a generator of this subgroup, and let h_m be an element of H such that $\pi_n(h_m) = c$. The proof will be complete once we show that $\{h_1, \dots, h_m\}$ is a basis for H.

Suppose $a_1 h_1 + \cdots + a_m h_m = 0$. Applying π_n to this equation yields $a_m c = 0$, so $a_m = 0$. Then $a_1 = \cdots = a_{m-1} = 0$ because of the independence of $\{h_1, \dots, h_{m-1}\}$, so $\{h_1, \dots, h_m\}$ is linearly independent. Now suppose $h \in H$ is arbitrary. Then $\pi_n(h) = ac$ for some integer a, so $h - ah_m \in H \cap \mathbb{Z}^{n-1}$. This element can be written as a linear combination of $\{h_1, \dots, h_{m-1}\}$, which shows that H is generated by $\{h_1, \dots, h_m\}$. $\square$

We will need to extend the notion of rank to finitely generated abelian groups that are not necessarily free abelian. To that end, we say that an element g of an abelian group G is a *torsion element* if $ng = 0$ for some nonzero $n \in \mathbb{Z}$. If $ng = n'g' = 0$, then $nn'(g + g') = 0$, so the set of all torsion elements is a subgroup G_{tor} of G, called the *torsion subgroup*. We say that G is *torsion free* if the only torsion element is 0. It is easy to check that the quotient group G/G_{tor} is torsion free.

Proposition 9.14. *Any abelian group that is finitely generated and torsion free is free abelian of finite rank.*

Proof. Suppose G is such a group. For any linearly independent subset $S \subset G$, we will extend our notation slightly and let $\mathbb{Z}\langle S\rangle$ denote the subgroup of G generated by S. It is easily seen to be free abelian with S as a basis, so this is consistent with our earlier notation.

The crux of the proof is the following claim: *There exists a nonzero integer n and a finite linearly independent set $S \subset G$ such that $nG \subset \mathbb{Z}\langle S\rangle$.* Assuming this, the rest of the proof goes as follows. Let $\varphi\colon G \to G$ be the homomorphism $\varphi(g) = ng$. It is injective because G is torsion free, and the claim implies that $\varphi(G) \subset \mathbb{Z}\langle S\rangle$. Thus G is isomorphic to the subgroup $\varphi(G)$ of the free abelian group $\mathbb{Z}\langle S\rangle$, so by Proposition 9.13, G is free abelian of finite rank.

We will prove the claim by induction on the number of elements in a generating set for G. If G is generated by one element g, the claim is true with $n = 1$, because the fact that G is torsion free implies that $\{g\}$ is a linearly independent set.

Now assume that the claim is true for any torsion-free group generated by $m-1$ elements, and suppose G is generated by a set $T = \{g_1, \dots, g_m\}$ with m elements. If T is linearly independent, we just take $S = T$. If not, there is a relation of the form $a_1 g_1 + \cdots + a_m g_m = 0$ with at least one of the coefficients, say a_m, not equal to zero. Letting G' denote the subgroup of G generated by $\{g_1, \dots, g_{m-1}\}$, this means that $a_m g_m \in G'$. Since G' is generated by $m-1$ elements, by induction there exist a nonzero integer n' and a finite linearly independent set $S \subset G'$ such that $n'G' \subset \mathbb{Z}\langle S\rangle$. Let $n = a_m n'$. Since G is generated by T, for any $g \in G$ we have

$$\begin{aligned} ng &= a_m n'(b_1 g_1 + \cdots + b_m g_m) \\ &= n'(a_m b_1 g_1 + \cdots + a_m b_{m-1} g_{m-1}) + n' b_m (a_m g_m). \end{aligned}$$

Both terms above are in $n'G' \subset \mathbb{Z}\langle S\rangle$. It follows that $nG \subset \mathbb{Z}\langle S\rangle$, which completes the proof. □

Now let G be any finitely generated abelian group. Because G/G_{tor} is finitely generated and torsion free, the preceding proposition implies that it is free abelian of finite rank. Thus we can define the rank of G to be the rank of G/G_{tor}.

Example 9.15. The rank of $\mathbb{Z}^n$ is n, and the rank of any finite group is 0 (since every element is a torsion element). The rank of a product group of the form $G = \mathbb{Z}^n \times \mathbb{Z}/\langle k_1\rangle \times \cdots \times \mathbb{Z}/\langle k_m\rangle$ is n, because $G_{\mathrm{tor}} = \mathbb{Z}/\langle k_1\rangle \times \cdots \times \mathbb{Z}/\langle k_m\rangle$.

Proposition 9.16. *If G is a free abelian group of finite rank and $f\colon G \to H$ is a surjective homomorphism, then* $\operatorname{rank} G = \operatorname{rank} H + \operatorname{rank}(\operatorname{Ker} f)$.

Proof. Write $K = \operatorname{Ker} f$. By Proposition 9.13, K is a finitely generated free abelian group, so we can choose elements $k_1, \dots, k_p \in G$ that form a basis for K.

Since f is surjective, it takes a set of generators for G to a set of generators for H; thus H is finitely generated and its rank is the rank of H/H_{tor}. Choose a basis $\widetilde{h}_1, \dots, \widetilde{h}_q$ for H/H_{tor}, and lift them to elements $h_1, \dots, h_q \in H$. By surjectivity, there are elements $g_j \in G$ such that $f(g_j) = h_j$.

The set $\{k_i, g_j\}$ is linearly independent, because a relation of the form $g = \sum_i m_i k_i + \sum_j n_j g_j = 0$ implies

$$0 = f(g) = \sum_j n_j f(g_j) = \sum_j n_j h_j.$$

Projecting this to the quotient group H/H_{tor}, we obtain a relation of the form $\sum_j n_j \widetilde{h}_j = 0$, which implies $n_j = 0$ for each j. Therefore, $g = \sum_i m_i k_i = 0$, so $m_i = 0$.

Let $\mathbb{Z}\langle k_i, g_j \rangle$ denote the subgroup of G generated by $\{k_i, g_j\}$. It is free abelian of rank $p + q = \operatorname{rank} K + \operatorname{rank} H$, so from Proposition 9.13 we conclude that $\operatorname{rank} K + \operatorname{rank} H \le \operatorname{rank} G$. The proof will be complete once we show the reverse inequality.

Because H_{tor} is a finitely generated torsion group, there is an integer N such that $Nt = 0$ for every $t \in H_{\mathrm{tor}}$. Let $g \in G$ be arbitrary, and let $[f(g)]$ denote the equivalence class of $f(g)$ in the quotient H/H_{tor}. Writing $[f(g)] = \sum_j n_j \widetilde{h}_j$, we have $[f(g - \sum_j n_j g_j)] = 0$, so $f(g - \sum_j n_j g_j) \in H_{\mathrm{tor}}$. This implies $N(g - \sum_j n_j g_j) \in K$, so we can write

$$Ng = \sum_j N n_j g_j + \sum_i m_i k_i.$$

Letting $\varphi\colon G \to G$ be the homomorphism $\varphi(g) = Ng$ as in the proof of Proposition 9.14, we have shown that $\varphi(G) \subset \mathbb{Z}\langle k_i, g_j \rangle$. Moreover, φ is injective because G is torsion free, so by Proposition 9.13 again we conclude that $\operatorname{rank} G \le p + q = \operatorname{rank} K + \operatorname{rank} H$. $\square$

Problems

9-1. Show that any free product of two or more nontrivial groups is infinite and nonabelian.

9-2. Show that the following groups have the given presentations.

(a) $\mathbb{Z}/\langle n\rangle \cong \langle \beta \mid \beta^n = 1\rangle$.

(b) $\mathbb{Z}/\langle m\rangle \times \mathbb{Z}/\langle n\rangle \cong \langle \beta, \gamma \mid \beta^m = 1,\ \gamma^n = 1,\ \beta\gamma = \gamma\beta\rangle$.

9-3. The *center* of a group G is the set Z of elements of G that commute with every element of G: $Z = \{g \in G : gh = hg \text{ for all } h \in G\}$. Show that a free group on two or more generators has center consisting of the identity alone.

9-4. Let G_1, G_2, H_1, H_2 be groups, and let $f_i\colon G_i \to H_i$ be group homomorphisms for $i = 1, 2$.

(a) Using the characteristic property of the free product, show that there is a unique homomorphism $f_1 * f_2\colon G_1 * G_2 \to H_1 * H_2$ such that the following diagram commutes for $i = 1, 2$:

$$\begin{tikzcd} G_1 * G_2 \arrow[r, "f_1 * f_2"] & H_1 * H_2 \\ G_i \arrow[u, "\iota_i"] \arrow[r, "f_i"'] & H_i, \arrow[u, "\iota_i'"'] \end{tikzcd}$$

where $\iota_i\colon G_i \to G_1 * G_2$ and $\iota_i'\colon H_i \to H_1 * H_2$ are the canonical injections.

(b) Show that $\operatorname{Ker} f_1 * f_2 = \operatorname{Im} j_1 * j_2$, where $j_i\colon \operatorname{Ker} f_i \hookrightarrow G_i$ is inclusion:

$$\operatorname{Ker} f_1 * \operatorname{Ker} f_2 \xrightarrow{j_1 * j_2} G_1 * G_2 \xrightarrow{f_1 * f_2} H_1 * H_2.$$

9-5. Show that the free abelian group on a set S is uniquely determined up to isomorphism by the characteristic property (Lemma 9.11(a)).

9-6. Suppose G is a free abelian group of finite rank. Show that every basis of G is finite.

10

The Seifert–Van Kampen Theorem

In this section we will develop the techniques needed to compute the fundamental groups of all compact surfaces, and a good many other spaces as well. The basic tool is the Seifert–Van Kampen theorem, which gives a formula for the fundamental group of a space that can be decomposed as the union of two open, path connected subsets whose intersection is also path connected.

In the first section we state a rather general version of the theorem. Then we examine two special cases in which the formula simplifies considerably. The first special case is that in which the intersection of the two subsets is simply connected: Then the theorem says that the fundamental group of the big space is the free product of the fundamental groups of its subspaces. As an application, we compute the fundamental groups of a wedge of spaces and of a graph (a one-dimensional simplicial complex). The second special case is that in which one of the two subsets is itself simply connected: Then the fundamental group of the big space is the quotient of the fundamental group of the non–simply connected piece by the path classes in the intersection. We use this formula to compute the fundamental groups of the compact surfaces.

After these applications, we give a detailed proof of the Seifert–Van Kampen theorem. Then in the last section of the chapter we revisit the classification of compact surfaces, and prove finally that the different surfaces on our list are all topologically distinct by showing that their fundamental groups are not isomorphic.

Statement of the Theorem

Here is the situation in which we will be able to compute fundamental groups. Suppose we are given a space X that is the union of two open subsets $U, V \subset X$, and suppose we can compute the fundamental groups of U, V, and $U \cap V$, each of which is path connected. As we will see below, any loop in X can be written as a product of loops, each of which lies in either U or V; such a loop can be thought of as representing an element of the free product $\pi_1(U) * \pi_1(V)$. But any loop in $U \cap V$ will represent only a single element of $\pi_1(X)$, even though it represents two distinct elements of the free product (one in $\pi_1(U)$ and one in $\pi_1(V)$). Thus the fundamental group of X can be thought of as the quotient of this free product modulo some relations coming from $\pi_1(U \cap V)$ that express this redundancy.

Let us set the stage for the precise statement of the theorem. Let X be any topological space, let $U, V \subset X$ be open subsets whose union is X and whose intersection is nonempty, and choose any base point $q \in U \cap V$. The four inclusion maps

$$\begin{tikzcd}
& U \arrow[dr, "k"] & \\
U \cap V \arrow[ur, "i"] \arrow[dr, "j"'] & & X \\
& V \arrow[ur, "l"'] &
\end{tikzcd}$$

induce fundamental group homomorphisms

$$\begin{tikzcd}
& \pi_1(U,q) \arrow[dr, "k_*"] & \\
\pi_1(U \cap V, q) \arrow[ur, "i_*"] \arrow[dr, "j_*"'] & & \pi_1(X,q). \\
& \pi_1(V,q) \arrow[ur, "l_*"'] &
\end{tikzcd}$$

Now insert the free product group $\pi_1(U,q) * \pi_1(V,q)$ into the middle of the picture, and let $\iota_U : \pi_1(U,q) \hookrightarrow \pi_1(U,q) * \pi_1(V,q)$ and $\iota_V : \pi_1(V,q) \hookrightarrow \pi_1(U,q) * \pi_1(V,q)$ be the canonical injections. By the characteristic property of the free product, k_* and l_* induce a homomorphism $\Phi\colon \pi_1(U,q) * \pi_1(V,q) \to \pi_1(X,q)$ such that the right half of the following diagram commutes:

$$\begin{tikzcd}
 & \pi_1(U,q) \arrow[d, "\iota_U"] \arrow[dr, "k_*"] & \\
\pi_1(U\cap V,q) \arrow[ur, "i_*"] \arrow[dr, "j_*"'] \arrow[r, dashed, "F"] & \pi_1(U,q) * \pi_1(V,q) \arrow[r, "\Phi"] & \pi_1(X,q). \\
 & \pi_1(V,q) \arrow[u, "\iota_V"'] \arrow[ur, "l_*"'] &
\end{tikzcd} \tag{10.1}$$

Finally, we define a map $F\colon \pi_1(U\cap V,q) \to \pi_1(U,q) * \pi_1(V,q)$ by setting $F(\gamma) = (i_*\gamma)^{-1}(j_*\gamma)$. ($F$ is not a homomorphism.) Let $\overline{F(\pi_1(U\cap V,q))}$ denote the normal closure of the image of F in $\pi_1(U,q) * \pi_1(V,q)$.

Theorem 10.1 (Seifert–Van Kampen). *Let X be a topological space. Suppose $U, V \subset X$ are open subsets whose union is X, and suppose U, V, and $U\cap V$ are path connected. Then, for any $q \in U\cap V$, the homomorphism Φ defined by* (10.1) *is surjective, and its kernel is $\overline{F(\pi_1(U\cap V,q))}$. Therefore,*

$$\pi_1(X,q) \cong \pi_1(U,q) * \pi_1(V,q) \Big/ \overline{F(\pi_1(U\cap V,q))}. \tag{10.2}$$

When the fundamental groups in question are finitely presented, the theorem has a useful reformulation in terms of generators and relations.

Corollary 10.2. *In addition to the hypotheses of the Seifert–Van Kampen theorem, assume that the fundamental groups of U, V, and $U\cap V$ have the following finite presentations:*

$$\begin{aligned}
\pi_1(U,q) &\cong \langle \alpha_1,\dots,\alpha_m \mid \rho_1,\dots,\rho_r\rangle;\\
\pi_1(V,q) &\cong \langle \beta_1,\dots,\beta_n \mid \sigma_1,\dots,\sigma_s\rangle;\\
\pi_1(U\cap V,q) &\cong \langle \gamma_1,\dots,\gamma_p \mid \tau_1,\dots,\tau_t\rangle.
\end{aligned}$$

Then $\pi_1(X,q)$ has the presentation

$$\begin{aligned}
&\pi_1(X,q)\\
&\quad\cong \langle \alpha_1,\dots,\alpha_m,\beta_1,\dots,\beta_n \mid \rho_1,\dots,\rho_r,\sigma_1,\dots,\sigma_s, u_1=v_1,\dots,u_p=v_p\rangle,
\end{aligned}$$

where for each $a = 1, \ldots, p$, u_a is an expression for $i_*\gamma_a \in \pi_1(U, q)$ in terms of the generators $\{\alpha_1, \ldots, \alpha_m\}$, and v_a similarly expresses $j_*\gamma_a \in \pi_1(V, q)$ in terms of $\{\beta_1, \ldots, \beta_n\}$.

The proofs of these two theorems are rather technical, so we will postpone them until later in the chapter. Before proving them, we will illustrate their use by computing a number of fundamental groups.

It is worth remarking here that the Seifert–Van Kampen theorem can be generalized to a covering of X by any number, finite or infinite, of path connected open sets containing the base point. This generalization can be found in [Sie92] or [Mas89].

Applications

All of our applications of the Seifert–Van Kampen theorem will be in special cases in which one of the sets U, V, or $U \cap V$ is simply connected.

First Special Case: Simply Connected Intersection

The first special case of the Seifert–Van Kampen theorem we will consider is that in which $U \cap V$ is simply connected. In that case, the group $\overline{F(\pi_1(U \cap V, q))}$ is trivial, so the following corollary is immediate.

Corollary 10.3. *Assume the hypotheses of the Seifert–Van Kampen theorem, and suppose in addition that $U \cap V$ is simply connected. Then Φ is an isomorphism between $\pi_1(U, q) * \pi_1(V, q)$ and $\pi_1(X, q)$.*

As our first application, we will compute the fundamental group of a wedge of spaces. Recall from Example 3.25 that the wedge of spaces $X_1, \ldots, X_n$ with base points $q_j \in X_j$ is the space $X_1 \vee \cdots \vee X_n$ defined as the quotient of $\coprod_j X_j$ by the equivalence relation generated by $q_1 \sim \cdots \sim q_n$.

A point q in a topological space X is said to be a *nondegenerate base point* if q has a neighborhood that admits a strong deformation retraction onto q and $\{q\}$ is closed in X. For example, any base point in a manifold is nondegenerate, because any Euclidean ball neighborhood admits a strong deformation retraction onto any point. (In more advanced treatments of homotopy theory a slightly more restrictive definition of nondegenerate base point is used, but this one will suffice for our purposes.)

Observe that inclusion of X_j into $\coprod_j X_j$ followed by projection onto the quotient induces continuous injective maps $\iota_j \colon X_j \hookrightarrow X_1 \vee \cdots \vee X_n$. If each base point q_j is nondegenerate, then these are closed maps and thus embeddings. Identifying each X_j with its image under ι_j, we will consider X_j as a subspace of $X_1 \vee \cdots \vee X_n$. We let $*$ denote the point in $X_1 \vee \cdots \vee X_n$ that is the equivalence class of the base points $q_1, \ldots, q_n$.

Lemma 10.4. *Suppose $q_i \in X_i$ is a nondegenerate base point for $i = 1, \dots, n$. Then $*$ is a nondegenerate base point in $X_1 \vee \cdots \vee X_n$.*

Proof. For each i, choose a neighborhood U_i of q_i that admits a strong deformation retraction $H_i \colon U_i \times I \to U_i$ onto $\{q_i\}$. Define a map $H \colon \coprod_i U_i \times I \to \coprod_i U_i$ by letting $H = H_i$ on $U_i \times I$. The restriction of the quotient map π to the saturated open set $\coprod_i U_i$ is a quotient map to a neighborhood U of $*$. Since $\pi \circ H$ respects the identifications made by π, it descends to the quotient and yields a strong deformation retraction of U onto $\{*\}$. To see that $\{*\}$ is closed in $X_1 \vee \cdots \vee X_n$, we need only observe that its inverse image under π is $\{q_1, \dots, q_n\}$, which is closed in the disjoint union because its intersection $\{q_i\}$ with X_i is closed in X_i by hypothesis. Thus $\{*\}$ is closed. □

Proposition 10.5. *Let $X_1, \dots, X_n$ be spaces with nondegenerate base points $q_j \in X_j$. The map*

$$\Phi \colon \pi_1(X_1, q_1) * \cdots * \pi_1(X_n, q_n) \to \pi_1(X_1 \vee \cdots \vee X_n, *)$$

induced by ${\iota_j}_* \colon \pi_1(X_j, q_j) \to \pi_1(X_1 \vee \cdots \vee X_n, *)$ is an isomorphism.

Proof. First consider the wedge of two spaces $X_1 \vee X_2$. We would like to use Corollary 10.3 to the Seifert–Van Kampen theorem with $U = X_1$, $V = X_2$, and $U \cap V = \{q\}$. The trouble is that these spaces are not open in $X_1 \vee X_2$, so the corollary does not apply directly. To remedy this, we replace them by slightly "thicker" spaces using the nondegenerate base point condition.

Choose neighborhoods W_i in which q_i is a strong deformation retract, and let $U = \pi(X_1 \amalg W_2)$, $V = \pi(X_2 \amalg W_1)$, where $\pi \colon X_1 \amalg X_2 \to X_1 \vee X_2$ is the quotient map (Figure 10.1). Since $X_1 \amalg W_2$ and $X_2 \amalg W_1$ are saturated open sets in $X_1 \amalg X_2$, the restriction of π to each of them is a quotient map and U and V are open in the wedge.

The key fact is that the three inclusion maps

$$\begin{aligned} \{*\} &\hookrightarrow U \cap V, \\ X_1 &\hookrightarrow U, \\ X_2 &\hookrightarrow V \end{aligned}$$

are all homotopy equivalences, because each subspace on the left-hand side above is a strong deformation retract of the corresponding right-hand side. For $U \cap V$, this follows immediately from the preceding lemma. For U, choose a strong deformation retraction $H_2 \colon W_2 \times I \to W_2$ of W_2 onto q_2, and define $G_1 \colon X_1 \amalg W_2 \times I \to X_1 \amalg W_2$ to be the identity on $X_1 \times I$ and H_2 on $W_2 \times I$; it descends to a strong deformation retraction of U onto X_1. A similar construction shows $V \simeq X_2$.

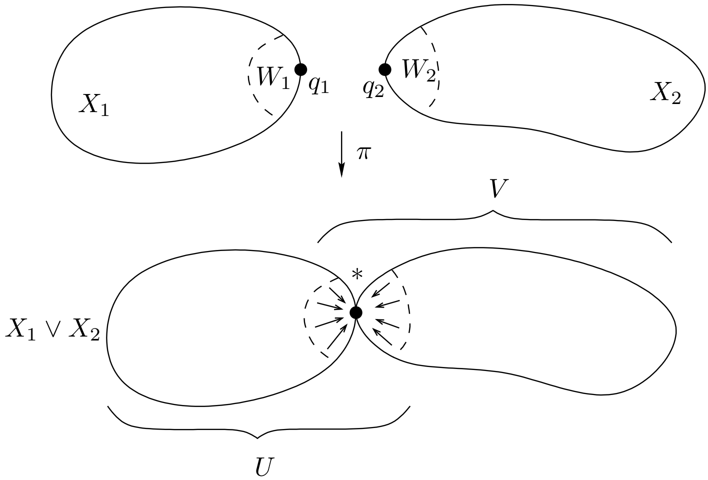


FIGURE 10.1. Computing the fundamental group of a wedge.

Because $U \cap V$ is contractible, Corollary 10.3 implies that the inclusion maps $U \hookrightarrow X_1 \vee X_2$ and $V \hookrightarrow X_1 \vee X_2$ induce an isomorphism

$$\pi_1(U, *) * \pi_1(V, *) \to \pi_1(X_1 \vee X_2, *).$$

Moreover, the injections $\iota_1 \colon X_1 \hookrightarrow U$ and $\iota_2 \colon X_2 \hookrightarrow V$, which are homotopy equivalences, induce isomorphisms $\pi_1(X_1, q_1) \to \pi_1(U, *)$ and $\pi_1(X_2, q_2) \to \pi_1(V, *)$. Composing these isomorphisms proves the proposition in the case $n = 2$. The case of $n > 2$ spaces follows by induction, because Lemma 10.4 guarantees that the hypotheses of the proposition are satisfied by X_1 and $X_2 \vee \cdots \vee X_n$. □

Example 10.6. The preceding proposition shows that the bouquet $\mathbb{S}^1 \vee \cdots \vee \mathbb{S}^1$ of n circles has fundamental group isomorphic to $\mathbb{Z} * \cdots * \mathbb{Z}$, which is a free group on n generators. In fact, it shows more: Since the isomorphism is induced by inclusion of each copy of $\mathbb{S}^1$ into the bouquet, we can write explicit generators of this free group. If α_i denotes the standard loop in the ith copy of $\mathbb{S}^1$, then the fundamental group of the bouquet is just the free group $\langle [\alpha_1], \ldots, [\alpha_n] \rangle$.

As a second application, we will compute the fundamental group of a finite graph. Let Γ be a finite connected graph, and choose a vertex v as base point. A *maximal tree* in Γ is a subgraph that is a tree and is not contained in any larger tree. The vertex v is contained in a (nonunique) maximal tree: Just start with a single edge containing v and keep adding edges until it is impossible to add another edge and still obtain a tree. Let $T \subset \Gamma$ be such a maximal tree.

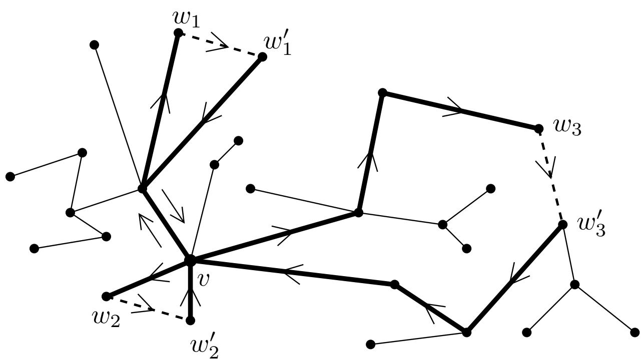


FIGURE 10.2. Generators for the fundamental group of a graph.

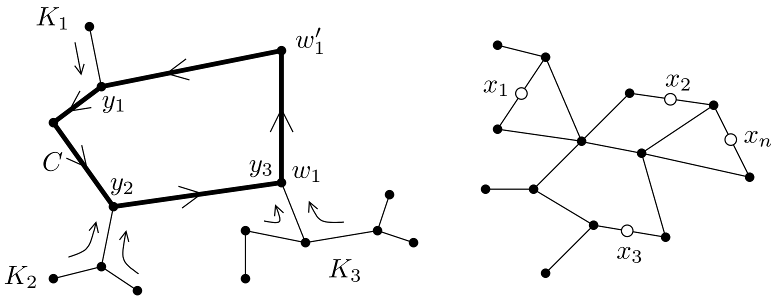
We will construct a set of generators for the fundamental group of Γ as follows. Let $\langle w_1, w_1' \rangle, \dots, \langle w_n, w_n' \rangle$ be the edges of Γ that are not in T (Figure 10.2). For each i, by maximality of T there is a reduced cycle in $T \cup \langle w_i, w_i' \rangle$ that does not lie in T. This cycle must therefore traverse the edge $\langle w_i, w_i' \rangle$, which implies that both of the vertices w_i and w_i' also belong to edges in T. Thus we can choose paths g_i and h_i in T from v to w_i and w_i', respectively. Let f_i denote the loop in Γ obtained by first following g_i from v to w_i, then traversing $\langle w_i, w_i' \rangle$, and then following h_i^{-1} from w_i' back to v. Note that the path class $[f_i]$ is independent of the choices of g_i and h_i, because any two paths in T with the same endpoints are path homotopic.

Theorem 10.7 (Fundamental Group of a Finite Graph). *The fundamental group of a finite connected graph Γ based at a vertex v is the free group on the path classes $[f_1], \dots, [f_n]$ constructed above.*

Proof. We prove the theorem by induction on n. If $n = 0$, then Γ is a tree and hence simply connected, so there is nothing to prove.

For $n = 1$, we must show that Γ is the infinite cyclic group generated by $[f_1]$. By assumption, there is a reduced cycle $(v_0, \dots, v_m)$ in Γ (Figure 10.3(a)). If $v_i = v_j$ for some $i \neq j$ other than $v_0 = v_m$, we can replace the cycle by the shorter one $(v_i, \dots, v_j)$. So we may assume that no vertex is repeated in this cycle except $v_0 = v_m$. This cycle must traverse the edge $\langle w_1, w_1' \rangle$, because otherwise it would be a reduced cycle in T. The subgraph $C \subset \Gamma$ consisting of the vertices $\{v_0, \dots, v_m\}$ and the intervening edges is homeomorphic to $\mathbb{S}^1$ because it is isomorphic as a simplicial complex to the complex $\mathcal{K}_m$ of Example 5.7(d). We will show that inclusion $C \subset \Gamma$ is a homotopy equivalence.

Let K be the union of all the edges in $\Gamma \smallsetminus C$ together with their vertices. Each component K_i of K is a connected subgraph of Γ contained in T, and is therefore a tree (since a reduced cycle in K_i would also be one in T). Moreover, each such component shares exactly one vertex y_i with



(a) $n = 1$. (b) $n > 1$.

FIGURE 10.3. Proof that the fundamental group of a graph is free.

C: If $K_i \cap C$ contained two vertices y_i, y_i', it would be possible to find a reduced cycle in T by following a reduced edge path in K_i from y_i to y_i' followed by the reduced edge path in C from y_i' to y_i that does not pass through $\langle w_1, w_1' \rangle$. (There must be at least one vertex in common because Γ is connected.)

Now define a strong deformation retraction of Γ onto C as follows: On each K_i, it is a strong deformation retraction of K_i onto y_i, which exists by Problem 7-13; and on C it is the identity. The resulting map is continuous by the gluing lemma, and shows that $\Gamma \simeq \mathbb{S}^1$.

It remains to show that the path class $[f_1]$ is a generator of $\pi_1(\Gamma, v)$. Let z be any vertex in C. The path a that starts at z and traverses each edge of C in order at constant speed is clearly path homotopic to the standard generator of $\mathbb{S}^1 \approx C$ (or its inverse). Choosing any path b from z to v yields an isomorphism $\Phi_b \colon \pi_1(\Gamma, z) \to \pi_1(\Gamma, v)$ as in Theorem 7.11. Thus a generator of $\pi_1(\Gamma, v)$ is $\Phi_b[a] = [b^{-1} \cdot a \cdot b]$. Since $b^{-1} \cdot a \cdot b$ is a path that goes from v to w_1, traverses $\langle w_1, w_1' \rangle$, and returns to v, it is homotopic to f_1. (Remember that the path class of f_1 is independent of which paths we choose from v to w_1 and w_1'.) This completes the proof in the case $n = 1$.

Now suppose $n > 1$, and assume that the theorem is true whenever there are fewer than n edges in the complement of a maximal tree. We will apply the Seifert–Van Kampen theorem in the following way. For each $i = 1, \dots, n$, choose a point x_i in the interior of the edge $\langle w_i, w_i' \rangle$ (Figure 10.3(b)). Let $U = \Gamma \setminus \{x_1, \dots, x_{n-1}\}$ and $V = \Gamma \setminus \{x_n\}$. Both U and V are open in Γ, and just as before it is easy to construct deformation retractions to show that $U \cap V \simeq T$, $U \simeq T \cup \langle w_n, w_n' \rangle$, and $V \simeq \Gamma \setminus \operatorname{Int}\langle w_n, w_n' \rangle$. By the inductive hypothesis, $\pi_1(V, v) = \langle [f_1], \dots, [f_{n-1}] \rangle$ and $\pi_1(U, v) = \langle [f_n] \rangle$. Since $U \cap V$ is simply connected, Corollary 10.3 shows that $\pi_1(\Gamma, v)$ is isomorphic to the free product of these two free groups, which is the free group on $[f_1], \dots, [f_n]$ as claimed. $\square$

Second Special Case: One Simply Connected Set

The other special case of the Seifert–Van Kampen theorem we will use is that in which one of the open sets, say U, is simply connected. In that case, diagram (10.1) simplifies considerably. Because the top group $\pi_1(U,q)$ is trivial, both the homomorphisms i_* and k_* are trivial, and the free product in the middle reduces to $\pi_1(V,q)$. Moreover, the homomorphism Φ is just equal to l_*, and the map F is just equal to j_*, so the entire diagram collapses to

$$\pi_1(U\cap V,q) \xrightarrow{j_*} \pi_1(V,q) \xrightarrow{l_*} \pi_1(X,q).$$

The conclusion of the theorem reduces immediately to the following corollary.

Corollary 10.8. *Assume the hypotheses of the Seifert–Van Kampen theorem, and suppose in addition that U is simply connected. Then inclusion $l\colon V \hookrightarrow X$ induces an isomorphism*

$$\pi_1(X,q) \cong \pi_1(V,q) \Big/ \overline{j_*\pi_1(U\cap V,q)},$$

where $\overline{j_*\pi_1(U\cap V,q)}$ is the normal closure of $j_*\pi_1(U\cap V,q)$ in $\pi_1(V,q)$. If the fundamental groups of V and $U\cap V$ have finite presentations

$$\pi_1(V,q) \cong \langle \beta_1,\dots,\beta_n \mid \sigma_1,\dots,\sigma_s\rangle,$$

$$\pi_1(U\cap V,q) \cong \langle \gamma_1,\dots,\gamma_p \mid \tau_1,\dots,\tau_t\rangle,$$

then $\pi_1(X,q)$ has the presentation

$$\pi_1(X,q) \cong \langle \beta_1,\dots,\beta_n \mid \sigma_1,\dots,\sigma_s, v_1,\dots,v_p\rangle,$$

where v_a is an expression for $j_*\gamma_a \in \pi_1(V,q)$ in terms of $\{\beta_1,\dots,\beta_m\}$.

We will give two applications of this corollary. The first is to give another proof that $\mathbb{S}^n$ is simply connected when $n \geq 2$.

Another proof of Theorem 8.7. As in the previous proof, we take $\mathbb{S}^n = U\cup V$ with $U = \mathbb{S}^n \smallsetminus \{N\}$ and $V = \mathbb{S}^n \smallsetminus \{S\}$, both of which are simply connected because they are homeomorphic to $\mathbb{R}^n$. Moreover, $U\cap V$ is path connected because it is homeomorphic to $\mathbb{R}^n \smallsetminus \{0\}$. Corollary 10.8 to the Seifert–Van Kampen theorem says that for any point $q \in U\cap V$, $\pi_1(\mathbb{S}^n,q)$ is isomorphic to a quotient of $\pi_1(V,q)$ by a certain subgroup. But this quotient is trivial because $\pi_1(V,q)$ is itself trivial. □

The next proposition will allow us to compute the fundamental groups of all compact surfaces. Now it will become clear why we chose similar notations for surface presentations and group presentations.

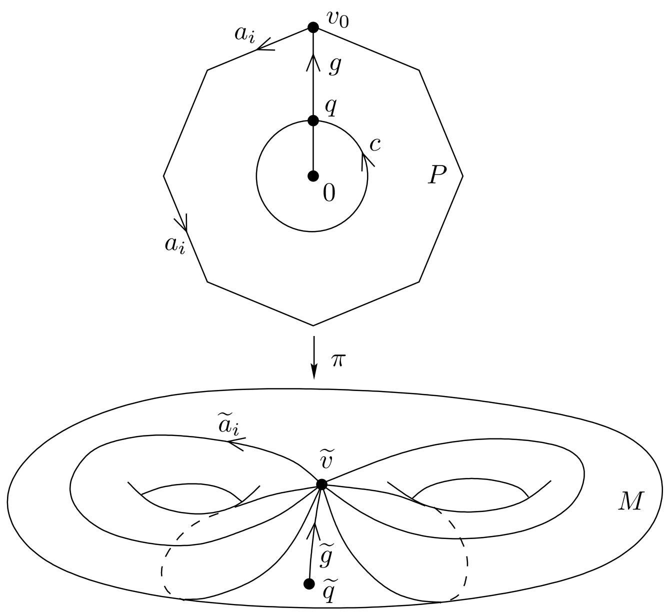


FIGURE 10.4. Proof of Proposition 10.9.

Proposition 10.9. *Let M be a 2-manifold determined by a polygonal surface presentation $\langle a_1, \ldots, a_n \mid W \rangle$ with one face, in which all vertices are identified to a single point. Then $\pi_1(M)$ has the presentation $\langle a_1, \ldots, a_n \mid W \rangle$.*

Proof. First we set up some notation (see Figure 10.4). Let P be a regular polygon in the plane with $2n$ sides, centered at the origin, and let $\pi\colon P \to M$ denote the quotient map determined by the given presentation. Set $U = \operatorname{Int} P$ and $V = P \smallsetminus \{0\}$, and let $\widetilde{U} = \pi(U)$, $\widetilde{V} = \pi(V) \subset M$. Since U and V are saturated open sets, the restrictions of π to U and V are quotient maps, and their images $\widetilde{U}, \widetilde{V}$ are open in M. Moreover, $\widetilde{U}$, $\widetilde{V}$, and $\widetilde{U} \cap \widetilde{V}$ are all path connected because they are images of path connected sets in P. Choose a base point $q \in U \cap V$ lying on the line segment from 0 to the vertex v_0 on the y-axis, and let $\widetilde{q} = \pi(q) \in \widetilde{U} \cap \widetilde{V}$. Finally, let $\widetilde{v} \in M$ denote the single vertex (the image of all the vertices of P). In general, we will use symbols without tildes to denote sets, points, or paths in P, and the same symbols with tildes to denote their images in M.

The restriction of π to U is a one-to-one quotient map and therefore a homeomorphism. Since U is convex, it is simply connected, and therefore $\widetilde{U}$ is simply connected as well. Thus we will be able to use Corollary 10.2 once we find presentations for the fundamental groups of $\widetilde{V}$ and $\widetilde{U} \cap \widetilde{V}$. The details are a bit involved, but the basic idea is just that $\widetilde{V}$ is homotopy equivalent to a bouquet of circles, so $\pi_1(\widetilde{V},\widetilde{q})$ is a free group on the n generators determined by $a_1,\dots,a_n$; and $\widetilde{U}\cap\widetilde{V}$ is homotopy equivalent to a circle, so $j_*\pi_1(\widetilde{U}\cap\widetilde{V},\widetilde{q})$ is the infinite cyclic group generated by the word W in these generators.

Consider first $\widetilde{V}$. Observe that its preimage $V\subset P$ is homotopy equivalent to ∂P, by the straight-line homotopy H that moves each point radially outward to the boundary. The map $\pi\times\mathrm{Id}\colon P\times I\to M\times I$ is a quotient map by Lemma 4.35 (or just by the closed map lemma). Since $V\times I$ is a saturated open subset of $P\times I$, the restriction of $\pi\times\mathrm{Id}$ to it is a quotient map. Because $\pi\circ H$ respects the identifications of this quotient map, it descends to a strong deformation retraction $\widetilde{H}$ of $\widetilde{V}$ onto $\pi(\partial P)$:

$$
\begin{tikzcd}
V\times I \arrow[r, "H"] \arrow[d, "\pi\times\mathrm{Id}"'] & V \arrow[d, "\pi"] \\
\widetilde{V}\times I \arrow[r, "\widetilde{H}"'] & \widetilde{V}.
\end{tikzcd}
$$

On ∂P, π identifies the edges in pairs and identifies all the vertices to a point, so $\pi(\partial P)$ is homeomorphic to a bouquet of n circles, one for each label a_i. (To verify this, you can, for example, construct quotient maps from $2n$ disjoint intervals to $\mathbb{S}^1\vee\cdots\vee\mathbb{S}^1$ and to $\pi(\partial P)$, both of which make the same identifications.) Therefore, $\pi_1(\widetilde{V},q)$ is a free group on n generators.

We need to identify these generators explicitly. For each i, let $\widetilde{a}_i$ denote a path in M that traverses the single edge labeled a_i (the image under π of two edges of P) in the indicated direction. Since all vertices of P project to $\widetilde{v}$, $\widetilde{a}_i$ is a loop based at $\widetilde{v}$. From Example 10.6, $\pi_1(\mathbb{S}^1\vee\cdots\vee\mathbb{S}^1,\widetilde{v})$ is freely generated by $[\widetilde{a}_1],\dots,[\widetilde{a}_n]$. The isomorphism $\pi_1(\mathbb{S}^1\vee\cdots\vee\mathbb{S}^1,\widetilde{v})\to\pi_1(\widetilde{V},\widetilde{v})$ induced by inclusion takes these generators to themselves, thought of as loops in $\widetilde{V}$, so $\pi_1(\widetilde{V},\widetilde{v})$ is the free group $\langle[\widetilde{a}_1],\dots,[\widetilde{a}_n]\rangle$.

We are really interested in the base point $\widetilde{q}$, not $\widetilde{v}$. Let g be the radial straight-line path from q to v_0, let $\widetilde{g}=\pi\circ g$, a path from $\widetilde{q}$ to $\widetilde{v}$, and consider the loops $\widetilde{a}'_i$ based at $\widetilde{q}$, defined by

$$\widetilde{a}'_i=\widetilde{g}\cdot\widetilde{a}_i\cdot\widetilde{g}^{-1}.$$

These are just the images of the loops $\widetilde{a}_i$ under the isomorphism from $\pi_1(\widetilde{V},\widetilde{v})$ to $\pi_1(\widetilde{V},\widetilde{q})$ provided by the path $\widetilde{g}^{-1}$ as in Theorem 7.11, so we conclude that $\pi_1(\widetilde{V},\widetilde{q})$ is the free group $\langle[\widetilde{a}'_1],\dots,[\widetilde{a}'_n]\rangle$.

Next we turn to $\widetilde{U}\cap\widetilde{V}$. Since π is one-to-one on $U\cap V$, $\widetilde{U}\cap\widetilde{V}$ is homeomorphic to $U\cap V=\operatorname{Int}P\smallsetminus\{0\}$. Let C be the circle centered at 0 and passing through q, and let $\widetilde{C}$ be its image in M. (We may assume that q is chosen close enough to 0 that C is contained in the interior of P.) A deformation retraction of $\operatorname{Int}P\smallsetminus\{0\}$ onto C is easily constructed by moving each point radially inward or outward toward C at constant speed. Thus

$\pi_1(\widetilde{U} \cap \widetilde{V}, \widetilde{q})$ is infinite cyclic, generated by the path class of any loop that goes around $\widetilde{C}$ once. Let c be a loop that goes counterclockwise around C, and let $\widetilde{c} = \pi \circ c$, a loop in $\widetilde{U} \cap \widetilde{V}$ based at $\widetilde{q}$. Then $\pi_1(\widetilde{U} \cap \widetilde{V}, \widetilde{q})$ is the infinite cyclic group $\langle [\widetilde{c}] \rangle$.

Now Corollary 10.8 to the Seifert–Van Kampen theorem says that $\pi_1(M, \widetilde{q}) \cong \langle [\widetilde{a}'_1], \ldots, [\widetilde{a}'_n] \mid b \rangle$, where b is an expression for $j_*[\widetilde{c}] \in \pi_1(\widetilde{V}, \widetilde{q})$ in terms of the generators $[\widetilde{a}'_1], \ldots, [\widetilde{a}'_n]$. So to complete the proof, we need only find such an expression.

The key observation is that $\widetilde{g}^{-1} \cdot \widetilde{c} \cdot \widetilde{g}$, a loop based at $\widetilde{v}$, is path homotopic in $\widetilde{V}$ to the loop $\widetilde{W}$ obtained from W by replacing each a_i by $\widetilde{a}_i$. To see this, we first work upstairs in P: Let H be the strong deformation retraction of V onto ∂P, and set

$$F(s,t) = H(g^{-1} \cdot c \cdot g(s), t).$$

The map F is a homotopy from $g^{-1} \cdot c \cdot g$ to a path that goes once around ∂P in the clockwise direction, and it descends to a homotopy in $\widetilde{V}$ from $\widetilde{g}^{-1} \cdot \widetilde{c} \cdot \widetilde{g}$ to $\widetilde{W}$. It follows that

$$\widetilde{c} \sim \widetilde{g} \cdot \widetilde{g}^{-1} \cdot \widetilde{c} \cdot \widetilde{g} \cdot \widetilde{g}^{-1} \sim \widetilde{g} \cdot \widetilde{W} \cdot \widetilde{g}^{-1} \sim \widetilde{W}',$$

where $\widetilde{W}'$ is obtained from W by replacing each a_i by $\widetilde{a}'_i$; the last equivalence follows by inserting $\widetilde{g}^{-1} \cdot \widetilde{g}$ between each pair of symbols in the word $\widetilde{W}$.

Thus we have shown that the generator $[\widetilde{c}]$ of $\pi_1(\widetilde{U} \cap \widetilde{V}, \widetilde{q})$ is mapped by inclusion to $[\widetilde{W}'] \in \pi_1(\widetilde{V}, \widetilde{q})$, where $[\widetilde{W}']$ is obtained from W by replacing each a_i by $[\widetilde{a}'_i]$. By Corollary 10.2, therefore, $\pi_1(M, \widetilde{q})$ has the presentation $\langle [\widetilde{a}'_1], \ldots, [\widetilde{a}'_n] \mid [\widetilde{W}'] \rangle$; relabeling the symbols a_i instead of $[\widetilde{a}'_i]$ yields the presentation given in the statement of the proposition. □

Example 10.10. Using the results of this section, we have the following presentations for the fundamental groups of compact, connected surfaces:

(a) $\pi_1(\mathbb{S}^2) \cong \langle\ \rangle$ (the trivial group).

(b) $\pi_1(\mathbb{T}^2 \# \cdots \# \mathbb{T}^2)$
$\cong \langle \beta_1, \gamma_1, \ldots, \beta_n, \gamma_n \mid \beta_1\gamma_1\beta_1^{-1}\gamma_1^{-1} \cdots \beta_n\gamma_n\beta_n^{-1}\gamma_n^{-1} = 1 \rangle$.

(c) $\pi_1(\mathbb{P}^2 \# \cdots \# \mathbb{P}^2) \cong \langle \beta_1, \ldots, \beta_n \mid \beta_1^2 \cdots \beta_n^2 = 1 \rangle$.

In particular, for the torus this gives $\pi_1(\mathbb{T}^2) \cong \langle \beta, \gamma \mid \beta\gamma = \gamma\beta \rangle$, which agrees with the result we derived earlier. In the case of the projective plane, this gives $\pi_1(\mathbb{P}^2) \cong \langle \beta \mid \beta^2 = 1 \rangle \cong \mathbb{Z}/\langle 2 \rangle$.

Proof of the Theorem

In this section we prove the Seifert–Van Kampen theorem (Theorem 10.1) and its corollary about finite presentations (Corollary 10.2).

Proof of Theorem 10.1. Because we will be considering paths and their homotopy classes in various spaces, for this proof we will refine our notation to specify explicitly where homotopies are assumed to lie. If a and b are paths in X that happen to lie in one of the subsets U, V, or $U \cap V$, we will use the notation

$$a \underset{U}{\sim} b, \quad a \underset{V}{\sim} b, \quad a \underset{U\cap V}{\sim} b, \quad a \underset{X}{\sim} b$$

to indicate that a is path homotopic to b in U, V, $U \cap V$, or X, respectively. We write $[a]_U$ for the path class of a in $\pi_1(U,q)$, and similarly for the other sets. Thus, for example, if a is a loop in $U \cap V$, the homomorphisms induced by the inclusions $i\colon U \cap V \hookrightarrow U$ and $k\colon U \hookrightarrow X$ can be written

$$i_*([a]_{U\cap V}) = [a]_U,$$

$$k_*([a]_U) = [a]_X.$$

We will have to consider two different types of products: path class multiplication within any one fundamental group, and word multiplication in the free product group. As usual, we will denote path and path class multiplication by a dot, as in

$$[a]_U \cdot [b]_U = [a \cdot b]_U.$$

To emphasize the distinction between the two products, we denote multiplication in the free product group by an asterisk, so, for example,

$$[a]_U * [b]_U * [c]_V = [a \cdot b]_U * [c]_V \in \pi_1(U,q) * \pi_1(V,q).$$

Then the map $\Phi\colon \pi_1(U,q) * \pi_1(V,q) \to \pi_1(X,q)$ can be written

$$\begin{aligned}
&\Phi([a_1]_U * [a_2]_V * \cdots * [a_{m-1}]_U * [a_m]_V) \\
&\quad = k_*[a_1]_U \cdot l_*[a_2]_V \cdot \cdots \cdot k_*[a_{m-1}]_U \cdot l_*[a_m]_V \\
&\quad = [a_1]_X \cdot [a_2]_X \cdot \cdots \cdot [a_{m-1}]_X \cdot [a_m]_X \\
&\quad = [a_1 \cdot a_2 \cdot \cdots \cdot a_{m-1} \cdot a_m]_X.
\end{aligned} \tag{10.3}$$

Let N denote the normal closure of $F(\pi_1(U \cap V, q))$ in $\pi_1(U,q) * \pi_1(V,q)$. We need to prove three things: (1) Φ is surjective; (2) $N \subset \operatorname{Ker}\Phi$; and (3) $\operatorname{Ker}\Phi \subset N$.

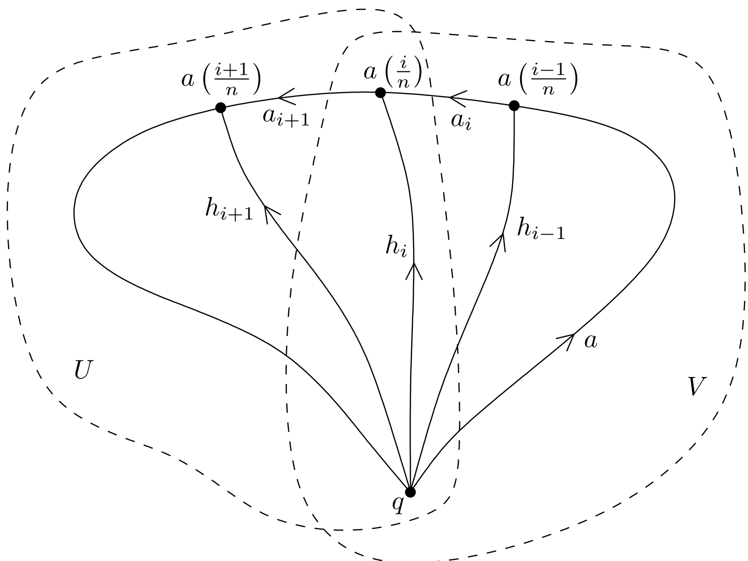


FIGURE 10.5. Proof that Φ is surjective.

STEP 1: Φ *is surjective.* Let $a\colon I \to X$ be any loop in X based at q. By the Lebesgue number lemma, we can choose n large enough that a maps each subinterval $[(i-1)/n, i/n]$ either into U or into V. (This is why it is important that the sets U and V be open.) Letting a_i denote the restriction of a to $[(i-1)/n, i/n]$ (reparametrized so that its parameter interval is I), the path class of a in X factors as

$$[a]_X = [a_1 \cdot \cdots \cdot a_n]_X.$$

The problem with this factorization is that the paths a_i are not loops in general. To remedy this, for each $i = 1, \ldots, n-1$, choose a path h_i from q to $a(i/n)$ (Figure 10.5). If $a(i/n) \in U \cap V$, choose h_i to lie entirely in $U \cap V$; otherwise, choose it to lie in whichever set U or V contains $a(i/n)$. (This is why the sets U, V, and $U \cap V$ must all be path connected.) Then set $\widetilde{a}_i = h_{i-1} \cdot a_i \cdot h_i^{-1}$ (where we let h_0 and h_n be the constant loop c_q), so that each $\widetilde{a}_i$ is a loop based at q and lying entirely in either U or V. It follows easily that a also factors as

$$[a]_X = [\widetilde{a}_1 \cdot \cdots \cdot \widetilde{a}_n]_X.$$

Now consider the element

$$\gamma = [\widetilde{a}_1]_U * [\widetilde{a}_2]_V * \cdots * [\widetilde{a}_n]_V \in \pi_1(U,q) * \pi_1(V,q),$$

where we choose either U or V for each $\widetilde{a}_i$ depending on which set contains its image. Then as in (10.3) above,

$$\Phi(\gamma) = [\widetilde{a}_1 \cdot \cdots \cdot \widetilde{a}_n]_X = [a]_X.$$

This proves that Φ is surjective.

STEP 2: *$N \subset \operatorname{Ker}\Phi$.* If we can show that $F(\pi_1(U \cap V, q))$ is contained in $\operatorname{Ker}\Phi$, then its normal closure N will be contained in $\operatorname{Ker}\Phi$ as well because $\operatorname{Ker}\Phi$ is normal.

Let $[a]_{U\cap V} \in \pi_1(U \cap V, q)$ be arbitrary. Then

$$\begin{aligned}
\Phi \circ F([a]_{U\cap V}) &= \Phi((i_*[a]_{U\cap V})^{-1} * (j_*[a]_{U\cap V})) \\
&= \Phi([a^{-1}]_U * [a]_V) \\
&= [a^{-1} \cdot a]_X \\
&= 1.
\end{aligned}$$

STEP 3: *$\operatorname{Ker}\Phi \subset N$.* This is the crux of the proof. Let

$$\gamma = [a_1]_U * [a_2]_V * \cdots * [a_k]_V \in \pi_1(U, q) * \pi_1(V, q)$$

be an arbitrary element of the free product, and suppose that $\Phi(\gamma) = 1$. Using (10.3) again, this means that

$$[a_1 \cdot \cdots \cdot a_k]_X = 1,$$

which is equivalent to

$$a_1 \cdot \cdots \cdot a_k \underset{X}{\sim} c_q.$$

We need to show that $\gamma \in N$.

Let $H\colon I \times I \to X$ be the path homotopy from $a_1 \cdot \cdots \cdot a_k$ to c_q in X. By the Lebesgue number lemma again, we can subdivide $I \times I$ into squares of side $1/n$ so that H maps each square $S_{ij} = [(i-1)/n, i/n] \times [(j-1)/n, j/n]$ either into U or into V.

Let v_{ij} denote the image under H of the vertex $(i/n, j/n)$; and let a_{ij} denote the restriction of H to the horizontal line segment $[(i-1)/n, i/n] \times \{j/n\}$, and b_{ij} the restriction to the vertical segment $\{i/n\} \times [(j-1)/n, j/n]$, both suitably reparametrized on I (see Figure 10.6).

The restriction of H to the bottom edge of $I \times I$, where $t = 0$, is equal to the path product $a_1 \cdot \cdots \cdot a_k$. By taking n to be a sufficiently large power of 2, we can ensure that the endpoints of the paths a_i in this product are

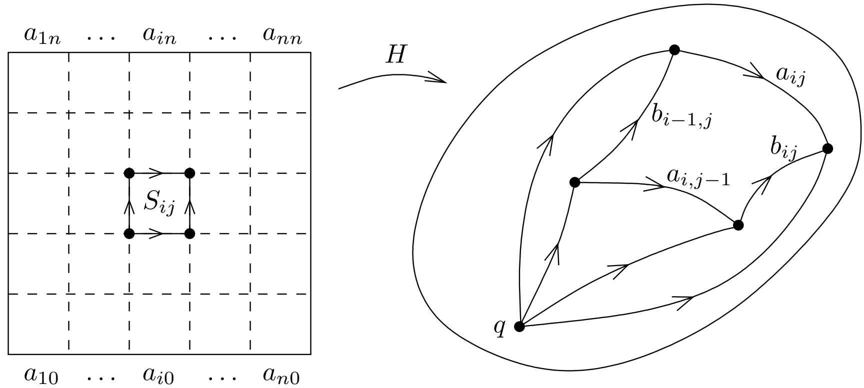


FIGURE 10.6. Proof that $\operatorname{Ker}\Phi \subset N$.

of the form i/n, so the path obtained by restricting H to the bottom edge of the square can also be written

$$H_0 \sim a_1 \cdot \cdots \cdot a_k \sim (a_{10} \cdot \cdots \cdot a_{p0}) \cdot \cdots \cdot (a_{r0} \cdot \cdots \cdot a_{n0}).$$

In the free product, this means that

$$\gamma = [a_{10} \cdot \cdots \cdot a_{p0}]_U * \cdots * [a_{r0} \cdot \cdots \cdot a_{n0}]_V.$$

We would like to factor this in the free product as $[a_{10}]_U * [a_{20}]_U * \cdots$ and so forth. But these paths are not loops based at q, so we cannot yet use this relation directly. This is easy to fix as in Step 1: For each i and j, choose a path h_{ij} from q to v_{ij}, staying in $U \cap V$ if $v_{ij} \in U \cap V$, and otherwise in U or V; if v_{ij} happens to be the base point q, choose h_{ij} to be the constant loop c_q. Then define loops

$$\widetilde{a}_{ij} = h_{i-1,j} \cdot a_{ij} \cdot h_{ij}^{-1}, \qquad \widetilde{b}_{ij} = h_{i,j-1} \cdot b_{ij} \cdot h_{ij}^{-1}, \tag{10.4}$$

each of which lies entirely in U or V. Then γ can be factored as

$$\gamma = [\widetilde{a}_{10}]_U * [\widetilde{a}_{20}]_U * \cdots * [\widetilde{a}_{n0}]_V. \tag{10.5}$$

The main idea of the proof is this: We will show that modulo N, the expression (10.5) for γ can be replaced by the corresponding expression obtained by restricting H to the top edge of the first row of squares:

$$\gamma \equiv [\widetilde{a}_{11}]_U * \cdots * [\widetilde{a}_{n1}]_V \pmod{N}.$$

Repeating this argument, we move up to the next row, and so forth by induction, until we obtain

$$\gamma \equiv [\widetilde{a}_{1n}]_U * \cdots * [\widetilde{a}_{nn}]_V \pmod N.$$

But the entire top edge of $I \times I$ is mapped by H to the point q, so each $\widetilde{a}_{in}$ is equal to the constant loop c_q, and this last product is equal to the identity. This shows that $\gamma \in N$, completing the proof.

Thus we need to prove the following inductive step: Assuming by induction that

$$\gamma \equiv [\widetilde{a}_{1,j-1}]_U * \cdots * [\widetilde{a}_{n,j-1}]_V \pmod N, \tag{10.6}$$

we need to show that γ is equivalent modulo N to the same expression with $j-1$ replaced by j.

First we observe the following simple fact: Suppose a is a loop in $U \cap V$. Then $[a]_U$ and $[a]_V$ are in the same coset in the free product modulo N, because

$$[a]_V * N = [a]_U * ([a]_U^{-1} * [a]_V) * N = [a]_U * F([a]_{U\cap V}) * N = [a]_U * N.$$

Since N is normal, this also implies $x * [a]_U * y * N = x * [a]_U * N * y = x * [a]_V * N * y = x * [a]_V * y * N$ for any x, y in the free product. Thus, as long as we are computing modulo N and a is a loop in $U \cap V$, we can freely interchange $[a]_U$ with $[a]_V$ wherever either appears.

Consider a typical square S_{ij}, and suppose for definiteness that H maps S_{ij} into V. The boundary of S_{ij}, traversed clockwise starting at the lower left corner, is mapped to the path $(b_{i-1,j} \cdot a_{ij}) \cdot (b_{ij}^{-1} \cdot a_{i,j-1}^{-1})$. By Lemma 7.12, this means that

$$a_{i,j-1} \underset{V}{\sim} b_{i-1,j} \cdot a_{ij} \cdot b_{ij}^{-1}. \tag{10.7}$$

Using the definition (10.4) of the loops $\widetilde{a}_{ij}$ and $\widetilde{b}_{ij}$, (10.7) yields

$$\begin{aligned} \widetilde{a}_{i,j-1} &= h_{i-1,j-1} \cdot a_{i,j-1} \cdot h_{i,j-1}^{-1} \\ &\underset{V}{\sim} h_{i-1,j-1} \cdot b_{i-1,j} \cdot a_{ij} \cdot b_{ij}^{-1} \cdot h_{i,j-1}^{-1} \\ &\underset{V}{\sim} \widetilde{b}_{i-1,j} \cdot \widetilde{a}_{ij} \cdot \widetilde{b}_{ij}^{-1}, \end{aligned} \tag{10.8}$$

since the interior factors of h_{ij} and $h_{i-1,j}$ and their inverses cancel out.

Now start with the expression (10.6) for γ. For each factor $[\widetilde{a}_{i,j-1}]_U$, check whether the square S_{ij} above it is mapped into U or V. If it is mapped into V, then $\widetilde{a}_{i,j-1}$ must map into $U \cap V$, and we can replace this factor by $[\widetilde{a}_{i,j-1}]_V$ modulo N. Correct each factor whose square maps into U similarly.

By (10.8), we can replace each such factor $[\widetilde{a}_{i,j-1}]_V$ by $[\widetilde{b}_{i-1,j}]_V * [\widetilde{a}_{ij}]_V * [\widetilde{b}_{ij}]_V^{-1}$, and similarly for the factors in U. Thus

$$\begin{aligned}\gamma &\equiv [\widetilde{b}_{0,j}]_U * [\widetilde{a}_{1j}]_U * [\widetilde{b}_{1j}]_U^{-1} * \cdots * [\widetilde{b}_{n-1,j}]_V * [\widetilde{a}_{nj}]_V * [\widetilde{b}_{nj}]_V^{-1} \pmod N\\ &\equiv [\widetilde{a}_{1j}]_U * \cdots * [\widetilde{a}_{nj}]_V \pmod N.\end{aligned}$$

Here we have used the facts that the interior $\widetilde{b}_{ij}$ factors all cancel each other out (replacing $[\widetilde{b}_{ij}]_U$ by $[\widetilde{b}_{ij}]_V$ when necessary), and $\widetilde{b}_{0j}$ and $\widetilde{b}_{nj}$ are both equal to the constant loop c_q. This completes the inductive step and thus the proof. □

Proof of Corollary 10.2. To simplify the notation, let A and B denote the free groups $\langle \alpha_1, \dots, \alpha_m \rangle$ and $\langle \beta_1, \dots, \beta_n \rangle$, respectively, and write $R = \{\rho_1, \dots, \rho_r\}$, $S = \{\sigma_1, \dots, \sigma_s\}$, and $G = \{u_1^{-1}v_1, \dots, u_p^{-1}v_p\}$, all considered as subsets of $A * B$. (Note that the relators τ_i in $U \cap V$ do not enter into either the statement of the corollary or its proof.) Then $\pi_1(U,q) \cong A/\overline{R}$ and $\pi_1(V,q) \cong B/\overline{S}$ by hypothesis. We will consider these isomorphisms as identifications, so Φ is a map from $(A/\overline{R}) * (B/\overline{S})$ to $\pi_1(X,q)$. Let $\pi\colon A * B \to (A/\overline{R}) * (B/\overline{S})$ denote the homomorphism induced by the projections $A \to A/\overline{R}$ and $B \to B/\overline{S}$ as in Problem 9-4.

The Seifert–Van Kampen theorem says that Φ is surjective and has kernel equal to $\overline{F(\pi_1(U \cap V, q))}$, which clearly contains $\overline{\pi(G)}$ by our choice of the set G. In fact, it is equal to $\overline{\pi(G)}$: To see this, we need only verify that $\overline{\pi(G)}$ includes all of $F(\pi_1(U \cap V, q))$, which is to say every element of the form $(i_*\gamma)^{-1}(j_*\gamma)$ for $\gamma \in \pi_1(U \cap V, q)$. Consider the quotient group $\big((A/\overline{R}) * (B/\overline{S})\big)/\overline{\pi(G)}$. By definition of G, each element of the form $u_i^{-1}v_i$ projects to 1 in this quotient, which is to say that u_i and v_i project to the same element. Given any $\gamma \in \pi_1(U \cap V, q)$, express it as a word in the generators $\gamma_1, \dots, \gamma_p$; then $i_*\gamma$ and $j_*\gamma$ will be expressed by the same word, but with γ_i replaced by u_i in one case and v_i in the other. This shows that $i_*\gamma$ and $j_*\gamma$ project to the same element in the quotient, which means that $(i_*\gamma)^{-1}(j_*\gamma) \in \overline{\pi(G)}$.

The composition

$$A * B \xrightarrow{\pi} (A/\overline{R}) * (B/\overline{S}) \xrightarrow{\Phi} \pi_1(X,q)$$

is also surjective. The corollary will follow from the first isomorphism theorem (Theorem A.13) once we show that its kernel is exactly $\overline{R \cup S \cup G}$.

Clearly, R and S are contained in $\operatorname{Ker}(\Phi \circ \pi)$. Also, $\pi(G) \subset F(\pi_1(U \cap V, q)$ by our choice of G, and thus $\pi(G) \subset \operatorname{Ker}\Phi$, which means that $G \in \operatorname{Ker}(\Phi \circ$

π) as well. Therefore, the kernel of $\Phi \circ \pi$ is a normal subgroup containing $R \cup S \cup G$, so $\overline{R \cup S \cup G} \subset \operatorname{Ker}(\Phi \circ \pi)$.

To prove the reverse inclusion, suppose $w \in A * B$ is a word such that $\Phi \circ \pi(w) = 1$. This means that $\pi(w) \in \operatorname{Ker}\Phi = \overline{\pi(G)} = \pi(\overline{G})$. (The last equality follows because the homomorphic image of a normal subgroup under a surjective map is normal; see Exercise A.25 in the Appendix.) In other words, $\pi(w) = \pi(g)$, where g is some element of $\overline{G}$. This just means that $wg^{-1} \in \operatorname{Ker}\pi$, which by Problem 9-4 is equal to $\overline{R} * \overline{S}$ (thought of as a subgroup of $A * B$). This can be rewritten as $w = hg$ for some $h \in \overline{R} * \overline{S}$, $g \in \overline{G}$. Since clearly $\overline{R \cup S \cup G}$ must contain $\overline{R} * \overline{S}$ and $\overline{G}$, this means that $w \in \overline{R \cup S \cup G}$, and the proof is complete. □

Distinguishing Manifolds

Now we are finally in a position to fill the gap in our classification of surfaces by showing that the different surfaces on our list are actually topologically distinct. We will do so by showing that their fundamental groups are not isomorphic. Even this is not completely straightforward, because it involves solving the isomorphism problem for certain finitely presented groups. But in this case we can reduce the problem to a much simpler problem involving abelian groups.

Given a group G, the *commutator subgroup* of G, denoted by $[G, G]$, is the normal closure of the set of all elements of the form $\alpha\beta\alpha^{-1}\beta^{-1}$ for $\alpha, \beta \in G$. Clearly, the quotient group $G/[G, G]$ is always abelian, and the commutator subgroup is trivial if and only if G itself is abelian. This quotient group is denoted by $\operatorname{Ab}(G)$ and called the *abelianization* of G. It is clear that isomorphic groups have isomorphic abelianizations. $\operatorname{Ab}(G)$ is the “largest” abelian quotient of G, or equivalently the largest abelian homomorphic image of G, in the sense that any other homomorphism into an abelian group factors through the abelianization, as the following characteristic property shows.

Theorem 10.11 (Characteristic Property of Abelianizations). *For any abelian group H and any homomorphism $\varphi\colon G \to H$, there exists a unique homomorphism $\widetilde{\varphi}\colon \operatorname{Ab}(G) \to H$ such that the following diagram commutes:*

$$\begin{tikzcd} G \arrow[r, "\varphi"] \arrow[d, "\pi"'] & H \\ \operatorname{Ab}(G). \arrow[ur, "\widetilde{\varphi}"'] & \end{tikzcd}$$

Exercise 10.1. Prove Theorem 10.11.

It is relatively easy to compute the abelianizations of our surface groups.

Theorem 10.12. *The fundamental groups of compact surfaces have the following abelianizations:*

$$\mathrm{Ab}(\pi_1(\mathbb{S}^2)) = \{1\};$$

$$\mathrm{Ab}(\pi_1(\underbrace{\mathbb{T}^2 \# \cdots \# \mathbb{T}^2}_{n})) \cong \mathbb{Z}^{2n};$$

$$\mathrm{Ab}(\pi_1(\underbrace{\mathbb{P}^2 \# \cdots \# \mathbb{P}^2}_{n})) \cong \mathbb{Z}^{n-1} \times \mathbb{Z}/\langle 2\rangle.$$

Proof. The case of the sphere is obvious, so consider first an orientable surface of genus n, and let

$$G = \langle \beta_1, \gamma_1, \dots, \beta_n, \gamma_n \mid \beta_1\gamma_1\beta_1^{-1}\gamma_1^{-1} \cdots \beta_n\gamma_n\beta_n^{-1}\gamma_n^{-1}\rangle$$

be the fundamental group. Define a map $\varphi\colon \mathrm{Ab}(G) \to \mathbb{Z}^{2n}$ as follows. Let $e_i = (0, \dots, 1, \dots, 0) \in \mathbb{Z}^{2n}$ (1 in the ith place), and set

$$\varphi(\beta_i) = e_i, \quad \varphi(\gamma_i) = e_{i+n}.$$

Thought of as a map from the free group $\langle\{\beta_i, \gamma_i\}\rangle$ into $\mathbb{Z}^{2n}$, this sends the element $\beta_1\gamma_1\beta_1^{-1}\gamma_1^{-1} \cdots \beta_n\gamma_n\beta_n^{-1}\gamma_n^{-1}$ to $(0, \dots, 0)$, so it descends to a homomorphism from G to $\mathbb{Z}^{2n}$. By the characteristic property of the abelianization, it also descends to a homomorphism (still denoted by φ) from $\mathrm{Ab}(G)$ to $\mathbb{Z}^{2n}$.

To go back the other way, define $\psi\colon \mathbb{Z}^{2n} \to \mathrm{Ab}(G)$ by

$$\psi(e_i) = \begin{cases} [\beta_i], & 1 \le i \le n, \\ [\gamma_{i-n}], & n+1 \le i \le 2n, \end{cases}$$

where the brackets on the right-hand side denote the equivalence class in $\mathrm{Ab}(G)$, and extend it to be a homomorphism. It is easy to check that φ and ψ are inverses of each other.

Next consider the connected sum of projective planes, and let $H = \langle \beta_1, \dots, \beta_n \mid \beta_1^2 \cdots \beta_n^2\rangle$. Write f for the nontrivial element of $\mathbb{Z}/\langle 2\rangle$, and define $\varphi\colon \mathrm{Ab}(H) \to \mathbb{Z}^{n-1} \times \mathbb{Z}/\langle 2\rangle$ by

$$\varphi(\beta_i) = \begin{cases} e_i, & 1 \le i \le n-1; \\ f - e_{n-1} - \cdots - e_1, & i = n. \end{cases}$$

As before, $\varphi(\beta_1^2 \cdots \beta_n^2) = (0, \dots, 0)$ by direct computation (noting that $f+f = 0$), so φ descends to $\mathrm{Ab}(H)$. The homomorphism $\psi\colon \mathbb{Z}^{n-1} \times \mathbb{Z}/\langle 2\rangle \to \mathrm{Ab}(H)$ defined by

$$\psi(e_i) = [\beta_i], \quad \psi(f) = [\beta_1 \cdots \beta_n]$$

is easily verified to be an inverse for φ. □

Corollary 10.13. *Any compact, connected surface is homeomorphic to exactly one of the surfaces $\mathbb{S}^2$, $\mathbb{T}^2 \# \cdots \# \mathbb{T}^2$, or $\mathbb{P}^2 \# \cdots \# \mathbb{P}^2$.*

Proof. First note that the sphere cannot be homeomorphic to a connected sum of tori or projective planes, because one has trivial fundamental group and the other does not. Next, if M is a connected sum of projective planes, then $\operatorname{Ab}(\pi_1(M))$ contains a nontrivial torsion element, whereas the abelianized fundamental groups of connected sums of tori are torsion free. Therefore, no connected sum of projective planes can be homeomorphic to a connected sum of tori. If M is a connected sum of n tori, then its abelianized fundamental group has rank $2n$. Thus the genus (i.e., the number of tori in the connected sum) can be recovered from the fundamental group, so the genus of an orientable surface is a topological invariant. Similarly, a connected sum of n projective planes has abelianized fundamental group of rank $n-1$, so once again the genus is a topological invariant. □

Now we can tie up the loose ends regarding the combinatorial invariants we discussed at the end of Chapter 6. Recall that a compact 2-manifold is said to be orientable if it admits an oriented presentation.

Corollary 10.14. *A connected sum of projective planes is not orientable.*

Proof. By the argument in Chapter 6, if a manifold admits an oriented presentation, then it is homeomorphic to a sphere or a connected sum of tori. The preceding corollary showed that a connected sum of projective planes is not homeomorphic to any of these surfaces. □

Corollary 10.15. *The Euler characteristic of a surface presentation is a topological invariant.*

Proof. Suppose $\mathcal{P}$ and $\mathcal{Q}$ are polygonal surface presentations such that $|\mathcal{P}| \approx |\mathcal{Q}|$. Each of these presentations can be transformed into one of the standard ones by elementary transformations, and since the surfaces represented by different standard presentations are not homeomorphic, both presentations must reduce to the same standard one. Since the Euler characteristic of a presentation is unchanged by elementary transformations, the two presentations must have had the same Euler characteristic to begin with. □

Because of this corollary, we can define the *Euler characteristic* of a compact surface M, denoted by $\chi(M)$, to be the Euler characteristic of any presentation of that surface.

Problems

10-1. Compute the fundamental group of each of the following spaces. (To "compute" a fundamental group means to give a presentation of the group together with a specific loop representing each generator.)

(a) A closed disk with two points removed.

(b) The projective plane with two points removed.

(c) The connected sum of n tori with one point removed.

(d) The connected sum of n tori with two points removed.

10-2. Give a purely algebraic proof that the groups $\langle \alpha, \beta \mid \alpha\beta\alpha\beta^{-1} \rangle$ and $\langle \rho, \gamma \mid \rho^2\gamma^2 \rangle$ are isomorphic. [Hint: Look at the Klein bottle for inspiration.]

10-3. Let n be an integer greater than 2. Construct a polygonal presentation whose geometric realization has a fundamental group that is cyclic of order n.

10-4. Suppose M and N are compact 2-manifolds. Show that any two connected sums of M and N are homeomorphic.

10-5. Compute the fundamental group of the complement of the three coordinate axes in $\mathbb{R}^3$. [Hint: This space is homotopy equivalent to the 2-sphere with six points removed.]

10-6. If M is a connected manifold of dimension at least 3 and $q \in M$, show that $\pi_1(M \smallsetminus \{q\}) \cong \pi_1(M)$.

10-7. Let M and N be connected n-manifolds, $n \geq 3$. Prove that the fundamental group of $M \# N$ is isomorphic to $\pi_1(M) * \pi_1(N)$.

10-8. Let P be the polyhedron of a finite simplicial complex $\mathcal{K}$, and assume that P is connected. Show that the fundamental group of P has a presentation with the same number of generators as $\pi_1(\mathcal{K}^{(1)})$ and one relation for each 2-simplex in $\mathcal{K}$. [Hint: First treat the case of a 2-dimensional complex by induction on the number of 2-simplices. Then use a similar argument to show that inclusion $\mathcal{K}^{(j)} \hookrightarrow \mathcal{K}^{(j+1)}$ induces a fundamental group isomorphism for $j \geq 2$.]

10-9. Let Γ be a finite connected graph. The Euler characteristic of Γ is $\chi(\Gamma) = V - E$, where V is the number of vertices and E is the number of edges. Show that the fundamental group of Γ is a free group on $1 - \chi(\Gamma)$ generators. Conclude that $\chi(\Gamma)$ is a homotopy invariant, i.e., that homotopy equivalent graphs have the same Euler characteristic. [Hint: First show by induction on the number of edges that the Euler characteristic of a finite tree is 1.]

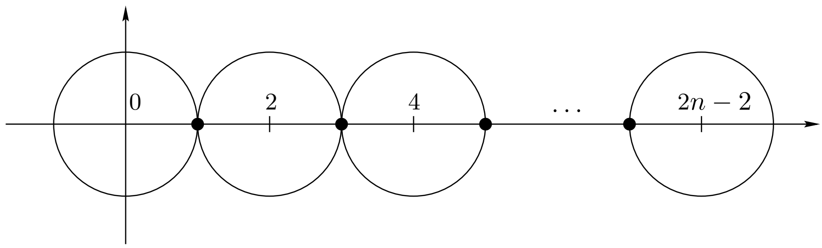


FIGURE 10.7. The space X_n of Problem 10-10.

10-10. Let X_n be the union of n circles of radius 1 centered at the points $\{0, 2, 4, \ldots, 2n-2\}$ in $\mathbb{C}$, which are pairwise tangent to each other along the x-axis (Figure 10.7). (Note that X_2 is homeomorphic to the figure eight space.) Prove that $\pi_1(X_n, 1)$ is a free group on n generators, and describe explicit loops representing the generators.

10-11. Show that a connected, compact surface M is nonorientable if and only if it contains a subset homeomorphic to the Möbius band.

10-12. Let $X \subset \mathbb{R}^3$ be the union of the unit 2-sphere with the line segment $\{(0,0,z) : -1 \le z \le 1\}$. Compute $\pi_1(X, N)$, where $N = (0,0,1)$ is the north pole, giving explicit generator(s).

10-13. Show that abelianization defines a functor from GROUP to AB. (You have to decide what the induced homomorphisms are.)

10-14. Given a group G, show that $\mathrm{Ab}(G)$ is the unique group that satisfies the characteristic property expressed in Theorem 10.11.

10-15. If G_1 and G_2 are groups, show that $\mathrm{Ab}(G_1 * G_2) \cong \mathrm{Ab}(G_1) \oplus \mathrm{Ab}(G_2)$. Conclude as a corollary that the abelianization of a free group on n generators is free abelian of rank n.

11
Covering Spaces

So far we have developed two general techniques for computing fundamental groups. The first is homotopy equivalence, which can often be used to show that one space has the same fundamental group as a simpler one. This was used in Chapter 7, for example, to show that every contractible space is simply connected, and that the fundamental group of the punctured plane is infinite cyclic. The second is the Seifert–Van Kampen theorem, which was used in Chapter 10 to compute the fundamental groups of spheres and compact surfaces.

The only other fundamental group we have computed is that of the circle, for which we used a technique that at first glance might seem to be rather ad hoc. The strategy for computing $\pi_1(S^1, 1)$ in Chapter 8 was to use the properties of the exponential quotient map ε to show that any loop based at 1 in the circle lifts to a path in $\mathbb{R}$ that ends at an integer; this integer is called the winding number of the loop, and different loops are path homotopic if and only if they have the same winding number. Another way to express this result is that lifting provides a one-to-one correspondence between the fiber of ε over 1 and the fundamental group of the circle.

The main ingredients in the proof were the three "lifting properties" of the circle: the path lifting property (Lemma 8.3), the homotopy lifting property (Lemma 8.4), and the unique lifting property (Lemma 8.2). These, in turn, followed from the basic fact that every point in the circle has an evenly covered neighborhood.

In this chapter we introduce a far-reaching generalization of these ideas, and show how the same techniques can be applied to a broad class of topological spaces. This brings us to the last major subject in the course:

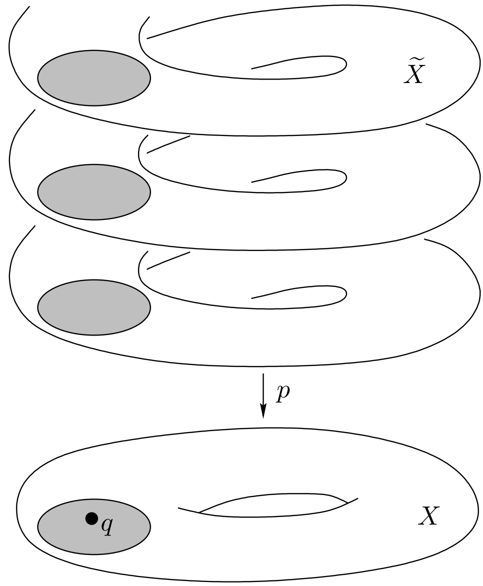


FIGURE 11.1. An evenly covered neighborhood of q.

covering spaces and covering maps. A covering map is a particular type of quotient map that has many of the same properties as the exponential quotient map. A careful study of covering maps will enable us to compute and analyze many more fundamental groups.

Definitions and Basic Properties

Let $\widetilde{X}$ and X be topological spaces, and let $p\colon \widetilde{X} \to X$ be a continuous map. A subset $U \subset X$ is said to be *evenly covered* by p if U is connected and open, and each component of $p^{-1}(U)$ is an open set that is mapped homeomorphically onto U by p (Figure 11.1). We usually visualize $p^{-1}(U)$ as a "stack of pancakes" that are projected down onto U by p. It is easy to see that any connected open subset of an evenly covered set is evenly covered.

A *covering map* is a continuous surjective map $p\colon \widetilde{X} \to X$ such that $\widetilde{X}$ is path connected and locally path connected, and every point $q \in X$ has an evenly covered neighborhood. If $p\colon \widetilde{X} \to X$ is a covering map, we call $\widetilde{X}$ a *covering space* of X, and X the *base* of the covering.

(Some authors define covering spaces more generally, omitting the requirement that $\widetilde{X}$ be locally path connected or path connected, or even sometimes omitting any connectivity requirement at all. In that case various connectivity hypotheses have to be added to the theorems below. We have chosen to include these hypotheses in the definition of covering maps, because most of the interesting results require them, such as the lifting criterion, the covering group structure theorem, and the classification of covering spaces, and this frees us from having to remember which connectivity hypotheses are necessary for which theorems. In any case, connected manifolds and most interesting spaces built from them will always satisfy the hypotheses.)

Example 11.1. The exponential quotient map $\varepsilon\colon \mathbb{R} \to \mathbb{S}^1$ given by $\varepsilon(x) = e^{2\pi i x}$ is a covering map; this is the content of Lemma 8.5.

Example 11.2. The nth power map $p_n\colon \mathbb{S}^1 \to \mathbb{S}^1$ given by $p_n(z) = z^n$ is also a covering map. For any $z_0 \in \mathbb{S}^1$, the set $U = \mathbb{S}^1 \smallsetminus \{-z_0\}$ has preimage equal to $\{z \in \mathbb{S}^1 : z^n \neq -z_0\}$, which has n components, each of which is an open arc mapped homeomorphically by p_n onto U.

Example 11.3. Define $E\colon \mathbb{R}^n \to \mathbb{T}^n$ by

$$E(x_1, \dots, x_n) = (\varepsilon(x_1), \dots, \varepsilon(x_n)),$$

where ε is the exponential quotient map of Example 11.1. Since a product of covering maps is a covering map (Problem 11-2), E is a covering map.

Example 11.4. Define a map $\pi\colon \mathbb{S}^n \to \mathbb{P}^n$ $(n \geq 1)$ by sending each point x in the sphere to the line through the origin and x, thought of as a point in $\mathbb{P}^n$. Then π is a covering map (Problem 11-1), and the fiber over each point of $\mathbb{P}^n$ is a pair of antipodal points $\{x, -x\}$.

Exercise 11.1. Let X_n be the union of n circles in $\mathbb{C}$ as described in Problem 10-10. Define a map $p\colon X_3 \to X_2$ by letting A, B, and C denote the circles centered at 0, 2, and 4, respectively (see Figure 11.2), and defining

$$p(z) = \begin{cases} z, & z \in A; \\ 2 - (z-2)^2, & z \in B; \\ 4 - z, & z \in C. \end{cases}$$

(In words, p is the identity on A, wraps B twice around itself, and reflects C onto A). Show that p is a covering map.

Lemma 11.5 (Elementary Properties of Covering Maps). *Every covering map is a local homeomorphism, an open map, and a quotient map. A one-to-one covering map is a homeomorphism.*

Exercise 11.2. Prove Lemma 11.5.

It is important to realize that a surjective local homeomorphism may not be a covering map, as the next example shows.

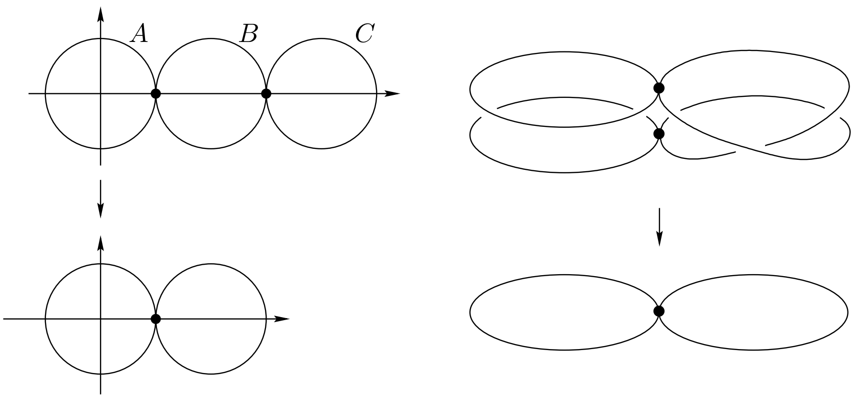


FIGURE 11.2. Two views of the map of Exercise 11.1.

Example 11.6. Let $\widetilde{X}$ be the interval $(0,2) \subset \mathbb{R}$, and define $f\colon \widetilde{X} \to \mathbb{S}^1$ by $f(x) = e^{2\pi i x}$ (Figure 11.3). Then f is a local homeomorphism (because it is the restriction of the covering map ε), and is clearly surjective. However, f is not a covering map, as is shown in the following exercise.

Exercise 11.3. Prove that the map f in the preceding example is not a covering map by showing that the point $1 \in \mathbb{S}^1$ has no evenly covered neighborhood.

Recall from Chapter 8 that a local section of a continuous map is a continuous right inverse defined on some open subset.

Lemma 11.7 (Existence of local sections). *Let $p\colon \widetilde{X} \to X$ be a covering map. Given any evenly covered open set $U \subset X$, any $q \in U$, and any $\widetilde{q}_0$ in the fiber over q, there exists a local section $\sigma\colon U \to \widetilde{X}$ such that $\sigma(q) = \widetilde{q}_0$.*

Proof. Let $\widetilde{U}_0$ be the component of $p^{-1}(U)$ containing $\widetilde{q}_0$. Since the restriction of p to $\widetilde{U}_0$ is a homeomorphism, we can just take $\sigma = (p|_{\widetilde{U}_0})^{-1}$. □

Proposition 11.8. *For any covering map $p\colon \widetilde{X} \to X$, the cardinality of the fibers $p^{-1}(q)$ is the same for all fibers.*

Proof. If U is any evenly covered open set in X, each component of $p^{-1}(U)$ contains exactly one point of each fiber. Thus, for any $q, q' \in U$, there are one-to-one correspondences

$$p^{-1}(q) \leftrightarrow \{\text{components of } p^{-1}(U)\} \leftrightarrow p^{-1}(q'),$$

which shows that the number of components is constant on U. It follows that the set of points $q' \in X$ such that $p^{-1}(q')$ has the same cardinality as $p^{-1}(q)$ is open.

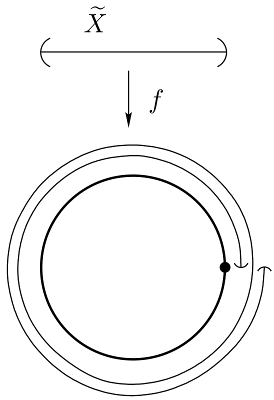


FIGURE 11.3. A surjective local homeomorphism that is not a covering map.

Now choose any point $q \in X$, and let A be the set of points in X whose fibers have cardinality equal to that of $p^{-1}(q)$. Then A is open by the above argument, and $X \smallsetminus A$ is open because it is a union of open sets (one for each cardinality not equal to that of $p^{-1}(q)$). Since X is connected and A is not empty, A must be equal to X. $\square$

If $p\colon \widetilde{X} \to X$ is a covering map, the cardinality of any fiber is called the *number of sheets* of the covering. For example, the nth power map of Example 11.2 is an n-sheeted covering; the map $\pi\colon \mathbb{S}^n \to \mathbb{P}^n$ of Example 11.4 is a two-sheeted covering; and the exponential quotient map $\varepsilon\colon \mathbb{R} \to \mathbb{S}^1$ has countably many sheets.

Lifting Properties

The key technical tools for working with covering spaces are the following three lifting properties, which are straightforward generalizations of the ones we proved for the circle in Chapter 8. In fact, the proofs for the circle apply almost verbatim to these more general propositions. We sketch the proofs in streamlined form here; if you remember the arguments given in Chapter 8, you can safely skip these proofs.

If $p\colon \widetilde{X} \to X$ is a covering map and $\varphi\colon B \to X$ is any continuous map, a *lift* of φ is a continuous map $\widetilde{\varphi}\colon B \to \widetilde{X}$ such that $p \circ \widetilde{\varphi} = \varphi$:

$$\begin{tikzcd} & \widetilde{X} \arrow[d, "p"] \\ B \arrow[ur, "\widetilde{\varphi}"] \arrow[r, "\varphi"'] & X. \end{tikzcd}$$

Proposition 11.9 (Unique Lifting Property). *Let $p\colon \widetilde{X} \to X$ be a covering map. Suppose B is connected, $\varphi\colon B \to X$ is continuous, and $\widetilde{\varphi}_1, \widetilde{\varphi}_2\colon B \to \widetilde{X}$ are lifts of φ that agree at some point of B. Then $\widetilde{\varphi}_1 \equiv \widetilde{\varphi}_2$.*

Proof. As in the proof of Lemma 8.2, it suffices to show that $\mathcal{S} = \{b \in B : \widetilde{\varphi}_1(b) = \widetilde{\varphi}_2(b)\}$ is open and closed in B.

For any $b \in \mathcal{S}$, let $U \subset X$ be an evenly covered neighborhood of $\varphi(b)$, and let U_α be the component of $p^{-1}(U)$ containing $\widetilde{\varphi}_1(b) = \widetilde{\varphi}_2(b)$. On the neighborhood $V = \widetilde{\varphi}_1^{-1}(U_\alpha) \cap \widetilde{\varphi}_2^{-1}(U_\alpha)$ of b, $\varphi = p \circ \widetilde{\varphi}_1 = p \circ \widetilde{\varphi}_2$. Since p is injective on U_α, this means $\widetilde{\varphi}_1 = \widetilde{\varphi}_2$ on V, so $\mathcal{S}$ is open.

On the other hand, for $b \notin \mathcal{S}$, if U is an evenly covered neighborhood of $\varphi(b)$, there are disjoint components U_1, U_2 of $p^{-1}(U)$ containing $\widetilde{\varphi}_1(b), \widetilde{\varphi}_2(b)$ such that p is a homeomorphism from each U_i to U. Letting $V = \widetilde{\varphi}_1^{-1}(U_1) \cap \widetilde{\varphi}_2^{-1}(U_2)$, we conclude that $\widetilde{\varphi}_1 \neq \widetilde{\varphi}_2$ on V, which shows that $\mathcal{S}$ is closed. □

Proposition 11.10 (Path Lifting Property). *Let $p\colon \widetilde{X} \to X$ be a covering map. Suppose $f\colon I \to X$ is any path, and $\widetilde{q}_0 \in \widetilde{X}$ is any point in the fiber of p over $f(0)$. Then there exists a unique lift $\widetilde{f}\colon I \to \widetilde{X}$ of f such that $\widetilde{f}(0) = \widetilde{q}_0$.*

Proof. By the Lebesgue number lemma, n can be chosen large enough that p maps each subinterval $[k/n, (k+1)/n]$ into an evenly covered open subset of X. Starting with $\widetilde{f}(0) = \widetilde{q}_0$, $\widetilde{f}$ is defined inductively by choosing an evenly covered neighborhood U_k containing $f[k/n, (k+1)/n]$, a local section $\sigma_k\colon U_k \to \widetilde{X}$ such that $\sigma_k(f(k/n)) = \widetilde{f}(k/n)$, and setting $\widetilde{f} = \sigma_k \circ f$ on $[k/n, (k+1)/n]$. Because $p \circ \widetilde{f} = (p \circ \sigma_k) \circ f = f$, this is indeed a lift, and it is unique by the unique lifting property. □

Proposition 11.11 (Homotopy Lifting Property). *Let $p\colon \widetilde{X} \to X$ be a covering map. Suppose $f_0, f_1\colon I \to X$ are path homotopic, and $\widetilde{f}_0, \widetilde{f}_1\colon I \to \widetilde{X}$ are lifts of f_0 and f_1 such that $\widetilde{f}_0(0) = \widetilde{f}_1(0)$. Then $\widetilde{f}_0 \sim \widetilde{f}_1$.*

Proof. If $H\colon f_0 \sim f_1$ is a path homotopy, by the Lebesgue number lemma we can choose n large enough that H maps each square of side $1/n$ into an evenly covered open set. Labeling the squares $S_{ij} = [i/n, (i+1)/n] \times [j/n, (j+1)/n]$, we define a lift $\widetilde{H}$ of H square by square along the bottom row, then along the next row, and so on by induction as in the proof of Lemma 8.3. On each square S_{ij}, set $\widetilde{H} = \sigma \circ H$, for an appropriate local section σ chosen so that the new definition of $\widetilde{H}$ matches the previous one at the corner point $(i/n, j/n)$. On a line segment where two such definitions overlap, both the old and new definitions are lifts of the path obtained by restricting H to that segment, starting at the same point. Thus they are equal by the unique lifting property.

On the left-hand and right-hand edges of $I \times I$, where $s = 0$ or $s = 1$, $\widetilde{H}$ is a lift of the constant loop and therefore constant. The restriction $\widetilde{H}_0$ to the bottom edge where $t = 0$ is a lift of f_0 starting at $\widetilde{f}_0(0)$, and therefore is equal to $\widetilde{f}_0$; and similarly $\widetilde{H}_1 = \widetilde{f}_1$. Thus $\widetilde{H}$ is the required path homotopy between $\widetilde{f}_0$ and $\widetilde{f}_1$. □

The next result is an immediate corollary of the homotopy lifting property.

Corollary 11.12 (The Monodromy Theorem). *Let $p\colon \widetilde{X} \to X$ be a covering map. Suppose f_0 and f_1 are path homotopic paths in X, and $\widetilde{f}_0$, $\widetilde{f}_1$ are lifts of them starting at the same point. Then $\widetilde{f}_0(1) = \widetilde{f}_1(1)$.*

Covering Maps and the Fundamental Group

Our next theorem characterizes the fundamental group homomorphism induced by a covering map.

Theorem 11.13 (Injectivity Theorem). *Let $p\colon \widetilde{X} \to X$ be a covering map. For any point $\widetilde{q} \in \widetilde{X}$, the induced homomorphism $p_*\colon \pi_1(\widetilde{X}, \widetilde{q}) \to \pi_1(X, p(\widetilde{q}))$ is injective.*

Proof. Suppose $[f] \in \pi_1(\widetilde{X}, \widetilde{q})$ is in the kernel of p_*. This means that $p_*[f] = [c_q]$, where $q = p(\widetilde{q})$, or in other words, $p \circ f \sim c_q$ in X. By the homotopy lifting property, therefore, any lifts of $p \circ f$ and c_q that start at the same point must be path homotopic in $\widetilde{X}$. Now, f is a lift of $p \circ f$ starting at $\widetilde{q}$, and the constant loop $c_{\widetilde{q}}$ is a lift of c_q starting at the same point; therefore, $f \sim c_{\widetilde{q}}$ in $\widetilde{X}$, which means that $[f] = 1$. □

This theorem shows that the fundamental group of a covering space can be identified with a subgroup of the fundamental group of the base. We call this the subgroup *induced* by the covering.

Example 11.14. Let $p\colon X_3 \to X_2$ be the covering map of Exercise 11.1, and choose 1 as base point in both X_3 and X_2. To compute the subgroup induced by p, we need to compute the action of p on the generators of $\pi_1(X_3, 1)$. Let a, b, c be loops that go once counterclockwise around each circle A, B, and C, starting at 1, 1, and 3, respectively; and let b_1 and b_2 be the lower and upper halves of b, so b_1 is a path from 1 to 3, b_2 is a path from 3 to 1, and $b \sim b_1 \cdot b_2$. Using the result of Problem 10-10, $\pi_1(X_3, 1)$ is the free group on $[a]$, $[b]$, and $[b_1 \cdot c \cdot b_1^{-1}]$, and $\pi_1(X_2, 1)$ is the free group on $[a]$ and $[b]$. The images of these generators under p_* are $[a]$, $[b]^2$, and $[b] \cdot [a]^{-1} \cdot [b]^{-1}$, so the subgroup induced by p is the subgroup of $\langle [a], [b] \rangle$ generated by these three elements.

As our first significant application of the theory we give a general solution to the *lifting problem* for covering maps: This is the problem of deciding, given a continuous map $\varphi\colon Y \to X$, whether φ admits a lift $\widetilde{\varphi}$ to a covering space $\widetilde{X}$ of X. The following theorem reduces this topological problem to an algebraic problem.

Theorem 11.15 (Lifting Criterion). *Suppose $p\colon \widetilde{X} \to X$ is a covering map. Let Y be a connected and locally path connected space, and let $\varphi\colon Y \to X$ be a continuous map. Given any points $y_0 \in Y$ and $\widetilde{q}_0 \in \widetilde{X}$ such that $p(\widetilde{q}_0) = \varphi(y_0)$, φ has a lift $\widetilde{\varphi}\colon Y \to \widetilde{X}$ satisfying $\widetilde{\varphi}(y_0) = \widetilde{q}_0$ if and only if the subgroup $\varphi_*\pi_1(Y, y_0)$ of $\pi_1(X, \varphi(y_0))$ is contained in $p_*\pi_1(\widetilde{X}, \widetilde{q}_0)$.*

Proof. The necessity of the algebraic condition is easy to prove (and, in fact, does not require any connectivity assumptions about Y). If $\widetilde{\varphi}$ satisfies the conditions in the statement of the theorem, the following diagram commutes:

$$\begin{tikzcd}
& \pi_1(\widetilde{X}, \widetilde{q}_0) \arrow[d, "p_*"] \\
\pi_1(Y, y_0) \arrow[ur, "\widetilde{\varphi}_*"] \arrow[r, "\varphi_*"'] & \pi_1(X, \varphi(y_0)).
\end{tikzcd}$$

Therefore, $\varphi_*\pi_1(Y, y_0) = p_*\widetilde{\varphi}_*\pi_1(Y, y_0) \subset p_*\pi_1(\widetilde{X}, \widetilde{q}_0)$.

To prove the converse, we will "lift φ along paths" using the path lifting property. If $\widetilde{\varphi}$ does exist, it will have the following property: For any point $y \in Y$ and any path f from y_0 to y, $\widetilde{\varphi} \circ f$ is a lift of $\varphi \circ f$ starting at $\widetilde{q}_0$, and $\widetilde{\varphi}(y)$ is equal to the terminal point of this path. We use this observation to *define* $\widetilde{\varphi}$: Namely, for any $y \in Y$, choose a path f from y_0 to y, let $\widetilde{\varphi \circ f}$ be the (unique) lift of the path $\varphi \circ f$ to a path in $\widetilde{X}$ starting at $\widetilde{q}_0$, and set

$$\widetilde{\varphi}(y) = \widetilde{\varphi \circ f}(1).$$

We need to show two things: (1) $\widetilde{\varphi}$ is well-defined, independently of the choice of the path f; and (2) $\widetilde{\varphi}$ is continuous. Then it is immediate from the definition that $p \circ \widetilde{\varphi}(y) = p \circ \widetilde{\varphi \circ f}(1) = \varphi \circ f(1) = \varphi(y)$, so $\widetilde{\varphi}$ is a lift of φ.

Claim 1: *$\widetilde{\varphi}$ is well-defined.* Suppose f and f' are two paths from y_0 to y (Figure 11.4). Then $f' \cdot f^{-1}$ is a loop based at y_0, so

$$\varphi_*[f' \cdot f^{-1}] \in \varphi_*\pi_1(Y, y_0) \subset p_*\pi_1(\widetilde{X}, \widetilde{q}_0).$$

This means that $[\varphi \circ (f' \cdot f^{-1})] = [p \circ g]$ for some loop g in $\widetilde{X}$ based at $\widetilde{q}_0$. Thus we have the following path homotopy in X:

$$p \circ g \sim \varphi \circ (f' \cdot f^{-1}) = (\varphi \circ f') \cdot (\varphi \circ f)^{-1},$$

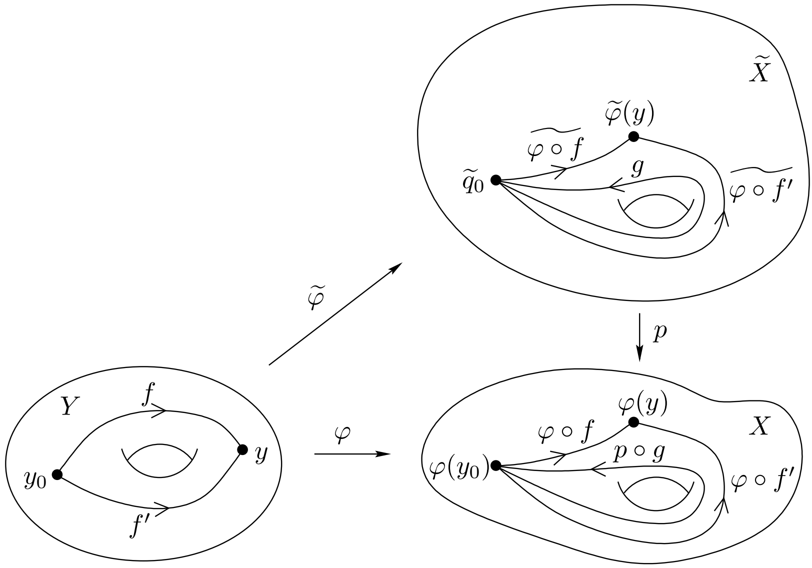


FIGURE 11.4. Proof that $\widetilde{\varphi}$ is well-defined.

which implies

$$(p \circ g) \cdot (\varphi \circ f) \sim (\varphi \circ f').$$

By the monodromy theorem, the lifts of these two paths starting at $\widetilde{q}_0$ have the same terminal points. Since the lift of $p \circ g$ is g, which starts and ends at $\widetilde{q}_0$, this implies

$$\widetilde{\varphi \circ f'}(1) = g \cdot \widetilde{\varphi \circ f}(1) = \widetilde{\varphi \circ f}(1),$$

so $\widetilde{\varphi}$ is well-defined.

CLAIM 2: *$\widetilde{\varphi}$ is continuous.* Before proving this, we will show that $\widetilde{\varphi}$ has one important property of a continuous map: It takes path connected sets to path connected sets. Let $V \subset Y$ be path connected, and $y_1, y_2 \in V$ be arbitrary. There is a path f in Y from y_0 to y_1, and a path g in V from y_1 to y_2 (Figure 11.5); by definition, $\widetilde{\varphi}$ maps the path $f \cdot g$ to the lift of $(\varphi \circ f) \cdot (\varphi \circ g)$. In particular, the lift of $\varphi \circ g$ is a path from $\widetilde{\varphi}(y_1)$ to $\widetilde{\varphi}(y_2)$ that is contained in $\widetilde{\varphi}(V)$. This proves that $\widetilde{\varphi}(V)$ is path connected.

To prove that $\widetilde{\varphi}$ is continuous, it suffices to show that each point in Y has a neighborhood on which $\widetilde{\varphi}$ is continuous. Let $y \in Y$ be arbitrary, let U be an evenly covered neighborhood of $\varphi(y)$, and let $\widetilde{U}$ be the component of $p^{-1}(U)$ containing $\widetilde{\varphi}(y)$ (Figure 11.6). If V is the path component of $\varphi^{-1}(U)$ containing y, the argument above shows that $\widetilde{\varphi}(V)$ is a connected

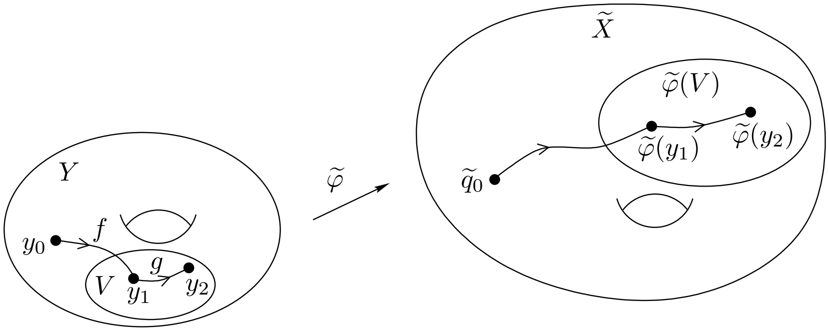


FIGURE 11.5. Proof that $\widetilde{\varphi}$ takes path connected sets to path connected sets.

subset of $p^{-1}(U)$, and must therefore be contained in $\widetilde{U}$. Since Y is locally path connected, V is open and thus is a neighborhood of y. Let $\sigma\colon U \to \widetilde{U}$ be the local section of p taking $\varphi(y)$ to $\widetilde{\varphi}(y)$, so $p \circ \sigma$ is the identity on U. The following equation holds on V:

$$p \circ \widetilde{\varphi} = \varphi = p \circ \sigma \circ \varphi.$$

Both $\widetilde{\varphi}$ and $\sigma \circ \varphi$ map V into $\widetilde{U}$, where p is injective, so this equation implies $\widetilde{\varphi} = \sigma \circ \varphi$ on V, which is a composition of continuous maps. □

The following corollaries are immediate.

Corollary 11.16. *If $p\colon \widetilde{X} \to X$ is a covering map and Y is a simply connected and locally path connected space, every continuous map $\varphi\colon Y \to X$ has a lift to $\widetilde{X}$. Given any point $y_0 \in Y$, the lift can be chosen to take y_0 to any point in the fiber over $\varphi(y_0)$.*

Corollary 11.17. *Suppose $p\colon \widetilde{X} \to X$ is a covering map and $\widetilde{X}$ is simply connected. For any connected and locally path connected space Y, a continuous map $\varphi\colon Y \to X$ has a lift to $\widetilde{X}$ if and only if φ_* is the trivial homomorphism for any base point $y_0 \in Y$. If this is the case, then the lift can be chosen to take y_0 to any point in the fiber over $\varphi(y_0)$.*

Example 11.18. Consider the n-sheeted covering of the circle given by the nth power map $p_n\colon \mathbb{S}^1 \to \mathbb{S}^1$. It is easy to check that the subgroup of $\pi_1(\mathbb{S}^1, 1)$ induced by p_n is the cyclic subgroup generated by $[\alpha]^n$. Thus, for

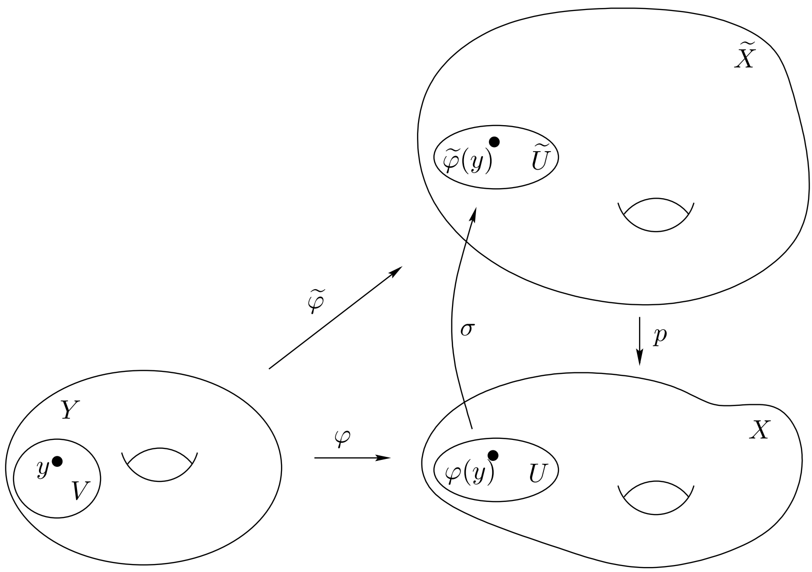


FIGURE 11.6. Proof that $\widetilde{\varphi}$ is continuous.

any integer m, there is a continuous map f making the diagram

$$\begin{tikzcd}
 & \mathbb{S}^1 \arrow[d, "p_n"] \\
\mathbb{S}^1 \arrow[ur, "f"] \arrow[r, "p_m"'] & \mathbb{S}^1
\end{tikzcd}$$

commute if and only if $m = nk$ for some integer k. If this is the case, the lift sending 1 to 1 is obviously $f = p_k$.

Dependence on Base Points

It is important to remember that in general, the subgroup induced by a covering depends not only on the covering but also on the choice of base point $\widetilde{q} \in \widetilde{X}$. As the next theorem shows, the subgroup may change when we change base point, but it can change only in a very limited way.

Theorem 11.19 (Conjugacy Theorem). *Let $p\colon \widetilde{X} \to X$ be a covering map. For any $q \in X$, as $\widetilde{q}$ varies over the fiber $p^{-1}(q)$, the set of subgroups $p_*\pi_1(\widetilde{X}, \widetilde{q}) \subset \pi_1(X, q)$ is exactly one conjugacy class.*

Proof. First we will show that given any $\widetilde{q}, \widetilde{q}\,' \in p^{-1}(q)$, the subgroups $p_*\pi_1(\widetilde{X}, \widetilde{q})$ and $p_*\pi_1(\widetilde{X}, \widetilde{q}\,')$ are conjugate. Let $\widetilde{g}$ be a path in $\widetilde{X}$ from $\widetilde{q}$ to

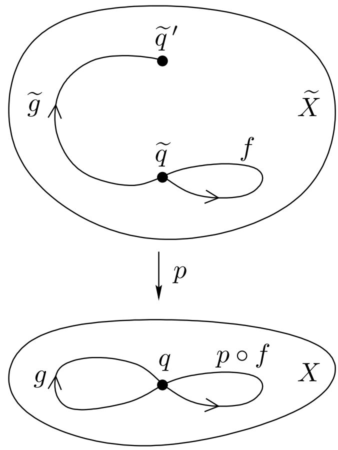


FIGURE 11.7. Proof of the conjugacy theorem.

$\widetilde{q}\,'$, and let $g = p \circ \widetilde{g}$, which is a loop in X based at q (Figure 11.7). We have four maps

$$\begin{array}{ccc} \pi_1(\widetilde{X}, \widetilde{q}) & \xrightarrow{\Phi_{\widetilde{g}}} & \pi_1(\widetilde{X}, \widetilde{q}\,') \\ \Big\downarrow p_* & & \Big\downarrow p_* \\ \pi_1(X, q) & \xrightarrow[\Phi_g]{} & \pi_1(X, q), \end{array} \tag{11.1}$$

where $\Phi_{\widetilde{g}}[f] = [\widetilde{g}^{-1}] \cdot [f] \cdot [\widetilde{g}]$, and Φ_g is defined similarly. This diagram commutes, because

$$\begin{aligned} p_*\Phi_{\widetilde{g}}[f] &= p_*[\widetilde{g}^{-1} \cdot f \cdot \widetilde{g}] \\ &= [p \circ (\widetilde{g}^{-1} \cdot f \cdot \widetilde{g})] \\ &= [(p \circ \widetilde{g}^{-1}) \cdot (p \circ f) \cdot (p \circ \widetilde{g})] \\ &= [g^{-1}] \cdot p_*[f] \cdot [g] \\ &= \Phi_g p_*[f]. \end{aligned}$$

This means that Φ_g maps the subgroup $p_*\pi_1(\widetilde{X}, \widetilde{q})$ into $p_*\pi_1(\widetilde{X}, \widetilde{q}\,')$, so we can replace the bottom two groups in (11.1) by the image groups under

p_* to obtain

$$\begin{tikzcd} \pi_1(\widetilde{X},\widetilde{q}) \arrow[r, "\Phi_{\widetilde{g}}"] \arrow[d, "p_*"'] & \pi_1(\widetilde{X},\widetilde{q}') \arrow[d, "p_*"] \\ p_*\pi_1(\widetilde{X},\widetilde{q}) \arrow[r, "\Phi_g"'] & p_*\pi_1(\widetilde{X},\widetilde{q}'). \end{tikzcd} \tag{11.2}$$

Now, in this diagram, both vertical maps are isomorphisms, and the top map is an isomorphism by Theorem 7.11. This means that the bottom map Φ_g is an isomorphism as well. But Φ_g is exactly the map that sends any subgroup onto its conjugate by $[g]^{-1}$, so this shows that $p_*\pi_1(\widetilde{X},\widetilde{q})$ and $p_*\pi_1(\widetilde{X},\widetilde{q}')$ are conjugate subgroups.

Conversely, let $\widetilde{q} \in p^{-1}(q)$, and suppose H is any subgroup of $\pi_1(X,q)$ conjugate to $p_*\pi_1(\widetilde{X},\widetilde{q})$. This means that there is some element $[g] \in \pi_1(X,q)$ such that $H = \Phi_g(p_*\pi_1(\widetilde{X},\widetilde{q}))$. If we let $\widetilde{g}$ be the lift of g starting at $\widetilde{q}$, and $\widetilde{q}' = \widetilde{g}(1)$, the above construction shows that $p_*\pi_1(\widetilde{X},\widetilde{q}') = \Phi_g(p_*\pi_1(\widetilde{X},\widetilde{q})) = H$. □

There is an important special case in which the subgroup $p_*\pi_1(\widetilde{X},\widetilde{q})$ does not depend on the choice of base point. We say that the covering $p\colon \widetilde{X} \to X$ is *normal* if $p_*\pi_1(\widetilde{X},\widetilde{q})$ is a normal subgroup of $\pi_1(X,p(\widetilde{q}))$ for each $\widetilde{q} \in \widetilde{X}$. This means, in particular, that for any fixed $q \in X$ the subgroup $p_*\pi_1(\widetilde{X},\widetilde{q})$ is independent of the choice of base point $\widetilde{q}$ in the fiber over q, because the only subgroup conjugate to a normal subgroup is itself. In fact, as the next lemma shows, as long as the induced subgroup is normal for one choice of $\widetilde{q} \in \widetilde{X}$, it is normal for all of them.

Lemma 11.20. *Let $p\colon \widetilde{X} \to X$ be a covering map, and suppose the subgroup induced by p is normal for one point $\widetilde{q} \in \widetilde{X}$. Then p is normal.*

Proof. Let $\widetilde{q},\widetilde{q}'$ be two points of $\widetilde{X}$, and let $q = p(\widetilde{q})$, $q' = p(\widetilde{q}')$. Let $\widetilde{g}$ be a path from $\widetilde{q}$ to $\widetilde{q}'$, and set $g = p \circ \widetilde{g}$, which is a path from q to q'. If we replace q by q' in the bottom right corner of diagram (11.1), the diagram still commutes, and the top and bottom rows are still isomorphisms. It follows from the commutativity of the diagram that Φ_g takes $p_*(\pi_1(\widetilde{X},\widetilde{q}))$ to $p_*(\pi_1(\widetilde{X},\widetilde{q}'))$. Since an isomorphism takes normal subgroups to normal subgroups, the result follows. □

Next we show that there is a natural right action of the fundamental group of the base on the fiber of any covering space. Recall from Chapter 3 that given an action of a group Γ on a set F, the orbit of a point $x \in F$ is the set of all images of x under elements of the group (for a right action, this is the set $\{x \cdot \gamma : \gamma \in \Gamma\}$), and the action is said to be transitive if each orbit is all of F.

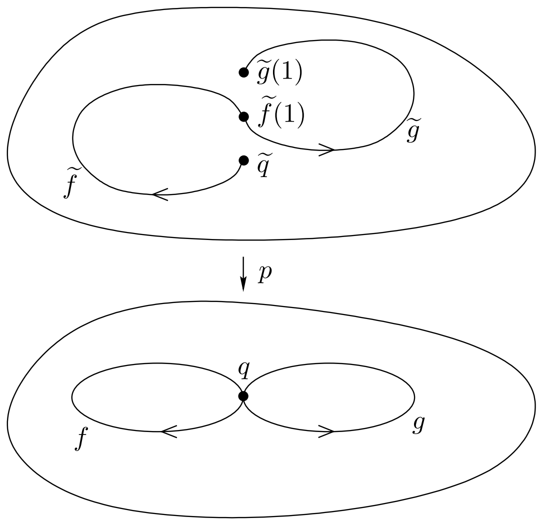


FIGURE 11.8. Proof that the fundamental group acts on the fiber.

Theorem 11.21 (Action of the Fundamental Group on a Fiber). *Suppose $p\colon \widetilde{X} \to X$ is a covering map and $q \in X$. There is a transitive right action of $\pi_1(X,q)$ on the fiber $p^{-1}(q)$, given by $\widetilde{q} \cdot [f] = \widetilde{f}(1)$, where $\widetilde{f}$ is the lift of f starting at $\widetilde{q} \in p^{-1}(q)$.*

Proof. If $\widetilde{q}$ is any point in the fiber over q, any path f starting at q has a lift to a path $\widetilde{f}$ starting at $\widetilde{q}$ by the path lifting property. The monodromy theorem guarantees that the endpoint $\widetilde{f}(1)$ depends only on the path class of f; therefore, $\widetilde{q} \cdot [f]$ is well-defined.

To see that this is a group action, we need to check two things:

(a) $\widetilde{q} \cdot [c_q] = \widetilde{q}$;

(b) $(\widetilde{q} \cdot [f]) \cdot [g] = \widetilde{q} \cdot ([f] \cdot [g])$.

For (a), just observe that the constant loop $c_{\widetilde{q}}$ is the unique lift of c_q starting at $\widetilde{q}$, and therefore $\widetilde{q} \cdot [c_q] = c_{\widetilde{q}}(1) = \widetilde{q}$. To prove the composition property (b), suppose f and g are two loops based at q. Let $\widetilde{f}$ be the lift of f starting at $\widetilde{q}$, so that $\widetilde{q} \cdot [f] = \widetilde{f}(1)$. Now, if $\widetilde{g}$ is the lift of g starting at $\widetilde{f}(1)$, then by definition, $(\widetilde{q} \cdot [f]) \cdot [g] = \widetilde{g}(1)$ (Figure 11.8). On the other hand, $\widetilde{f} \cdot \widetilde{g}$ is clearly the lift of $f \cdot g$ starting at $\widetilde{q}$. This means that

$$\begin{aligned}\widetilde{q} \cdot ([f] \cdot [g]) &= \widetilde{q} \cdot [f \cdot g] \\ &= (\widetilde{f} \cdot \widetilde{g})(1) \\ &= \widetilde{g}(1).\end{aligned}$$

Finally, to prove that the action is transitive, just note that any two points $\widetilde{q}, \widetilde{q}'$ in the fiber over q are joined by a path $\widetilde{f}$ because $\widetilde{X}$ is path connected. Setting $f = p \circ \widetilde{f}$, it is immediate that $\widetilde{f}$ is the lift of f starting at $\widetilde{q}$, and therefore $\widetilde{q} \cdot [f] = \widetilde{q}'$. □

Corollary 11.22. *Let $p\colon \widetilde{X} \to X$ be a covering map, and suppose $\widetilde{X}$ is simply connected. The number of sheets of the covering is equal to the cardinality of the fundamental group of X.*

Proof. Choose a base point $q \in X$ and a point $\widetilde{q}$ in the fiber over q, and consider the map $\pi_1(X, q) \to p^{-1}(q)$ given by $[f] \mapsto \widetilde{q} \cdot [f]$. It is surjective because the action of the fundamental group is transitive. To show that it is injective, suppose that $\widetilde{q} \cdot [f] = \widetilde{q} \cdot [g]$. This means that the lifts $\widetilde{f}$ and $\widetilde{g}$ starting at $\widetilde{q}$ end at the same point. Since $\widetilde{X}$ is simply connected, $\widetilde{f} \sim \widetilde{g}$, and therefore $[f] = p_*[\widetilde{f}] = p_*[\widetilde{g}] = [g]$. □

Example 11.23. Since $\pi\colon \mathbb{S}^n \to \mathbb{P}^n$ is a two-sheeted covering and $\mathbb{S}^n$ is simply connected, Corollary 11.22 shows that $\pi_1(\mathbb{P}^n)$ is a two-element group, which must therefore be isomorphic to $\mathbb{Z}/\langle 2 \rangle$.

Corollary 11.24. *If X is a simply connected space, any covering map $p\colon \widetilde{X} \to X$ is a homeomorphism.*

Proof. The injectivity theorem shows that $\widetilde{X}$ is also simply connected. Thus Corollary 11.22 shows that the cardinality of the fibers is 1, so p is a one-to-one covering map and therefore a homeomorphism. □

The Covering Group

In this section we introduce the group of covering transformations of a covering space, and explore its relation to the fundamental groups of the base and the covering space.

Suppose $p\colon \widetilde{X} \to X$ is a covering map. A homeomorphism $\varphi\colon \widetilde{X} \to \widetilde{X}$ is called a *covering transformation* if $p \circ \varphi = p$:

$$\begin{tikzcd} \widetilde{X} \arrow[rr, "\varphi"] \arrow[dr, "p"'] & & \widetilde{X} \arrow[dl, "p"] \\ & X. & \end{tikzcd}$$

Covering transformations are also variously known as *deck transformations* or *automorphisms* of the covering.

Let $\mathcal{C}_p(\widetilde{X})$ denote the set of all covering transformations of $\widetilde{X}$ with respect to p. It is easy to verify that the composition of two covering transformations, the inverse of a covering transformation, and the identity map of $\widetilde{X}$ are all covering transformations; thus $\mathcal{C}_p(\widetilde{X})$ is a group, called the *covering group* or the *automorphism group* of the covering. It acts on $\widetilde{X}$ (on the left) in a natural way, and the definition of covering transformations implies that each orbit is a subset of a single fiber.

Example 11.25. For the covering $\varepsilon\colon \mathbb{R} \to \mathbb{S}^1$, the integral translations $x \mapsto x + k$ for $k \in \mathbb{Z}$ are easily seen to be covering transformations. More generally, for any integers $(k_1, \dots, k_n)$, the translation $(x_1, \dots, x_n) \mapsto (x_1 + k_1, \dots, x_n + k_n)$ is a covering transformation of $E\colon \mathbb{R}^n \to \mathbb{T}^n$. We will prove below that these are all of them.

Example 11.26. If $\pi\colon \mathbb{S}^n \to \mathbb{P}^n$ is the covering map of Example 11.4, then the antipodal map $A(x) = -x$ is a covering transformation. We will see shortly that $\mathcal{C}_\pi(\mathbb{P}^n)$ is the two-element group $\{\mathrm{Id}, A\}$.

Proposition 11.27 (Properties of the Covering Group). *Let $p\colon \widetilde{X} \to X$ be a covering map.*

(*a*) *If two covering transformations agree at one point, they are identical.*

(*b*) *The covering group acts freely and continuously on $\widetilde{X}$: If $\varphi(\widetilde{q}) = \widetilde{q}$ for some $\widetilde{q} \in \widetilde{X}$, then $\varphi = \mathrm{Id}_{\widetilde{X}}$.*

(*c*) *For any $q \in X$, each covering transformation permutes the points of the fiber $p^{-1}(q)$.*

(*d*) *For any evenly covered open set $U \subset X$, each covering transformation permutes the components of $p^{-1}(U)$.*

Proof. Note that a covering transformation φ is, in particular, a lift of p:

$$\begin{tikzcd} & \widetilde{X} \arrow[d, "p"] \\ \widetilde{X} \arrow[ur, "\varphi"] \arrow[r, "p"'] & X. \end{tikzcd}$$

Thus (a) follows from the unique lifting property. The covering group acts continuously because each covering transformation is continuous by definition; the fact that it acts freely follows from (a) by comparing φ with the identity. Part (c) follows from the fact that if $\widetilde{q} \in p^{-1}(q)$, then $p(\varphi(\widetilde{q})) = p(\widetilde{q}) = p$, so φ takes the fiber over q to itself; since the same is true of φ^{-1}, φ acts as a permutation of the fiber. To prove (d), let U be an evenly covered open set, and let U_α be a component of $p^{-1}(U)$. Since $\varphi(U_\alpha)$ is a connected subset of $p^{-1}(U)$, it must be contained in a single component; applying the same argument to φ^{-1} shows that $\varphi(U_\alpha)$ is exactly a component. □

Example 11.28. Consider again the covering group of $\varepsilon\colon \mathbb{R} \to \mathbb{S}^1$. Let $\varphi \in \mathcal{C}_\varepsilon(\mathbb{R})$ be arbitrary. If we set $n = \varphi(0)$, then both φ and the translation $x \mapsto x + n$ are covering transformations taking 0 to n, and are therefore equal by Proposition 11.27(a). Thus the covering group of $\varepsilon\colon \mathbb{R} \to \mathbb{S}^1$ is equal to $\mathbb{Z}$ acting on $\mathbb{R}$ by integral translations. By a similar argument, the covering group of $E\colon \mathbb{R}^n \to \mathbb{T}^n$ is $\mathbb{Z}^n$ acting by translations, and the covering group of $\pi\colon \mathbb{S}^n \to \mathbb{P}^n$ is equal to the cyclic group of order 2 generated by the antipodal map.

Because of Proposition 11.27, the action of the covering group on $\widetilde{X}$ is completely determined by the action on any fiber. However, unlike the action of the fundamental group on a fiber that we defined in Theorem 11.21, the action of the covering group on fibers is not transitive in general. It is often useful to have a criterion for deciding when two points in a fiber are in the same orbit.

Proposition 11.29 (Orbit Criterion). *Let $p\colon \widetilde{X} \to X$ be a covering map.*

(a) If $\widetilde{q}$, $\widetilde{q}' \in \widetilde{X}$ are two points in the same fiber $p^{-1}(q)$, there exists a covering transformation taking $\widetilde{q}$ to $\widetilde{q}'$ if and only if the induced subgroups $p_*\pi_1(\widetilde{X},\widetilde{q})$ and $p_*\pi_1(\widetilde{X},\widetilde{q}')$ are equal.

(b) $\mathcal{C}_p(\widetilde{X})$ acts transitively on each fiber if and only if p is a normal covering.

Proof. If there exists φ such that $\varphi(\widetilde{q}) = \widetilde{q}'$, then $\varphi_*\colon \pi_1(\widetilde{X},\widetilde{q}) \to \pi_1(\widetilde{X},\widetilde{q}')$ is an isomorphism, so $p_*\pi_1(\widetilde{X},\widetilde{q}) = p_*\varphi_*\pi_1(\widetilde{X},\widetilde{q}) = p_*\pi_1(\widetilde{X},\widetilde{q}')$. Conversely, if the two subgroups are equal, then the lifting criterion yields a lift $\widetilde{p}\colon \widetilde{X} \to \widetilde{X}$ satisfying $p \circ \widetilde{p} = p$ and $\widetilde{p}(\widetilde{q}) = \widetilde{q}'$. Reversing the roles of $\widetilde{q}$ and $\widetilde{q}'$, we get a lift $\widetilde{p}'$ satisfying $\widetilde{p}'(\widetilde{q}') = \widetilde{q}$. To show that $\widetilde{p}$ and $\widetilde{p}'$ are inverses of each other, note that $\widetilde{p}' \circ \widetilde{p}$ and $\mathrm{Id}_{\widetilde{X}}$ are lifts of p taking $\widetilde{q}$ to itself, and thus are equal by the unique lifting property, and similarly $\widetilde{p} \circ \widetilde{p}' = \mathrm{Id}_{\widetilde{X}}$. Therefore, $\widetilde{p}$ is the required covering transformation.

Now suppose p is normal. This means that for any $\widetilde{q}, \widetilde{q}'$ in the same fiber, $p_*\pi_1(\widetilde{X},\widetilde{q}) = p_*\pi_1(\widetilde{X},\widetilde{q}')$, so by part (a) there is a covering transformation taking $\widetilde{q}$ to $\widetilde{q}'$. Conversely, if $\mathcal{C}_p(\widetilde{X})$ acts transitively on the fiber $p^{-1}(q)$, the groups $p_*\pi_1(\widetilde{X},\widetilde{q})$ coincide for all $\widetilde{q} \in p^{-1}(q)$, which is to say that p is normal. □

The next theorem is the central result concerning the relationship between covering spaces and fundamental groups. It gives an explicit formula for the covering group in terms of the fundamental groups of the covering space and the base, and can be used to compute the fundamental groups of certain spaces from properties of their coverings. For normal coverings, the theorem simply says that the covering group is isomorphic to the quotient of the fundamental group of the base by the subgroup induced by the covering. The statement for general groups is somewhat more complicated, and involves the following algebraic notion: If G is a group and $H \subset G$ is a subgroup, the *normalizer* of H in G, denoted by $N(H)$, is the set of all elements $\gamma \in G$ such that $\gamma^{-1}H\gamma = H$. The normalizer $N(H)$ is easily seen to be a subgroup of G; it is in fact the largest subgroup in which H is normal.

Theorem 11.30 (Covering Group Structure Theorem). *Suppose $p\colon \widetilde{X} \to X$ is a covering map and $\widetilde{q} \in \widetilde{X}$. The covering group $\mathcal{C}_p(\widetilde{X})$ is isomorphic to the quotient*

$$\frac{N(p_*\pi_1(\widetilde{X},\widetilde{q}))}{p_*\pi_1(\widetilde{X},\widetilde{q})}.$$

The isomorphism is induced by the map $\alpha\colon N(p_*\pi_1(\widetilde{X},\widetilde{q})) \to \mathcal{C}_p(\widetilde{X})$ that sends $[f]$ to the unique covering transformation φ taking $\widetilde{q}$ to $\widetilde{q} \cdot [f]$.

Before we give the proof of this important theorem, which is rather technical, let us derive some immediate consequences to illustrate its utility.

Corollary 11.31 (Normal Case). *If $p\colon \widetilde{X} \to X$ is a normal covering, $\widetilde{q} \in \widetilde{X}$, and $q = p(\widetilde{q})$, then $\mathcal{C}_p(\widetilde{X}) \cong \pi_1(X,q)/p_*\pi_1(\widetilde{X},\widetilde{q})$.*

Corollary 11.32 (Simply Connected Case). *If $p\colon \widetilde{X} \to X$ is a covering map and $\widetilde{X}$ is simply connected, then for any $\widetilde{q} \in \widetilde{X}$ the map α of Theorem 11.30 is an isomorphism from $\pi_1(X,q)$ to $\mathcal{C}_p(\widetilde{X})$, where $q = p(\widetilde{q})$.*

Example 11.33. Since the covering group of $\varepsilon\colon \mathbb{R} \to \mathbb{S}^1$ is infinite cyclic and $\mathbb{R}$ is simply connected, Corollary 11.32 yields another proof that the fundamental group of the circle is infinite cyclic. In fact, if you look back carefully at the proof in Chapter 8, you will see that this is really the same proof we gave there, decorated with some fancier terminology.

Example 11.34. Because the covering group of $\pi\colon \mathbb{S}^n \to \mathbb{P}^n$ is the two-element group $\{\mathrm{Id}, A\}$, Corollary 11.32 gives another proof that $\pi_1(\mathbb{P}^n) \cong \mathbb{Z}/\langle 2\rangle$.

Proof of the covering group structure theorem. Write $H = p_*\pi_1(\widetilde{X}, \widetilde{q}) \subset \pi_1(X, q)$. We will show that $\alpha\colon N(H) \to \mathcal{C}_p(\widetilde{X})$ is a surjective homomorphism whose kernel is H; by the first isomorphism theorem, this proves the result.

Let $[g] \in N(H)$ be arbitrary, and let $\widetilde{q}\,' = \widetilde{q} \cdot [g]$ as defined in Theorem 11.21. Recall that $\widetilde{q}\,'$ is the terminal point of the lift $\widetilde{g}$ of g starting at $\widetilde{q}$. The key claim is that there exists a covering transformation $\varphi \in \mathcal{C}_p(\widetilde{X})$ such that $\varphi(\widetilde{q}) = \widetilde{q}\,'$.

By the orbit criterion, to prove the claim it suffices to show that $p_*\pi_1(\widetilde{X}, \widetilde{q}\,') = p_*\pi_1(\widetilde{X}, \widetilde{q})$. Let $\Phi_{\widetilde{g}}\colon \pi_1(\widetilde{X}, \widetilde{q}) \to \pi_1(\widetilde{X}, \widetilde{q}\,')$ be the isomorphism determined by the path $\widetilde{g}$ as in Theorem 7.11. From the commutative diagram (11.2) in the proof of the conjugacy theorem, we conclude that

$$\begin{aligned} p_*\pi_1(\widetilde{X}, \widetilde{q}\,') &= p_*\Phi_{\widetilde{g}}\pi_1(\widetilde{X}, \widetilde{q}) \\ &= \Phi_g p_*\pi_1(\widetilde{X}, \widetilde{q}) \\ &= [g]^{-1} \cdot H \cdot [g] = H \\ &= p_*\pi_1(\widetilde{X}, \widetilde{q}). \end{aligned}$$

Thus there exists a covering transformation φ such that $\varphi(\widetilde{q}) = \widetilde{q}\,'$; it is necessarily unique by Proposition 11.27(a). Define $\alpha[g] = \varphi$.

To show that α is a homomorphism, let $[g_1], [g_2] \in N(H)$, and write $\alpha[g_i] = \varphi_i$, so that φ_i is a covering transformation satisfying $\varphi_i(\widetilde{q}) = \widetilde{g}_i(1)$. Let $\varphi_{12} = \alpha[g_1 \cdot g_2]$, so $\varphi_{12}(\widetilde{q}) = \widetilde{g_1 \cdot g_2}(1)$. We need to show that $\varphi_{12} = \varphi_1 \circ \varphi_2$. It suffices to show that these two covering transformations agree at one point, so let us show that $\varphi_{12}(\widetilde{q}) = \varphi_1 \circ \varphi_2(\widetilde{q})$, or equivalently $\widetilde{g_1 \cdot g_2}(1) = \varphi_1(\widetilde{g}_2(1))$.

Now, the lift $\widetilde{g}_2$ of g_2 is a path in $\widetilde{X}$ starting at $\widetilde{q}$. Because $p \circ \varphi_1 = p$, the image $\varphi_1 \circ \widetilde{g}_2$ of $\widetilde{g}_2$ under φ_1 is also a lift of g_2, but this one starts at $\varphi_1(\widetilde{q}) = \widetilde{g}_1(1)$ (Figure 11.9). Thus the path product $\widetilde{g}_1 \cdot (\varphi_1 \circ \widetilde{g}_2)$ makes sense, and is the lift of $g_1 \cdot g_2$ starting at $\widetilde{q}$. In summary,

$$\begin{aligned} \varphi_{12}(\widetilde{q}) &= \widetilde{g_1 \cdot g_2}(1) \\ &= \widetilde{g}_1 \cdot (\varphi_1 \circ \widetilde{g}_2)(1) \\ &= \varphi_1 \circ \widetilde{g}_2(1) \\ &= \varphi_1(\varphi_2(\widetilde{q})), \end{aligned}$$

which was to be proved.

To show that α is surjective, let $\varphi \in \mathcal{C}_p(\widetilde{X})$ be arbitrary, let $\widetilde{q}\,' = \varphi(\widetilde{q})$, and let $\widetilde{g}$ be a path in $\widetilde{X}$ from $\widetilde{q}$ to $\widetilde{q}\,'$. Then $g = p \circ \widetilde{g}$ is a loop in X. Moreover,

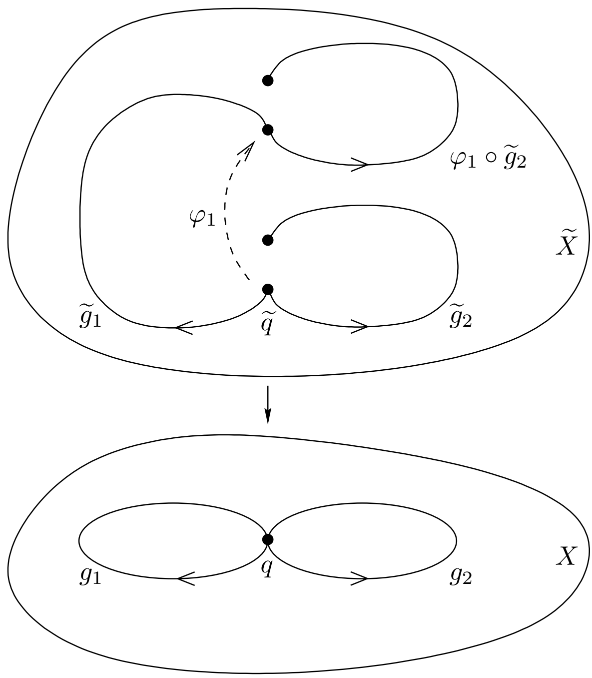


FIGURE 11.9. Proof that α is a homomorphism.

the orbit criterion shows that $p_*\pi_1(\widetilde{X},\widetilde{q}) = p_*\pi_1(\widetilde{X},\widetilde{q}\,')$ because $\widetilde{q}$ and $\widetilde{q}\,'$ are in the same orbit. On the other hand, from (11.2), $p_*\pi_1(\widetilde{X},\widetilde{q}\,') = \Phi_g p_*\pi_1(\widetilde{X},\widetilde{q})$. Thus $\Phi_g p_*\pi_1(\widetilde{X},\widetilde{q}) = p_*\pi_1(\widetilde{X},\widetilde{q})$, which is to say that $[g] \in N(H)$, and the construction above gives $\alpha[g] = \varphi$.

Finally, we need to show that $\operatorname{Ker}\alpha = H$. Let $[g] \in N(H)$, let $\widetilde{g}$ be the lift of g starting at $\widetilde{q}$, and write $\varphi = \alpha[g]$. Then φ is the identity transformation if and only if $\varphi(\widetilde{q}) = \widetilde{g}(1) = \widetilde{q}$, which means that $\widetilde{g}$ is a loop in $\widetilde{X}$; so φ is the identity if and only if $[g] = [p \circ \widetilde{g}] = p_*[\widetilde{g}]$ for some $[\widetilde{g}] \in \pi_1(\widetilde{X},\widetilde{q})$, i.e., $[g] \in H$. $\square$

Problems

11-1. Prove that for any $n \geq 1$ the map $\pi\colon \mathbb{S}^n \to \mathbb{P}^n$ defined in Example 11.4 is a covering map.

11-2. Show that a finite product of covering maps is a covering map: If $p_i : \widetilde{X}_i \to X_i$ are covering maps for $i = 1, \dots, n$, then so is the map

$$p_1 \times \cdots \times p_n : \widetilde{X}_1 \times \cdots \times \widetilde{X}_n \to X_1 \times \cdots \times X_n.$$

11-3. Suppose $p\colon \widetilde{X} \to X$ is a covering map.

(a) If $\widetilde{X}$ is an n-manifold and X is Hausdorff, show that X is an n-manifold.

(b) If X is an n-manifold, show that $\widetilde{X}$ is an n-manifold.

11-4. Suppose $p\colon \widetilde{X} \to X$ is a covering map and X is a compact manifold. Show that $\widetilde{X}$ is compact if and only if p is a finite-sheeted covering.

11-5. Let S be the following subset of $\mathbb{C}^2$:

$$S = \{(z, w) : w^2 = z,\ w \neq 0\}.$$

(It is the graph of the two-valued complex square root "function" described in Chapter 1, with the origin removed.) Show that the projection $\pi_1 \colon \mathbb{C}^2 \to \mathbb{C}$ onto the first coordinate restricts to a two-sheeted covering map $p\colon S \to \mathbb{C} \smallsetminus \{0\}$.

11-6. Show that there is a two-sheeted covering of the Klein bottle by the torus.

11-7. Let $\widetilde{M}$, M, and N be connected manifolds of dimension n and suppose $p\colon \widetilde{M} \to M$ is a k-sheeted covering map. Show that there exists a k-sheeted covering of $M \mathbin{\#} N$ by the connected sum of $\widetilde{M}$ with k copies of N. [Hint: Choose the ball to be cut out of M to lie inside an evenly covered neighborhood.]

11-8. Show that every nonorientable compact surface of genus n has a two-sheeted covering by an orientable one of genus $n-1$. [Hint: Use Problem 11-7 and induction.]

11-9. Show that a proper local homeomorphism between connected, path connected, and locally compact Hausdorff spaces is a covering map.

11-10. A continuous map $f\colon \mathbb{S}^1 \to \mathbb{S}^1$ is said to be *odd* if $f(-z) = -f(z)$ for all $z \in \mathbb{S}^1$, and *even* if $f(z) = f(-z)$ for all $z \in \mathbb{S}^1$. Show that every odd map has odd degree, as follows.

(a) Let $p_2\colon \mathbb{S}^1 \to \mathbb{S}^1$ be the two-sheeted covering map of Example 11.2. If f is odd, show that there exists a continuous map $g\colon \mathbb{S}^1 \to \mathbb{S}^1$ such that $\deg f = \deg g$ and the following diagram commutes:

$$\begin{array}{ccc} \mathbb{S}^1 & \xrightarrow{f} & \mathbb{S}^1 \\ {\scriptstyle p_2}\downarrow & & \downarrow{\scriptstyle p_2} \\ \mathbb{S}^1 & \xrightarrow[g]{} & \mathbb{S}^1. \end{array}$$

(b) If $\deg f$ is even, show that g lifts to a map $\widetilde{g}\colon \mathbb{S}^1 \to \mathbb{S}^1$ such that $p_2 \circ \widetilde{g} = g$.

(c) Show that $\widetilde{g} \circ p_2$ and f are both lifts of $g \circ p_2$ that agree at either $(1,0)$ or $(-1,0)$, so they are equal everywhere; derive a contradiction.

11-11. Show that every even map $f\colon \mathbb{S}^1 \to \mathbb{S}^1$ has even degree.

11-12. Prove the *ham sandwich theorem*: If two pieces of bread and one piece of ham are placed arbitrarily in space, all three pieces can be cut in half with a single slice of the knife. (If you do not like ham, you may wish to substitute tofu.) More precisely, given three disjoint, bounded, connected open subsets $U_1, U_2, U_3 \subset \mathbb{R}^3$, there exists a plane that simultaneously bisects all three, in the sense that the plane divides $\mathbb{R}^3$ into two half spaces H^+ and H^- such that for each i, $U_i \cap H^+$ has the same volume as $U_i \cap H^-$. [Hint: For any $\omega \in \mathbb{S}^2$, show that there are unique real numbers $(\lambda_1, \lambda_2, \lambda_3)$ such that the plane through $\lambda_i \omega$ and orthogonal to ω bisects U_i. If there does not exist a plane bisecting all three sets, define a map $F\colon \mathbb{S}^2 \to \mathbb{S}^1$ by

$$F(\omega) = \frac{(\lambda_1 - \lambda_2) + i(\lambda_2 - \lambda_3)}{\sqrt{(\lambda_1 - \lambda_2)^2 + (\lambda_2 - \lambda_3)^2}}.$$

Show that F is continuous, and $F \circ \iota_{\mathbb{S}^1}$ contradicts the result of Problem 11-10, where $\iota_{\mathbb{S}^1} : \mathbb{S}^1 \hookrightarrow \mathbb{S}^2$ is the inclusion map. You may assume that there is a volume function Vol assigning a nonnegative real number to each open set in $\mathbb{R}^3$ and satisfying the following properties: The volume of a set is unchanged by translations or rotations; the volumes of balls, cylinders, and rectangular solids are given by the usual formulas; and if $U \subset V$ then $\mathrm{Vol}(U) \le \mathrm{Vol}(V)$.]

11-13. Let $p\colon X_3 \to X_2$ be the covering map of Exercise 11.1.

(a) Determine the covering group $\mathcal{C}_p(X_3)$.

(b) Determine whether p is a normal covering.

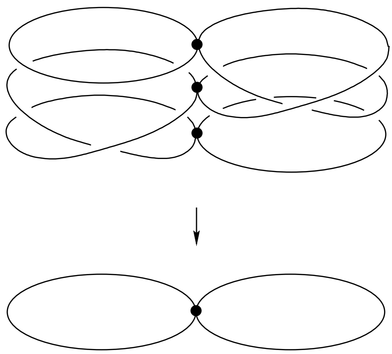


FIGURE 11.10. The covering map of Problem 11-14.

(c) For each of the following maps $f\colon \mathbb{S}^1 \to X_2$, determine whether f has a lift to X_3 taking 1 to 1.

i. $f(z) = z$.

ii. $f(z) = z^2$.

iii. $f(z) = 2 - z$.

iv. $f(z) = 2 - z^2$.

11-14. Let X_4 be the union of four circles described in Problem 10-10, and let $p\colon X_4 \to X_2$ be the covering map indicated schematically in Figure 11.10. Answer the questions of Problem 11-13 for this covering.

11-15. Let $\mathcal{E}$ be the figure eight space of Example 7.21, and let X be the union of the x-axis with infinitely many unit circles centered at $\{2\pi k + i : k \in \mathbb{Z}\}$. Let $p\colon X \to \mathcal{E}$ be the map that sends each circle in X onto the upper circle in $\mathcal{E}$ by translating in the x-direction and sends the x-axis onto the lower circle by $x \mapsto ie^{ix} - i$. You may accept without proof that p is a covering map.

(a) Identify the subgroup $p_*\pi_1(X, 0)$ of $\pi_1(\mathcal{E}, 0)$ in terms of the generators for $\pi_1(\mathcal{E}, 0)$.

(b) Compute the covering group $\mathcal{C}_p(X)$.

(c) Determine whether p is a normal covering.

11-16. This problem shows that the hypothesis that Y is locally path connected is necessary for the lifting criterion (Theorem 11.15) to hold. Let X be the topologist's sine curve (Example 4.10), and let Y be the union of X with a path in the plane from $(1, \sin 1)$ to $(0, 1)$ that intersects X only at those two points.

(a) Show that Y is simply connected.

(b) Show that there is a map $f\colon Y \to \mathbb{S}^1$ that has no lift to $\mathbb{R}$.

11-17. Suppose X is a compact polyhedron and $p\colon \widetilde{X} \to X$ is a covering map.

(a) Show that X and $\widetilde{X}$ admit triangulations such that p is induced by a simplicial map. [Hint: Use barycentric subdivision.]

(b) Suppose $\mathcal{K}$, $\widetilde{\mathcal{K}}$ are finite complexes such that $|\mathcal{K}| = X$, $|\widetilde{\mathcal{K}}| = \widetilde{X}$, and p is induced by a simplicial map from $\widetilde{\mathcal{K}}$ to $\mathcal{K}$. If p is an n-sheeted covering, show that $\chi(\widetilde{\mathcal{K}}) = n\chi(\mathcal{K})$.

12
Classification of Coverings

The main thrust of the preceding chapter was to learn about fundamental groups by studying covering maps. In this chapter we reverse the process and explore what there is to be learned from the fundamental group about the existence and uniqueness of covering spaces. The key idea is provided by the conjugacy theorem of the preceding chapter: Each covering space of X determines a conjugacy class of subgroups in the fundamental group of X.

We begin with the uniqueness question. In the first section of the chapter we define isomorphisms of covering spaces, and show that two covering spaces are isomorphic if and only if they induce the same conjugacy class of subgroups.

Then we address the existence question. The ultimate goal is to show that for a sufficiently nice space X (any connected manifold, for example), every conjugacy class of subgroups of $\pi_1(X, q)$ corresponds to some covering. This is accomplished in several stages. First we show that X has a unique simply connected covering space, called its "universal covering space." Then we show how to construct coverings as quotients of a given space by certain group actions. The dénouement is the last theorem of the chapter, which puts together all the preceding results to give a complete classification of all coverings of X up to isomorphism: They are in one-to-one correspondence with conjugacy classes of subgroups of the fundamental group of X. We illustrate the theory by determining the universal covering spaces of all the compact surfaces and classifying all the coverings of the torus.

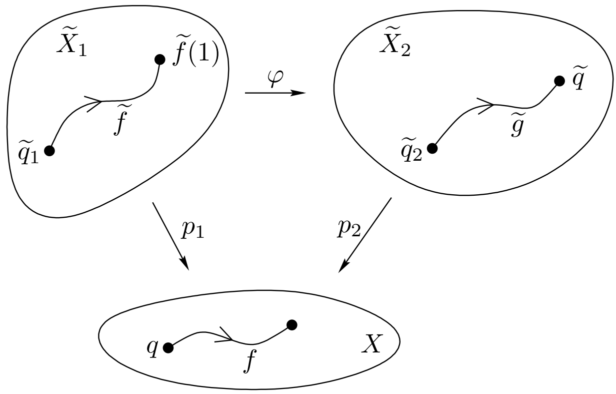


FIGURE 12.1. Proof that a covering homomorphism is surjective.

Covering Homomorphisms

In this section we examine the question of how to tell when two covering spaces are "the same." As usual, we consider two coverings the same if they are related by a suitable isomorphism. We begin by defining some terms.

Let X be a space, and let $p_1 : \widetilde{X}_1 \to X$, $p_2 : \widetilde{X}_2 \to X$ be two coverings of X. A *covering homomorphism* from p_1 to p_2 is a continuous map $\varphi : \widetilde{X}_1 \to \widetilde{X}_2$ such that $p_2 \circ \varphi = p_1$:

$$\begin{tikzcd} \widetilde{X}_1 \arrow[rr, "\varphi"] \arrow[dr, "p_1"'] & & \widetilde{X}_2 \arrow[dl, "p_2"] \\ & X. & \end{tikzcd}$$

A covering homomorphism that is also a homeomorphism is said to be an *isomorphism* of coverings. It is easy to see that in this case the inverse map is also a covering homomorphism. We say two coverings are *isomorphic* if there is an isomorphism between them; this is an equivalence relation on the set of coverings of X. Note that an isomorphism from a covering to itself is just a covering transformation.

An interesting feature of covering homomorphisms is that they are themselves covering maps, as the following lemma shows.

Lemma 12.1. *Let $p_1 : \widetilde{X}_1 \to X$ and $p_2 : \widetilde{X}_2 \to X$ be coverings of X, and let φ be a covering homomorphism from p_1 to p_2. Then φ is a covering map.*

Proof. First we show that φ is surjective. Let $\widetilde{q} \in \widetilde{X}_2$ be arbitrary. Choose some $\widetilde{q}_1 \in \widetilde{X}_1$, and let $\widetilde{q}_2 = \varphi(\widetilde{q}_1) \in \widetilde{X}_2$, $q = p_1(\widetilde{q}_1) = p_2(\widetilde{q}_2) \in X$ (Figure 12.1). There is a path $\widetilde{g}$ in $\widetilde{X}_2$ from $\widetilde{q}_2$ to $\widetilde{q}$. Let $f = p_2 \circ \widetilde{g}$, which is a

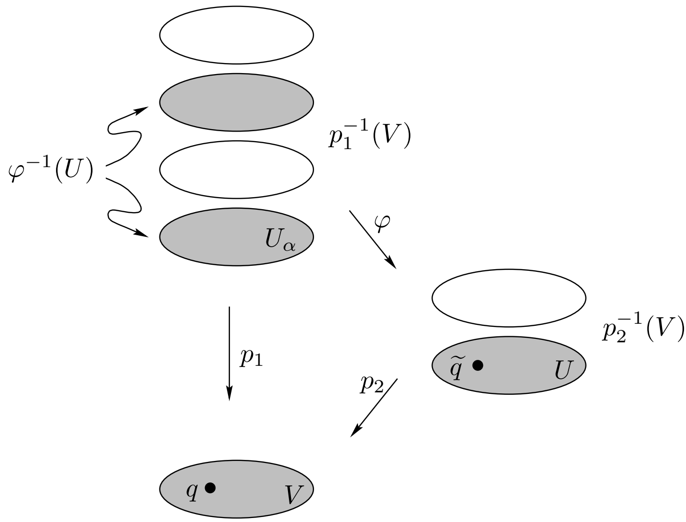


FIGURE 12.2. An evenly covered neighborhood of $\widetilde{q}$.

path in X starting at q, and let $\widetilde{f}$ be the unique lift of f to a path in $\widetilde{X}_1$ starting at $\widetilde{q}_1$. Consider now the path $\varphi \circ \widetilde{f}$ in $\widetilde{X}_2$. Its initial point is $\varphi \circ \widetilde{f}(0) = \varphi(\widetilde{q}_1) = \widetilde{q}_2$, and it satisfies $p_2 \circ \varphi \circ \widetilde{f} = p_1 \circ \widetilde{f} = f$, so $\varphi \circ \widetilde{f}$ is the lift of f to $\widetilde{X}_2$ starting at $\widetilde{q}_2$. By the unique lifting property, this means that $\varphi \circ \widetilde{f} = \widetilde{g}$, so

$$\varphi(\widetilde{f}(1)) = \widetilde{g}(1) = \widetilde{q},$$

which shows that φ is surjective.

To show that φ is a covering map, let $\widetilde{q} \in \widetilde{X}_2$ be arbitrary; let $q = p_2(\widetilde{q}) \in X$; let $U_1, U_2 \subset X$ be neighborhoods of q that are evenly covered by p_1 and p_2, respectively; and let V be the component of $U_1 \cap U_2$ containing q. Thus V is a neighborhood of q that is evenly covered by both p_1 and p_2.

Let U be the component of $p_2^{-1}(V)$ containing $\widetilde{q}$. We need to show that the components of $\varphi^{-1}(U)$ are mapped homeomorphically onto U by φ. Consider the restrictions of p_1 and φ to the "stack of pancakes" $p_1^{-1}(V)$ (Figure 12.2). Since U is both open and closed in $p_2^{-1}(V)$, it follows that $\varphi^{-1}(U)$ is both open and closed in $p_1^{-1}(V)$, and is thus a union of components. On any such component U_α, the following diagram commutes:

$$\begin{tikzcd}
U_\alpha \arrow[dd, "p_1"'] \arrow[dr, "\varphi"] & \\
& U \arrow[dl, "p_2"] \\
V. &
\end{tikzcd}$$

Since p_1 and p_2 are homeomorphisms in this diagram, so is φ. □

The key to determining when two covering spaces are isomorphic is to decide when there are covering homomorphisms between them. This question is answered by the following theorem.

Theorem 12.2 (Covering Homomorphism Criterion). *Suppose $p_1 \colon \widetilde{X}_1 \to X$ and $p_2 \colon \widetilde{X}_2 \to X$ are two coverings of X, and $\widetilde{q}_1 \in \widetilde{X}_1$, $\widetilde{q}_2 \in \widetilde{X}_2$ are base points such that $p_1(\widetilde{q}_1) = p_2(\widetilde{q}_2) = q \in X$. There exists a covering homomorphism from p_1 to p_2 taking $\widetilde{q}_1$ to $\widetilde{q}_2$ if and only if $p_{1*}\pi_1(\widetilde{X}_1, \widetilde{q}_1) \subset p_{2*}\pi_1(\widetilde{X}_2, \widetilde{q}_2)$.*

Proof. A covering homomorphism from p_1 to p_2 can also be viewed as a lift of p_1:

$$\begin{tikzcd} & \widetilde{X}_2 \arrow[d, "p_2"] \\ \widetilde{X}_1 \arrow[ur, "\varphi"] \arrow[r, "p_1"'] & X. \end{tikzcd} \tag{12.1}$$

Thus both the necessity and the sufficiency of the subgroup condition follow from the lifting criterion (Theorem 11.15). □

Example 12.3. Let $p_n \colon \mathbb{S}^1 \to \mathbb{S}^1$ be the nth power map defined in Example 11.2. The subgroup of $\pi_1(\mathbb{S}^1, 1)$ induced by p_n is the cyclic subgroup generated by $[\alpha]^n$ (Example 11.18). By the covering homomorphism criterion, there is a homomorphism from p_m to p_n if and only if m is divisible by n; the homomorphism in that case is just $p_{m/n}$.

Example 12.4. Consider the following two coverings of $\mathbb{T}^2$: $E \colon \mathbb{R}^2 \to \mathbb{T}^2$ is the covering of Example 11.3 (the product of two copies of $\varepsilon \colon \mathbb{R} \to \mathbb{S}^1$); and $p \colon \mathbb{S}^1 \times \mathbb{R} \to \mathbb{T}^2$ is given by $p(z, y) = (z, \varepsilon(y))$. Writing $\pi_1(\mathbb{T}^2) \cong \langle\beta\rangle \times \langle\gamma\rangle$, we see that $E_*\pi_1(\mathbb{R}^2)$ is trivial, while $p_*\pi_1(\mathbb{S}^1 \times \mathbb{R}) = \langle\beta\rangle \times \{1\}$. Therefore, there exists a covering homomorphism from E to p. (Why do the base points not matter?) It is easy to check that $\varphi(x, y) = (\varepsilon(x), y)$ is such a homomorphism.

The following theorem completely solves the uniqueness question for covering spaces up to isomorphism.

Theorem 12.5 (Covering Isomorphism Theorem). *Two coverings $p_1 \colon \widetilde{X}_1 \to X$ and $p_2 \colon \widetilde{X}_2 \to X$ are isomorphic if and only if for some $q \in X$ and base points $\widetilde{q}_1 \in p_1^{-1}(q)$ and $\widetilde{q}_2 \in p_2^{-1}(q)$, the induced subgroups $p_{1*}\pi_1(\widetilde{X}_1, \widetilde{q}_1)$ and $p_{2*}\pi_1(\widetilde{X}_2, \widetilde{q}_2)$ are conjugate in $\pi_1(X, q)$. If this is the case, these subgroups are conjugate for every such q, $\widetilde{q}_1$, and $\widetilde{q}_2$.*

Proof. If there exists an isomorphism $\varphi\colon \widetilde{X}_1 \to \widetilde{X}_2$, choose $\widetilde{q}_1 \in \widetilde{X}_1$ arbitrarily and set $\widetilde{q}_2 = \varphi(\widetilde{q}_1)$. The covering homomorphism criterion applied to φ and φ^{-1} guarantees that the two subgroups $p_{1*}\pi_1(\widetilde{X}_1, \widetilde{q}_1)$ and $p_{2*}\pi_1(\widetilde{X}_2, \widetilde{q}_2)$ are contained in each other, so they are equal. Thus by the conjugacy theorem (Theorem 11.19), the subgroups associated with any other choices of base points in the same fibers are conjugate.

Conversely, suppose the two subgroups are conjugate for some choice of q, $\widetilde{q}_1$, and $\widetilde{q}_2$. By the conjugacy theorem, we can change to a new base point $\widetilde{q}_2' \in \widetilde{X}_2$ such that $p_{2*}\pi_1(\widetilde{X}_2, \widetilde{q}_2') = p_{1*}\pi_1(\widetilde{X}_1, \widetilde{q}_1)$. Then by the covering homomorphism criterion there exist homomorphisms φ from p_1 to p_2 and ψ from p_2 to p_1, with $\varphi(\widetilde{q}_1) = \widetilde{q}_2'$ and $\psi(\widetilde{q}_2') = \widetilde{q}_1$. The composite map $\psi \circ \varphi$ is a covering transformation of p_1 that fixes $\widetilde{q}_1$, so it is the identity. Similarly, $\varphi \circ \psi$ is the identity, so φ is the required isomorphism. □

The Universal Covering Space

When the results of the preceding section are applied to simply connected covering spaces, they yield some extremely useful results.

Proposition 12.6 (Properties of Simply Connected Coverings).

(*a*) *Let $p\colon \widetilde{X} \to X$ be a covering map with $\widetilde{X}$ simply connected. If $p_1\colon \widetilde{X}_1 \to X$ is any covering, there exists a covering map $\widetilde{p}\colon \widetilde{X} \to \widetilde{X}_1$ such that the following diagram commutes:*

$$\begin{tikzcd} \widetilde{X} \arrow[dd, "p"'] \arrow[dr, "\widetilde{p}"] & \\ & \widetilde{X}_1 \arrow[dl, "p_1"] \\ X. & \end{tikzcd}$$

(*b*) *Any two simply connected coverings of the same space are isomorphic.*

Proof. Since the trivial subgroup is contained in every other subgroup, part (a) follows from the covering homomorphism criterion and the fact that every covering homomorphism is a covering map. Part (b) follows immediately from the covering isomorphism theorem. □

Part (a) of this proposition says that a simply connected covering space covers every other covering space of X. Because of this, any covering of X by a simply connected space $\widetilde{X}$ (which is unique by (b)) is called a *universal covering*, and $\widetilde{X}$ is called the *universal covering space* of X.

Example 12.7. The universal covering space of the n-torus is $\mathbb{R}^n$, because we constructed a covering map $E\colon \mathbb{R}^n \to \mathbb{T}^n$ in Example 11.3. The universal covering space of $\mathbb{P}^n$ is $\mathbb{S}^n$, by the covering map π of Example 11.4.

As the next theorem shows, every "reasonable" space, including every manifold, has a universal covering space. We say that a space X is *locally simply connected* if it admits a basis of simply connected open sets. Clearly, a locally simply connected space is locally path connected, because simply connected sets are path connected. Any manifold is locally simply connected, because it has a basis of Euclidean balls.

Theorem 12.8 (Existence of the Universal Covering Space). *Every connected and locally simply connected topological space (in particular, every connected manifold) has a universal covering space.*

Proof. To get an idea how to proceed, suppose for a moment that X does have a universal covering $p\colon \widetilde{X} \to X$. The key fact is that once we choose base points $\widetilde{q}_0 \in \widetilde{X}$ and $q_0 = p(\widetilde{q}_0) \in X$, the fiber $p^{-1}(q)$ over any $q \in X$ is in one-to-one correspondence with path classes from q_0 to q. To see why, define a map E from the set of such path classes to $p^{-1}(q)$ by sending $[f]$ to the terminal point of the lift of f starting at $\widetilde{q}_0$. Since lifts of homotopic paths have the same terminal point by the monodromy theorem, E is well-defined. E is surjective, because given any $\widetilde{q}$ in the fiber over q, there is a path $\widetilde{f}$ from $\widetilde{q}_0$ to $\widetilde{q}$, and then $p \circ \widetilde{f}$ is a path from q_0 to q whose lift ends at $\widetilde{q}$. Injectivity of E follows from the fact that $\widetilde{X}$ is simply connected: If f_1, f_2 are two paths from q_0 to q whose lifts $\widetilde{f}_1, \widetilde{f}_2$ end at the same point, then $\widetilde{f}_1$ and $\widetilde{f}_2$ are path homotopic, and therefore so are $f_1 = p \circ \widetilde{f}_1$ and $f_2 = p \circ \widetilde{f}_2$.

Now let X be any space satisfying the hypotheses of the theorem, and choose any base point $q_0 \in X$. Guided by the observation in the preceding paragraph, we define $\widetilde{X}$ to be the set of path classes of paths in X starting at q_0, and define $p\colon \widetilde{X} \to X$ by $p[f] = f(1)$, which is well-defined because path homotopic paths have the same terminal point. We will prove that $\widetilde{X}$ has the required properties in a series of steps.

STEP 1: *Topologize $\widetilde{X}$.* We define a topology on $\widetilde{X}$ by constructing a basis. For any $[f] \in \widetilde{X}$ and any simply connected open set $U \subset X$ containing $f(1)$, define the set $[f \cdot U] \subset \widetilde{X}$ by

$$[f \cdot U] = \{[f \cdot a] : a \text{ is a path in } U \text{ starting at } f(1)\}.$$

Let $\mathcal{B}$ denote the collection of all such sets $[f \cdot U]$; we will show that $\mathcal{B}$ is a basis. First, since X is locally simply connected, for any $[f] \in \widetilde{X}$ there exists a simply connected open set U containing $f(1)$, and clearly $[f] \in [f \cdot U]$. Thus the union of all the sets in $\mathcal{B}$ is $\widetilde{X}$.

To check the intersection condition, suppose $[h] \in \widetilde{X}$ is in the intersection of two basis sets $[f \cdot U], [g \cdot V] \in \mathcal{B}$. This means that $h \sim f \cdot a \sim g \cdot b$, where a is a path in U and b is a path in V (Figure 12.3). Let W be a simply connected neighborhood of $h(1)$ contained in $U \cap V$ (such a neighborhood exists because X has a basis of simply connected open sets). If $[h \cdot c]$ is any element of $[h \cdot W]$, then $[h \cdot c] = [f \cdot a \cdot c] \in [f \cdot U]$ because $a \cdot c$ is a path in

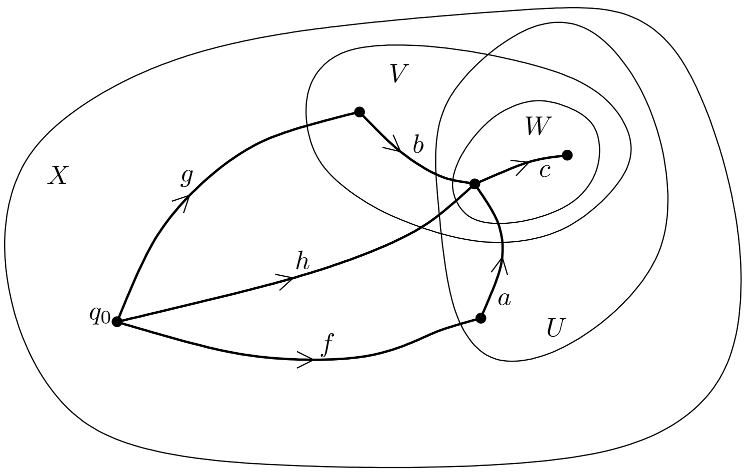


FIGURE 12.3. Proof that the collection of sets $[f \cdot U]$ is a basis.

U. Similarly, $[h \cdot c] = [g \cdot b \cdot c] \in [g \cdot V]$. Thus $[h \cdot W]$ is a basis set contained in $[f \cdot U] \cap [g \cdot V]$, which proves that $\mathcal{B}$ is a basis. From now on, we endow $\widetilde{X}$ with the topology generated by $\mathcal{B}$.

STEP 2: *$\widetilde{X}$ is path connected.* Let $[f] \in \widetilde{X}$ be arbitrary. We will show that there is a path in $\widetilde{X}$ from $\widetilde{q}_0$ to $[f]$, where $\widetilde{q}_0 = [c_{q_0}]$.

For any $0 \leq t \leq 1$, define $f_t \colon I \to X$ by

$$f_t(s) = f(ts),$$

so f_t is a path in X from q_0 to $f(t)$. Then define $\widetilde{f} \colon I \to \widetilde{X}$ by

$$\widetilde{f}(t) = [f_t].$$

Clearly, $\widetilde{f}(0) = [f_0] = \widetilde{q}_0$, and $\widetilde{f}(1) = [f_1] = [f]$. So we need only show that $\widetilde{f}$ is continuous; for this it suffices to show that the inverse image under $\widetilde{f}$ of any basis open set $[h \cdot U] \subset \widetilde{X}$ is open. Let $t_0 \in I$ be a point such that $\widetilde{f}(t_0) \in [h \cdot U]$ (Figure 12.4). This means that $f_{t_0} \sim h \cdot c$ for some path c lying in U, and in particular that $f(t_0) = f_{t_0}(1) \in U$. For any $0 \leq t \leq 1$, define a path $f_{t_0 t}$ by

$$f_{t_0 t}(s) = f(t_0 + s(t - t_0)).$$

This path just follows f from $f(t_0)$ to $f(t)$, so $f_{t_0} \cdot f_{t_0 t}$ is easily seen to be path homotopic to f_t.

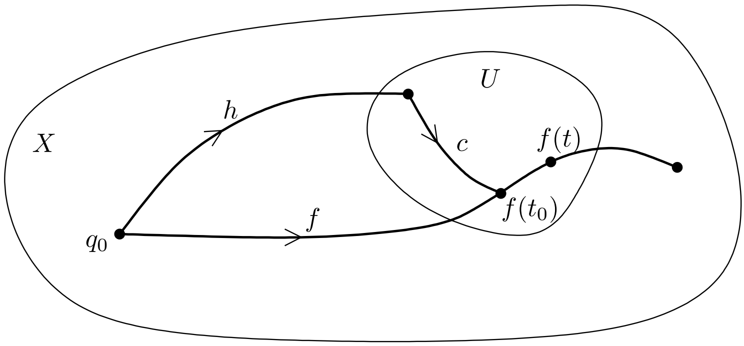


FIGURE 12.4. Proof that $\widetilde{X}$ is path connected.

By continuity of f, there is some $\delta > 0$ such that $f(t_0 - \delta, t_0 + \delta) \subset U$. If $t \in (t_0 - \delta, t_0 + \delta)$, then

$$f_t \sim f_{t_0} \cdot f_{t_0 t} \sim h \cdot c \cdot f_{t_0 t},$$

from which it follows that

$$\widetilde{f}(t) = [f_t] = [h \cdot c \cdot f_{t_0 t}] \in [h \cdot U].$$

This shows that $\widetilde{f}^{-1}[h \cdot U]$ contains the set $(t_0 - \delta, t_0 + \delta)$, so $\widetilde{f}$ is continuous.

STEP 3: *p is a covering map.* Let $U \subset X$ be any simply connected open set. We will show that U is evenly covered.

Choose any point $q_1 \in U$. We begin by showing that $p^{-1}(U)$ is the disjoint union of the sets $[f \cdot U]$ as $[f]$ varies over all the distinct path classes from q_0 to q_1. It is obvious from the definition of p that $p[f \cdot U] \subset U$, so $\bigcup_{[f]} [f \cdot U] \subset p^{-1}(U)$. Conversely, if $[g] \in p^{-1}(U)$, then $g(1) = p[g] \in U$, so there is a path b in U from $g(1)$ to q_1, and $[g] = [g \cdot b \cdot b^{-1}] \in [(g \cdot b) \cdot U]$. This proves that $p^{-1}(U) = \bigcup_{[f]} [f \cdot U]$.

This shows, in particular, that p is continuous: X has a basis of simply connected open sets, and the inverse image under p of any such set is a union of basis sets and therefore open. And p is clearly surjective, because each $q \in X$ is equal to $p[g]$ for any path g from q_0 to q.

Next we show that p is a homeomorphism from each set $[f \cdot U]$ to U. It is surjective because for each $q \in U$ there is a path a from $f(1)$ to q in U, so $q = p[f \cdot a] \in p[f \cdot U]$. To see that it is injective, let $[g], [g'] \in [f \cdot U]$, and suppose $p[g] = p[g']$, or in other words, $g(1) = g'(1)$ (Figure 12.5). Then by definition of $[f \cdot U]$, $g \sim f \cdot a$ and $g' \sim f \cdot a'$ for some paths a, a' in U from $f(1)$ to $g(1)$. Since U is simply connected, $a \sim a'$ and therefore $[g] = [g']$. Finally, p is an open map because it takes basis open sets to open sets, and therefore $p \colon [f \cdot U] \to U$ is a homeomorphism.

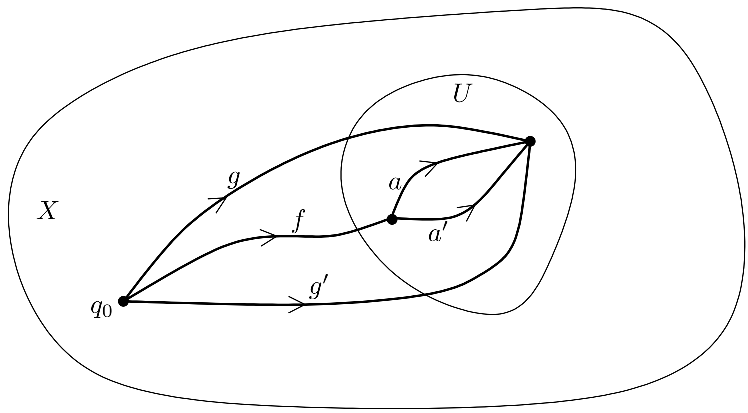


FIGURE 12.5. Proof that p is injective on $[f \cdot U]$.

Each set $[f \cdot U]$ is open by definition, and each is path connected because it is homeomorphic to the path connected set U. It follows that $\widetilde{X}$ is locally path connected. To complete the proof that p is a covering map, we need to show that for any two paths f and f' from q_0 to q_1, the sets $[f \cdot U]$ and $[f' \cdot U]$ are either equal or disjoint. If they are not disjoint, there exists $[g] \in [f \cdot U] \cap [f' \cdot U]$, so $g \sim f \cdot a \sim f' \cdot a'$ for paths a, a' in U from q_1 to $g(1)$. Since U is simply connected, $a \sim a'$, which implies $f \sim f'$ and therefore $[f \cdot U] = [f' \cdot U]$.

STEP 4: *$\widetilde{X}$ is simply connected.* Suppose $F\colon I \to \widetilde{X}$ is any loop based at $\widetilde{q}_0$. Let $f = p \circ F$, so F is a lift of f. If we write $\widetilde{f}(t) = [f_t]$ as in Step 3, then $p \circ \widetilde{f}(t) = p[f_t] = f_t(1) = f(t)$, so $\widetilde{f}$ is also a lift of f starting at $\widetilde{q}_0$. By the unique lifting property, $F = \widetilde{f}$. Since F is a loop,

$$[c_{q_0}] = \widetilde{q}_0 = F(1) = \widetilde{f}(1) = [f_1] = [f],$$

so f is null homotopic. By the homotopy lifting property, this means that F is null homotopic as well. □

A careful study of this proof shows that it does not really need the full strength of the hypothesis that X is locally simply connected. Each time we use the fact that a loop in a small open set $U \subset X$ is null homotopic, all we really need to know is that it is null homotopic *in* X. For this reason, it is traditional to make the following definition: A space X is *semilocally simply connected* if it admits a basis of open sets U with the property that every loop in U is null homotopic in X. It can be shown that a connected, locally path connected space admits a universal covering space if and only if it is semilocally simply connected (see [Mas89] or [Sie92]). Since our motivation for studying the fundamental group is to understand manifolds, we have no need for this extra generality.

Once you have understood the proof of the existence of the universal covering space of a space X, you should forget the complicated construction of $\widetilde{X}$ in terms of path classes, and just think of $\widetilde{X}$ as a simply connected space with a covering map to X. The uniqueness theorem tells us that all the relevant properties of $\widetilde{X}$ can be derived from these facts.

Proper Group Actions

The next step in classifying coverings is to start with a space Y and develop a technique for constructing spaces *covered by* Y. In the next section we will apply this to the universal covering space in order to derive a classification theorem for coverings of a given space X.

To get an idea how to construct spaces covered by Y, let us suppose $p\colon \widetilde{X} \to X$ is a normal covering. (The restriction to normal coverings will not be a limitation in the end: For reasons that will soon become apparent, the construction in this section will produce only normal coverings, but in the next section we will be able to use them to produce *all* coverings of a given space.)

As we observed in the previous chapter, the covering group $\mathcal{C}_p(\widetilde{X})$ acts continuously and freely on $\widetilde{X}$ (on the left). The orbit criterion (Proposition 11.29) says that $\mathcal{C}_p(\widetilde{X})$ acts transitively on each fiber when p is normal, so the identifications made by p are exactly those determined by the equivalence relation $x \sim y$ if and only if $y = \varphi(x)$ for some $\varphi \in \mathcal{C}_p(\widetilde{X})$. Since p is a quotient map by Lemma 11.5, X is homeomorphic to the orbit space determined by the left action of $\mathcal{C}_p(\widetilde{X})$ on $\widetilde{X}$ (see Chapter 3).

Now let Y be any space, and suppose we are given a left action by a group Γ on Y. Our aim in this section is to describe conditions under which the quotient map $\pi\colon Y \to Y/\Gamma$ onto the orbit space is a covering map whose covering group is Γ. Note that this construction can produce only normal coverings, because Γ acts transitively on the fibers of any orbit space by definition.

Not every group action yields a covering map, of course. Clearly, the action must be continuous and free (Proposition 11.27(b)). Moreover, every point of a covering space Y has a neighborhood (one of the "pancakes" over an evenly covered open set) whose images under the covering group are all disjoint, which places a strong restriction on the actions we can consider.

A simple condition that will guarantee that a group action has the requisite properties in all cases of interest to us is the following. A continuous action of a topological group Γ on a space Y is said to be *proper* if the map $\Gamma \times Y \to Y \times Y$ given by $(g, y) \mapsto (y, g \cdot y)$ is a proper map, i.e., if the inverse image of any compact set under this map is compact.

Proper actions have the following useful alternative characterizations, at least for discrete group actions on locally compact Hausdorff spaces, the only type we will be concerned with. For any $g \in \Gamma$ and any subset $K \subset Y$, we let $g \cdot K = \{g \cdot y : y \in K\}$.

Proposition 12.9. *For a discrete group Γ acting on a locally compact Hausdorff space Y, the following are equivalent:*

(*a*) *Γ acts properly.*

(*b*) *For any compact set $K \subset Y$, $K \cap (g \cdot K) = \varnothing$ for all but finitely many $g \in \Gamma$.*

(*c*) *For every $y, y' \in Y$, there exist neighborhoods U of y and U' of y' such that $U \cap (g \cdot U') = \varnothing$ for all but finitely many $g \in \Gamma$.*

Proof. We will show (a) $\implies$ (b) $\implies$ (c) $\implies$ (a). Assume first that the action of Γ is proper, and let $\Phi\colon \Gamma \times Y \to Y \times Y$ denote the proper map $\Phi(g, y) = (y, g \cdot y)$. Given any compact set $K \subset Y$, the set

$$\Phi^{-1}(K \times K) = \{(g, y) \in \Gamma \times Y : y \in K,\ \ g \cdot y \in K\}$$

is compact. Thus its projection onto Γ is compact and therefore finite. But this projection includes all $g \in \Gamma$ such that $K \cap (g \cdot K) \neq \varnothing$, so this proves (a) $\implies$ (b).

Now suppose (b) holds. Because Y is locally compact Hausdorff, any points $y, y' \in Y$ have precompact neighborhoods U and U', respectively. Let K be the compact set $\overline{U} \cup \overline{U'}$. Then the set of $g \in \Gamma$ such that $K \cap (g \cdot K) = \varnothing$ is finite, which implies (c).

Finally, if (c) holds, let L be an arbitrary compact subset of $Y \times Y$. For any $(y, y') \in L$, choose neighborhoods U of y and U' of y' as in (c), so $U \times U'$ is a neighborhood of (y, y'). The set of such product neighborhoods as (y, y') ranges over L is an open cover of L, so finitely many such neighborhoods $U_1 \times U_1', \dots, U_m \times U_m'$ cover L. For each i, the set $S_i = \{g \in \Gamma : U_i \cap (g^{-1} \cdot U_i') \neq \varnothing\}$ is finite. Let $S = S_1 \cup \dots \cup S_m$ and $K = \pi_1(L) \subset Y$. Then it is straightforward to check that $\Phi^{-1}(L)$ is contained in the compact set $S \times K$. Since Y is Hausdorff, L is closed in $Y \times Y$; and since Φ is continuous, $\Phi^{-1}(L)$ is a closed subset of a compact set and thus compact. □

Corollary 12.10. *If a discrete group Γ acts freely and properly on a locally compact Hausdorff space Y, every point $y \in Y$ has a neighborhood U such that $U \cap (g \cdot U) = \varnothing$ unless $g = 1$.*

Proof. Taking $y = y'$ in Proposition 12.9(c), we obtain neighborhoods U and U' of y such that $U \cap (g \cdot U') = \varnothing$ except for finitely many group elements $1, g_1, \dots, g_m \in \Gamma$. Since the action is free and Y is Hausdorff, for each g_i there are disjoint neighborhoods W_i of y and W_i' of $g_i \cdot y$. Let

$$\widetilde{U} = U \cap U' \cap W_1 \cap (g_1^{-1} \cdot W_1') \cap \cdots \cap W_m \cap (g_m^{-1} \cdot W_m').$$

We will show that $\widetilde{U}$ has the required properties.

First consider $g = g_i$ for some i. If $y \in \widetilde{U} \subset g_i^{-1} \cdot W_i'$, then $g_i \cdot y \in W_i'$, which is disjoint from W_i and therefore from $\widetilde{U}$. Thus $\widetilde{U} \cap (g_i \cdot \widetilde{U}) = \varnothing$. On the other hand, if $g \in \Gamma$ is not the identity and not one of the g_i's, then for any $y \in \widetilde{U} \subset U'$, we have $g \cdot y \in g \cdot U'$, which is disjoint from U and therefore also from $\widetilde{U}$. □

A group action possessing the property expressed in this corollary, or that expressed in part (c) of Proposition 12.9, or something closely related to these (depending on whom you read) has traditionally been called *properly discontinuous.* This is a particularly unfortunate term, because the group actions we are interested in are all continuous, so one is forced to speak of a "continuous properly discontinuous action." We will avoid the problem by working only with proper actions, which have many important applications in topology and geometry, and are quite sufficient as long as we confine our attention to locally compact Hausdorff spaces.

Theorem 12.11. *Let Y be a connected, locally path connected, locally compact Hausdorff space (for example, a connected manifold), and suppose a discrete group Γ acts continuously, freely, and properly on Y. Then Y/Γ is Hausdorff, the quotient map $\pi\colon Y \to Y/\Gamma$ is a normal covering map, and $\mathcal{C}_\pi(Y) = \Gamma$, considered as a group of homeomorphisms of Y.*

Proof. Clearly, π is surjective and continuous. In fact, it is an open map, for the following reason: If $U \subset Y$ is open, then $\pi^{-1}(\pi(U))$ is the union of all sets of the form $g \cdot U$ as g ranges over Γ. This is a union of open sets and therefore open, so $\pi(U)$ is open.

To show that π is a covering map, let $y \in Y$, and choose a neighborhood U of y as in Corollary 12.10. Let $\widetilde{V} \subset U$ be the component of U containing y; clearly, $\widetilde{V}$ still has the property that its images under Γ are disjoint. Let $V = \pi(\widetilde{V})$, which is open in Y/Γ because π is an open map.

Now, $\pi^{-1}(V)$ is the union of the disjoint connected open sets $g \cdot \widetilde{V}$ for $g \in \Gamma$, so to show that π is a covering it remains only to show that π is a homeomorphism from each such set onto V. Because for each $g \in \Gamma$, $g\colon \widetilde{V} \to g \cdot \widetilde{V}$ is a homeomorphism and the diagram

$$\begin{tikzcd}
\widetilde{V} \arrow[rr, "g"] \arrow[dr, "\pi"'] & & g \cdot \widetilde{V} \arrow[dl, "\pi"] \\
& V &
\end{tikzcd}$$

commutes, it suffices to show that $\pi\colon \widetilde{V} \to V$ is a homeomorphism. It is surjective, continuous, and open; and it is injective because $\pi(v) = \pi(v')$ for

$v, v' \in \widetilde{V}$ implies $v' = g \cdot v$ for some $g \in \Gamma$, so $v = v'$ because $\widetilde{V} \cap (g \cdot \widetilde{V}) = \varnothing$ when $g \neq 1$. This proves that π is a covering map.

If $g \in \Gamma$, then $x \mapsto g \cdot x$ is a covering transformation, since $\pi(g \cdot x) = \pi(x)$ by definition; thus $\Gamma \subset \mathcal{C}_\pi(Y)$. By construction, Γ acts transitively on each fiber, so π is a normal covering. If φ is any covering transformation, choose $y \in Y$ and let $y' = \varphi(y)$. Then there is some $g \in \Gamma$ such that $g \cdot y = y'$; since φ and $x \mapsto g \cdot x$ are covering transformations that agree at a point, they are equal. Thus Γ is the full covering group.

To show that the quotient space is Hausdorff, let $\Phi(g, y) = (y, g \cdot y)$ as in the proof of Proposition 12.9. Since Φ is proper, it is closed by Proposition 4.32, so $\Phi(\Gamma \times Y)$ is a closed subset of $Y \times Y$. Let $x, x' \in Y/\Gamma$ be distinct points. Choosing $y, y' \in Y$ such that $\pi(y) = x$ and $\pi(y') = x'$, the fact that $x \neq x'$ means that $(y, y') \notin \Phi(\Gamma \times Y)$. Therefore, (y, y') has a product neighborhood $U \times U' \subset Y \times Y$ that is disjoint from $\Phi(\Gamma \times Y)$. Since π is open, $\pi(U)$ and $\pi(U')$ are neighborhoods of x and x', respectively. Any point $z \in \pi(U) \cap \pi(U')$ would satisfy $z = \pi(v) = \pi(v')$ for some $v \in U$ and $v' \in U'$. But this would mean that $v' = g \cdot v$ for some $g \in \Gamma$, so $(v, v') = (v, g \cdot v) \in \Phi(\Gamma \times Y)$, which contradicts the fact that $U \times U'$ is disjoint from the image of Φ. Thus $\pi(U) \cap \pi(U') = \varnothing$, which shows that Y/Γ is Hausdorff. □

Corollary 12.12. *Let $\widetilde{M}$ be a connected n-manifold on which a discrete group Γ acts continuously, freely, and properly. Then $\widetilde{M}/\Gamma$ is an n-manifold, and the quotient map $\pi \colon \widetilde{M} \to \widetilde{M}/\Gamma$ is a normal covering map.*

Proof. We know from Theorem 12.11 that π is a normal covering map and $\widetilde{M}/\Gamma$ is Hausdorff, and therefore it is a manifold by Problem 11-3(a). □

Example 12.13 (Lens Spaces). By identifying $\mathbb{R}^4$ with $\mathbb{C}^2$ in the obvious way, we can consider $\mathbb{S}^3$ as the following subset of $\mathbb{C}^2$:

$$\mathbb{S}^3 = \{(z_1, z_2) \in \mathbb{C}^2 : |z_1|^2 + |z_2|^2 = 1\}.$$

Fix a pair of relatively prime integers $1 \leq m < n$, and define an action of $\mathbb{Z}/\langle n \rangle$ on $\mathbb{S}^3$ by

$$k \cdot (z_1, z_2) = \big(e^{2\pi i k/n} z_1, e^{2\pi i k m/n} z_2\big).$$

It can easily be checked that this action is free, and it is proper because $\mathbb{Z}/\langle n \rangle$ is a finite group (Problem 12-6). The orbit space $\mathbb{S}^3/(\mathbb{Z}/\langle n \rangle)$ is thus a compact 3-manifold whose universal covering space is $\mathbb{S}^3$ and whose fundamental group is isomorphic to $\mathbb{Z}/\langle n \rangle$. This manifold, denoted by $L(n, m)$, is called a *lens space*.

A particularly important example of a free proper action arises when we consider a topological group G and a discrete subgroup Γ (that is, a subgroup that is a discrete subspace). Recall from Chapter 3 that left translation defines a left action of Γ on G whose quotient is the coset space G/Γ.

Proposition 12.14. *Let Γ be a discrete subgroup of a connected, locally path connected, locally compact Hausdorff topological group G. Then Γ acts freely and properly on G by left translations, so the quotient map $\pi\colon G \to G/\Gamma$ is a normal covering map.*

Proof. As we observed in Example 3.35(e), G acts freely on itself, so the restriction of this action to Γ is certainly free. We will show that the action is proper by showing that it satisfies property (b) of Proposition 12.9.

Let $K \subset G$ be any compact set. If $\gamma \in \Gamma$ is an element such that $K \cap \gamma K \neq \varnothing$, then there exist $g_1, g_2 \in K$ such that $g_1 = \gamma g_2$, which is to say $\gamma \in KK^{-1} = \{g_1 g_2^{-1} : g_1, g_2 \in K\}$. This set KK^{-1} is compact because it is the image of $K \times K$ under the continuous map from $G \times G$ to G given by $(g_1, g_2) \mapsto g_1 g_2^{-1}$. Because Γ is discrete, there can be only finitely many elements of Γ in KK^{-1}. □

Corollary 12.15. *Suppose G and H are connected, locally path connected, locally compact Hausdorff topological groups, and $\varphi\colon G \to H$ is a surjective continuous homomorphism with discrete kernel. If φ is an open or closed map, then it is a normal covering map.*

Proof. Let $\Gamma = \operatorname{Ker}\varphi$. By the preceding proposition, the quotient map $\pi\colon G \to G/\Gamma$ is a normal covering map. The assumption that φ is either open or closed implies that it is a quotient map, and by the first isomorphism theorem the identifications made by φ are precisely those made by π. Thus the result follows from the uniqueness of quotient spaces. □

Example 12.16 (Coverings of the Torus). For any integers a, b, c, d such that $ad - bc \neq 0$, consider the map $p\colon \mathbb{T}^2 \to \mathbb{T}^2$ given by $p(z, w) = (z^a w^b, z^c w^d)$. This is easily seen to be a surjective continuous homomorphism, and it is a closed map by the closed map lemma. Once we show that it has discrete kernel, it will follow from the preceding corollary that it is a normal covering map.

Let A denote the invertible linear transformation of $\mathbb{R}^2$ whose matrix is $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$. Then we have a commutative diagram

$$\begin{tikzcd}
\mathbb{R}^2 \arrow[r, "A"] \arrow[d, "E"'] & \mathbb{R}^2 \arrow[d, "E"] \\
\mathbb{T}^2 \arrow[r, "p"'] & \mathbb{T}^2
\end{tikzcd} \tag{12.2}$$

where $E(x,y) = (e^{2\pi i x}, e^{2\pi i y})$ is the universal covering map of the torus. To identify $\operatorname{Ker} p$, note that

$$\begin{aligned} p \circ E(x,y) = (1,1) &\iff E \circ A(x,y) = (1,1) \\ &\iff A(x,y) \in \mathbb{Z}^2 \\ &\iff (x,y) \in A^{-1}(\mathbb{Z}^2), \end{aligned}$$

where $A^{-1}(\mathbb{Z}^2)$ denotes the additive subgroup $\{A^{-1}(m,n) : (m,n) \in \mathbb{Z}^2\}$ of $\mathbb{R}^2$. Because E is surjective, this shows that $\operatorname{Ker} p = E \circ A^{-1}(\mathbb{Z}^2)$.

Since A^{-1} has rational entries, it follows easily that each element of $\operatorname{Ker} p$ has finite order in $\mathbb{T}^2$. Moreover, since $\mathbb{Z}^2$ is generated (as a group) by the two elements $(1,0)$ and $(0,1)$, $\operatorname{Ker} p$ is generated by their images under $E \circ A^{-1}$. An abelian group that is generated by finitely many elements of finite order is easily seen to be finite; in particular, it is discrete.

Application: Universal Coverings of Higher Genus Surfaces

As another application of the theory of proper group actions, we will show that the unit disk $\mathbb{B}^2 \subset \mathbb{C}$ is the universal covering space of all the orientable surfaces of genus $n \geq 2$. The construction is rather involved, so we will describe the main steps and leave some of the details for you to work out. Some of these steps can be done a bit more straightforwardly if you know a little about Riemannian metrics and their geodesics, but we will not assume any such knowledge. We will, however, assume a passing acquaintance with complex analysis, at least enough to understand what it means for a function to be complex analytic.

We begin by describing a special metric on the disk. For $z_1, z_2 \in \mathbb{B}^2$, define

$$d(z_1, z_2) = \cosh^{-1}\left(1 + \frac{2|z_1 - z_2|^2}{(1 - |z_1|^2)(1 - |z_2|^2)}\right).$$

This is a metric, called the *hyperbolic metric*. (The only property of a metric that is not straightforward to check is the triangle inequality; a way to prove it is indicated in Problem 12-8.)

The disk with this metric, called the *hyperbolic disk*, is one model of non-Euclidean plane geometry. The "straight lines" in this geometry, called *hyperbolic geodesics*, are the intersections with the disk of Euclidean circles and lines meeting the unit circle orthogonally (Figure 12.6). (A line segment through the origin can be thought of as the limiting case of a circular arc as the radius goes to infinity.) It is easy to check that "two points determine a line": That is, given any two points in the disk, there is a unique hyperbolic geodesic passing through both points.

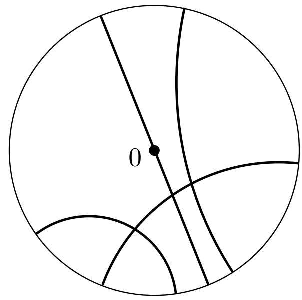


FIGURE 12.6. Hyperbolic geodesics.

The most interesting feature of the hyperbolic metric is that it is preserved by a transitive group action. Let α and β be complex numbers with $|\alpha|^2 - |\beta|^2 > 0$, and define

$$\varphi(z) = \frac{\alpha z + \beta}{\overline{\beta} z + \overline{\alpha}}. \tag{12.3}$$

A straightforward calculation shows that φ is continuous, takes the disk to itself, and preserves the hyperbolic metric in the sense that $d(\varphi(z_1), \varphi(z_2)) = d(z_1, z_2)$ for all $z_1, z_2 \in \mathbb{B}^2$. Any such map is called a *Möbius transformation* of the disk, and the set $\mathcal{M}$ of all such maps is a group under composition, called the *Möbius group* of the disk. It can be identified with the group of matrices of the form $\left(\begin{smallmatrix} \alpha & \beta \\ \overline{\beta} & \overline{\alpha} \end{smallmatrix}\right)$ and so is a topological group acting continuously on $\mathbb{B}^2$.

Möbius transformations take geodesics to geodesics, as can be seen by substituting $\varphi(z)$ for z in the equation defining a circle or line intersecting the boundary of the disk orthogonally, and noting that it reduces to another equation of one of the same types. In fact, the same computation shows that a Möbius transformation takes the intersection of the disk with *any* Euclidean circle or line to another set of one of the same forms.

One special case worth noting is that any rotation of the disk $z \mapsto e^{i\theta} z$ is a Möbius transformation with $\alpha = e^{i\theta/2}$ and $\beta = 0$, so the hyperbolic metric is invariant under rotations. In fact, any Möbius transformation that takes the origin to itself must be of this form, because (12.3) reduces to $\varphi(z) = (\alpha/\overline{\alpha})z$ in that case. Observe also that the hyperbolic distance from the origin to z depends only on $|z|$, so any metric ball $B_r(0)$ about the origin is actually a Euclidean disk centered at 0, and its boundary is a Euclidean circle. Since Möbius transformations preserve hyperbolic distance and take circles to circles, it follows that every metric ball is a Euclidean disk. (Its Euclidean center may not be the same as its hyperbolic center, however). It also follows that the hyperbolic metric generates the Euclidean topology.

The left action of $\mathcal{M}$ on the disk defined by (12.3) is transitive because any $z_0 \in \mathbb{B}^2$ is carried to 0 by the Möbius transformation

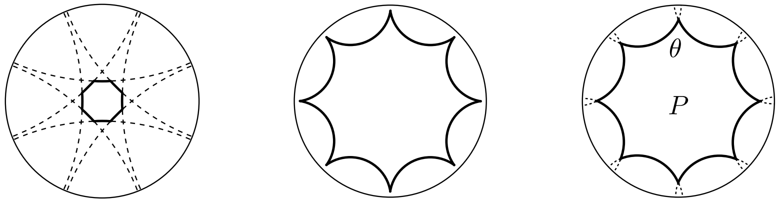
$$\varphi(z) = \frac{z - z_0}{1 - \bar{z}_0 z}. \tag{12.4}$$

In fact, more is true: Given any two *pairs* of points z_0, z_1 and z_0', z_1' such that $d(z_0, z_1) = d(z_0', z_1')$, there is a unique Möbius transformation taking z_0 to z_0' and z_1 to z_1' (and therefore taking the geodesic segment joining z_0, z_1 to the one joining z_0', z_1'). To prove this, let $\psi = \rho \circ \varphi$, where φ is the transformation (12.4) and ρ is a rotation moving $\varphi(z_1)$ to the positive x-axis, so that ψ takes z_0 to 0 and z_1 to some $\lambda > 0$. Similarly, there is a transformation ψ' taking z_0' to 0 and z_1' to $\lambda' > 0$. Since Möbius transformations preserve distances, λ and λ' are at the same distance from 0 along the positive x-axis and therefore must be equal, so $\psi'^{-1} \circ \psi$ is the transformation we seek. It is unique because if γ is any Möbius transformation taking z_0 to z_0' and z_1 to z_1', the composition $\psi' \circ \gamma \circ \psi^{-1}$ fixes 0 and therefore must be a rotation, and since it also fixes λ, it must be the identity, which implies $\gamma = \psi'^{-1} \circ \psi$.

Each Möbius transformation φ is complex analytic with nowhere vanishing derivative. Multiplication by the complex derivative $\varphi'(z_0)$ defines a linear map from $\mathbb{C}$ to $\mathbb{C}$, which can be interpreted geometrically as the action of φ on tangent vectors to curves: For any differentiable parametrized curve $f\colon (-\varepsilon, \varepsilon) \to \mathbb{B}^2$ with $f(0) = z_0$, the chain rule gives $(\varphi \circ f)'(0) = \varphi'(z_0) f'(0)$. Thus φ acts on tangent vectors by multiplying them by the nonzero complex number $\varphi'(z_0)$, and since all tangent vectors are rotated through the same angle, every Möbius transformation is *conformal*, meaning it preserves angles between tangent vectors. (We will also be considering angles between geodesics, by which we always mean angles between their tangent vectors.) In particular, if $\varphi(z) = e^{i\theta} z$ is rotation through an angle θ, then $\varphi'(0) = e^{i\theta}$ rotates tangent vectors through the same angle. It follows that the only Möbius transformation that fixes the origin and fixes the direction of a tangent vector at the origin is the identity. In fact, a Möbius transformation that fixes any point and a tangent direction at that point must be the identity, because conjugation with a transformation taking the fixed point to 0 yields a transformation that fixes 0 and a tangent direction at 0.

Now let M be a compact orientable surface of genus $n \geq 2$. We will show that there is a discrete subgroup $\Gamma \subset \mathcal{M}$ acting freely and properly on $\mathbb{B}^2$ such that M is homeomorphic to $\mathbb{B}^2/\Gamma$. It follows from Theorem 12.11 that the universal covering space of M is $\mathbb{B}^2$.

Recall from Chapter 6 the standard polygonal presentation of M as a quotient of a polygonal region with $4n$ sides whose edges are identified in pairs. We will realize M as a quotient of a compact region in $\mathbb{B}^2$ bounded by a *geodesic polygon*, that is, the union of finitely many geodesic segments. We begin by constructing a $4n$-sided geodesic polygon whose edges have



$\theta \approx \pi - \pi/2n.$ $\theta \approx 0.$ $\theta = \pi/2n.$

FIGURE 12.7. Geodesic polygons with interior angles $0 < \theta < \pi - \pi/2n$.

equal lengths and meet at equal angles (a *regular geodesic polygon*). Start with $4n$ points $(z_0, z_1, \ldots, z_{4n} = z_0)$ equally spaced on some circle about the origin. Because the hyperbolic metric is invariant under rotations, the geodesic segments joining z_j and z_{j+1} for $j = 0, \ldots, 4n - 1$ all have the same length and meet at equal angles, so their union is a regular geodesic polygon. As the radius of the circle goes to zero, these geodesics approach line segments through the origin, and define small regular geodesic polygons whose interior angles are very close to what they would be in the Euclidean case, namely $\pi - \pi/2n$ (see Figure 12.7). As the points get farther from the origin, the arcs become nearly tangent to each other, defining geodesic polygons with interior angles very near zero. By continuity, somewhere in between there is a polygon whose interior angles are exactly $\theta = \pi/2n$. (Note that this does not work when $n = 1$, so we cannot construct a covering of the torus in this manner.)

Let P be the compact subset of $\mathbb{B}^2$ consisting of this regular geodesic polygon together with the bounded component of its complement. Choose one vertex v_0, and label the edges $a_1, b_1, a_1^{-1}, b_1^{-1}, \ldots, a_n, b_n, a_n^{-1}, b_n^{-1}$ in counterclockwise order starting from v_0. (See Figure 12.8, but ignore the vertex labels other than v_0 for now.) For each edge pair a_j, a_j^{-1}, there is a unique Möbius transformation α_j that takes the edge labeled a_j^{-1} onto the one labeled a_j, with the initial vertex of one going to the initial vertex of the other. Similarly, let β_j be the transformation taking b_j to b_j^{-1} and respecting the initial and terminal vertices. Let $\Gamma \subset \mathcal{M}$ be the subgroup generated by $\{\alpha_j, \beta_j : j = 1, \ldots, 4n\}$. We will call the generators α_j, β_j, and their inverses *edge pairing transformations*.

One important property of the edge pairing transformations is easy to verify: If σ is any edge pairing transformation, then $P \cap \sigma(P)$ consists of exactly one edge of P. To see why, suppose σ takes an edge e to another edge e'. Then clearly, $P \cap \sigma(P)$ contains e'. Note that the complement of any geodesic in $\mathbb{B}^2$ has exactly two components, which we may call the *sides* of the geodesic. Because P is connected and lies on one side of each of its edges, the same is true of $\sigma(P)$. Using conformality and following what σ does to a vector that is perpendicular to e and points into P, it is easy to check that

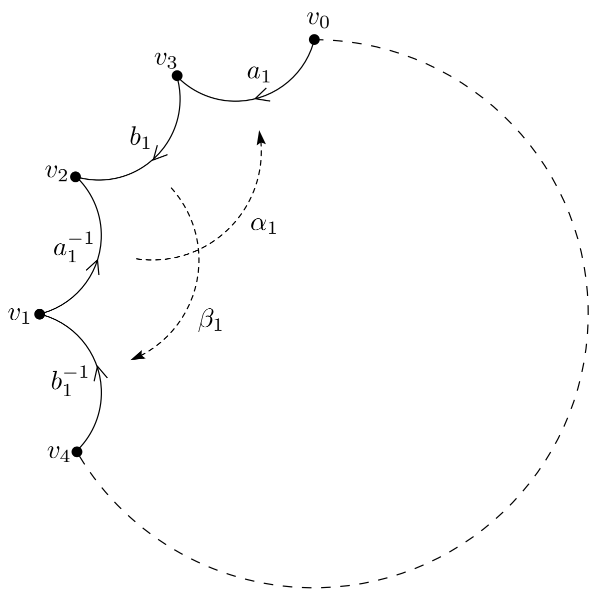


FIGURE 12.8. Edge pairing transformations.

$\sigma(P)$ lies on the opposite side of e' from P, and therefore $P \cap \sigma(P)$ consists of exactly the edge e'. Because P is obviously homeomorphic to a regular Euclidean polygon, the quotient of P by the identifications determined by the edge pairing transformations is homeomorphic to M. Let $p\colon P \to M$ denote the quotient map.

Theorem 12.17. *The group Γ is discrete and acts freely and properly on $\mathbb{B}^2$, and the quotient $\mathbb{B}^2/\Gamma$ is homeomorphic to M. The restriction of this quotient map to P is p.*

Proof. The first thing we will prove is that the edge pairing transformations satisfy the same relation as the generators of the fundamental group of M:

$$\alpha_1 \circ \beta_1 \circ \alpha_1^{-1} \circ \beta_1^{-1} \circ \cdots \circ \alpha_n \circ \beta_n \circ \alpha_n^{-1} \circ \beta_n^{-1} = \mathrm{Id}\,. \tag{12.5}$$

Actually, it will be more convenient to prove the equivalent identity obtained by inversion:

$$\beta_n \circ \alpha_n \circ \beta_n^{-1} \circ \alpha_n^{-1} \circ \cdots \circ \beta_1 \circ \alpha_1 \circ \beta_1^{-1} \circ \alpha_1^{-1} = \mathrm{Id}\,. \tag{12.6}$$

To simplify the notation, let us write the sequence of transformations on the left-hand side of (12.6) as $\sigma_{4n} \circ \cdots \circ \sigma_2 \circ \sigma_1$.

By definition, $\sigma_1 = \alpha_1^{-1}$ takes v_0, the initial vertex of the edge labeled a_1, to the initial vertex of the edge labeled a_1^{-1}. If we label the vertices in

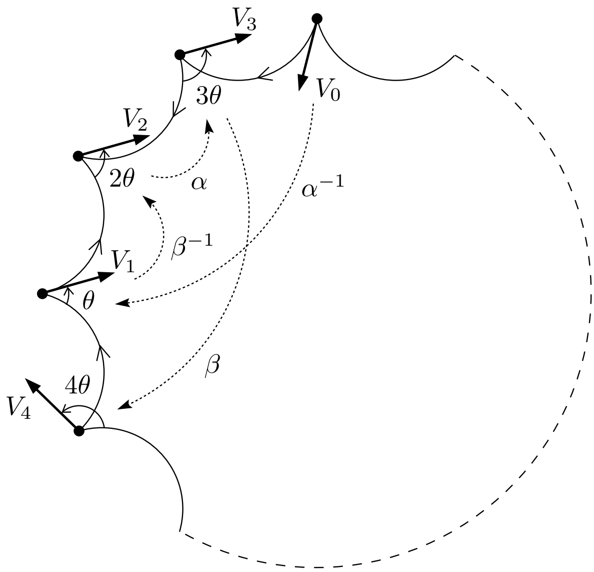


FIGURE 12.9. Images of a vector V_0 under edge pairing transformations.

counterclockwise order starting from v_0 as v_0, v_3, v_2, v_1, v_4 as in Figure 12.8, it is easy to check one step at a time that σ_j takes v_{j-1} to v_j for $j = 1, \dots, 4$. Since v_4 is also the initial vertex of the edge labeled a_2, we can continue by induction to number all the remaining vertices v_5 through $v_{4n} = v_0$ in such a way that $\sigma_j(v_{j-1}) = v_j$. In particular, $\sigma_{4n} \circ \cdots \circ \sigma_2 \circ \sigma_1(v_0) = v_0$. To show that this composition is the identity, it suffices to show that it fixes a tangent direction at v_0.

For any vertex v_j, we will measure angles of vectors at v_j from the edge adjacent to v_j in the counterclockwise direction (so we measure from a_1 at v_0, from b_1^{-1} at v_1, etc.). Positive angles will always be understood to mean counterclockwise rotation from that edge. Let $\theta = \pi/2n$ be the measure of the interior angles of P.

Let V_0 be a nonzero vector that makes an angle of 0 at v_0 (see Figure 12.9), and for $j = 1, \dots, 4n$ let V_j be the image of V_0 under $\sigma_j \circ \cdots \circ \sigma_1$, so that σ_j takes V_{j-1} to V_j. We will prove the following claim: *For each j, the angle of V_j at v_j is $j\theta$.* For $j = 0$ this is immediate from the definition of V_0. For $j = 1$, note that $\sigma_1 = \alpha_1^{-1}$ takes a_1 to a_1^{-1}, and therefore takes V_0 to a vector V_1 that points in the direction of a_1^{-1}, which makes an angle θ with b_1^{-1}. Next, since Möbius transformations preserve angles, the image V_2 of V_1 under $\sigma_2 = \beta_1^{-1}$ makes an angle θ with b_1, which is the same as an angle 2θ with a_1^{-1}. A similar analysis shows that the angles of V_3 and V_4 are 3θ and 4θ, respectively, and the claim is then proved for all j by induction. In particular, the angle of V_{4n} is $4n\theta = 2\pi$, so V_{4n} points in the same direction as V_0. This completes the proof of (12.5).

Now we have to prove that Γ is discrete and acts freely and properly on $\mathbb{B}^2$. It seems to be impossible to prove this by directly analyzing the action of Γ, so instead we resort to a rather circuitous trick due originally to Poincaré. We will construct "by hand" a covering space of M that ought to be its universal covering space, as a union of infinitely many copies of P—one for each element of $\pi_1(M)$—with "adjacent" copies glued together by the identifications determined by the edge pairing transformations. Only later will we show that this space is homeomorphic to $\mathbb{B}^2$, and therefore is simply connected and so is in fact the universal covering space.

Let G be the abstract group with presentation $\langle \alpha_1, \beta_1, \dots, \alpha_n, \beta_n \mid \alpha_1\beta_1\alpha_1^{-1}\beta_1^{-1}\cdots\alpha_n\beta_n\alpha_n^{-1}\beta_n^{-1}\rangle$, which is isomorphic to $\pi_1(M)$. Let $\sim$ be the equivalence relation on $G \times P$ generated by all relations of the form $(g, \sigma(z)) \sim (g\sigma, z)$, where σ is an edge pairing transformation and both z and $\sigma(z)$ are points in ∂P. Give G the discrete topology, and let $\widetilde{M}$ denote the quotient space $G \times P/\sim$. We will denote the equivalence class of (g, z) in $\widetilde{M}$ by $[g, z]$, and the quotient map by $\pi\colon G \times P \to \widetilde{M}$.

Left translation in the G factor defines a natural continuous action of G on $G \times P$. This respects the identifications made by π, so it descends to a continuous action of G on $\widetilde{M}$, satisfying $g' \cdot [g, z] = [g'g, z]$. This action is free, because $(g'g, z) \sim (g, z)$ only when $g' = 1$.

The subset $\widetilde{P} = \pi(\{1\} \times P) = \{[1, z] : z \in P\}$ of $\widetilde{M}$ is homeomorphic to P (why?), and $\widetilde{M}$ is the union of the sets $g \cdot \widetilde{P} = \{[g, z] : z \in P\}$ as g ranges over G. Each of these sets is a homeomorphic copy of P in $\widetilde{M}$, and the copies $g \cdot \widetilde{P}$ and $g' \cdot \widetilde{P}$ intersect in an edge precisely when g and g' differ by a single edge pairing transformation. Since there are only finitely many such transformations, this means in particular that each set $g \cdot \widetilde{P}$ intersects only finitely many others.

Because $\sim$ identifies only points (g, z) with $z \in \partial P$, the fiber of π over any point $[g_0, z_0]$ for $z_0 \in \operatorname{Int} P$ consists of exactly one point $(g_0, z_0) \in G \times P$. If z_0 is in ∂P but is not a vertex, then z_0 lies on one edge, and there is exactly one edge pairing transformation σ that identifies that edge with another edge; thus the fiber over $[g_0, z_0]$ is exactly two points (g_0, z_0) and $(g\sigma^{-1}, \sigma(z_0))$. If z_0 is a vertex of P, then by the argument at the beginning of the proof there is a sequence of edge pairing transformations $\sigma_1, \dots, \sigma_{4n}$ (possibly a cyclic permutation of the sequence we considered earlier) such that the points $z_j = \sigma_j \circ \cdots \circ \sigma_1(z_0)$ are the vertices of P, so the fiber over $[g_0, z_0]$ consists of the $4n$ points $(g_0\sigma_1^{-1}, z_1)$, $(g_0\sigma_1^{-1}\sigma_2^{-1}, z_2)$, $\dots$, $(g_0\sigma_1^{-1}\cdots\sigma_{4n}^{-1}, z_{4n}) = (g_0, z_0)$.

There is a natural continuous map $\widetilde{p}\colon \widetilde{M} \to M$ given by $\widetilde{p}[g,z] = p(z)$, obtained from $p \circ \pi_2$ by passing to the quotient:

$$\begin{tikzcd} G \times P \arrow[r, "\pi_2"] \arrow[d, "\pi"'] & P \arrow[d, "p"] \\ \widetilde{M} \arrow[r, "\widetilde{p}"'] & M. \end{tikzcd}$$

Clearly, $\widetilde{p}$ is surjective, because $\widetilde{p}(\widetilde{P}) = M$. It is a quotient map for the following reason: If $U \subset \widetilde{M}$ is an open set that is saturated with respect to $\widetilde{p}$, then $\pi^{-1}(U) \subset G \times P$ is open and saturated with respect to $\widetilde{p} \circ \pi = p \circ \pi_2$, and since $p \circ \pi_2$ is a quotient map, it follows that $\widetilde{p}(U) = \widetilde{p} \circ \pi(\pi^{-1}(U)) = p \circ \pi_2(\pi^{-1}(U))$ is open. You can check that the fibers of $\widetilde{p}$ are precisely the orbits of G in $\widetilde{M}$, so we can identify M with the orbit space $\widetilde{M}/G$. We wish to show that $\widetilde{p}$ is actually a covering map.

To show that $\widetilde{p}$ is a covering, by Theorem 12.11 it suffices to show that $\widetilde{M}$ is connected, locally path connected, locally compact, and Hausdorff, and that the action of G on $\widetilde{M}$ is proper. Connectedness is easy: If σ is an edge pairing transformation taking edge e to edge e', then the connected sets $\widetilde{P}$ and $\sigma \cdot \widetilde{P}$ have the points $[1, \sigma(z)] = [\sigma, z]$ in common for $z \in e$, so $\widetilde{P} \cup (\sigma \cdot \widetilde{P})$ is connected. By induction, any set of the form $\widetilde{P} \cup (\sigma_1 \cdot \widetilde{P}) \cup \cdots \cup (\sigma_m \cdots \sigma_1) \cdot \widetilde{P}$ is connected. Since $\widetilde{M}$ is the union of all such sets, and they all have points of $\widetilde{P}$ in common, $\widetilde{M}$ is connected.

To prove the other properties of $\widetilde{M}$, we first need to introduce some more maps. Let $\tau\colon G \to \Gamma$ be the homomorphism that sends each generator α_i or β_i to itself (thought of as an element of Γ), which is well-defined because (12.5) holds in Γ. The map $G \times P \to \mathbb{B}^2$ defined by $(g,z) \mapsto \tau(g)z$ is continuous and respects the identifications made by $\sim$, so it descends to a continuous map $\delta\colon \widetilde{M} \to \mathbb{B}^2$ given by $\delta[g,z] = \tau(g)z$. It takes the action of G on $\widetilde{M}$ over to the action of Γ on $\mathbb{B}^2$, in the sense that

$$\delta(g \cdot x) = \tau(g) \circ \delta(x). \tag{12.7}$$

The most important feature of $\widetilde{M}$ is that every $x \in \widetilde{M}$ has a neighborhood U with the following properties:

(i) $\overline{U}$ is mapped homeomorphically by δ onto a closed hyperbolic ball $\overline{B}_\varepsilon(\delta(x)) \subset \mathbb{B}^2$.

(ii) $\delta(U) = B_\varepsilon(\delta(x))$.

(iii) U intersects the sets $g \cdot \widetilde{P}$ for only finitely many $g \in G$.

We will call any such set U a *regular hyperbolic neighborhood* of x.

From the existence of regular hyperbolic neighborhoods it follows immediately that

- $\widetilde{M}$ is locally compact and locally path connected, because each regular hyperbolic neighborhood has these properties.

- $\widetilde{M}$ is Hausdorff: Let $x, x' \in \widetilde{M}$, and let U, U' be regular hyperbolic neighborhoods of them. If $x' \notin U$, then shrinking U a bit if necessary we may assume $x' \notin \overline{U}$, so that U and $U' \smallsetminus \overline{U}$ are disjoint neighborhoods of x and x'. On the other hand, if $x' \in U$, then the inverse images under $\delta|_U$ of disjoint neighborhoods of $\delta(x)$ and $\delta(x')$ are open sets separating x and x'.

- The action of G on $\widetilde{M}$ is proper: With x, x', U, and U' as above, there can be at most finitely many $g \in G$ such that $U \cap (g \cdot U') \neq \varnothing$, because U and U' intersect only finitely many of the sets $g \cdot \widetilde{P}$.

Thus, to complete the proof that $\widetilde{p}$ is a covering map, we need only prove the existence of a regular hyperbolic neighborhood of each point.

Let $x = [g_0, z_0]$ be an arbitrary point of $\widetilde{M}$. The fiber over x consists of finitely many points of the form (g_j, z_j), where $z_j = \sigma_j \circ \cdots \circ \sigma_1(z_0)$ for some (possibly empty) sequence of edge transformations $\sigma_1, \dots, \sigma_j$ and $g_j = g_0\sigma_1^{-1} \cdots \sigma_j^{-1}$. (The fiber contains one, two, or $4n$ such points depending on whether z_0 is an interior point, an edge point, or a vertex.) Choose $\varepsilon > 0$ smaller than half the distance from z_0 to any edge that does not contain z_0. Let $W \subset G \times P$ be the union of the sets $\{g_j\} \times (B_\varepsilon(z_j) \cap P)$, and let $U = \pi(W)$. Because W is a saturated open set, U is a neighborhood of x in $\widetilde{M}$. Similarly, $\overline{W}$ is the union of the sets $\{g_j\} \times (\overline{B}_\varepsilon(z_j) \cap P)$, a saturated closed set, so $\pi(\overline{W}) = \overline{U}$. Clearly, U intersects $g \cdot \widetilde{P}$ for only finitely many g.

To complete the proof that U is a regular hyperbolic neighborhood, we need to show that δ is a homeomorphism from $\overline{U}$ to $\overline{B}_\varepsilon(z_0)$ taking U to $B_\varepsilon(z_0)$. Since the diagram

$$\begin{tikzcd}
\overline{U} \arrow[r, "\delta"] \arrow[d, "g"'] & \overline{B}_\varepsilon(\delta(x)) \arrow[d, "\tau(g)"] \\
\overline{g \cdot U} \arrow[r, "\delta"'] & \overline{B}_\varepsilon(\delta(g \cdot x))
\end{tikzcd}$$

commutes for each $g \in G$ and the vertical maps are homeomorphisms, it suffices to prove this for $x = [1, z_0] \in \widetilde{P}$. We consider three cases.

Case I: $z_0 \in \operatorname{Int} P$. In this case, $\overline{U} \subset \widetilde{P}$, and it is immediate from the definitions that δ is one-to-one on $\overline{U}$, $\delta(\overline{U}) = \overline{B}_\varepsilon(z_0)$, and $\delta(U) = B_\varepsilon(z_0)$. Since $\overline{U}$ is the image under π of a compact set, it is compact, so $\delta\colon \overline{U} \to \overline{B}_\varepsilon(z_0)$ is a homeomorphism by the closed map lemma.

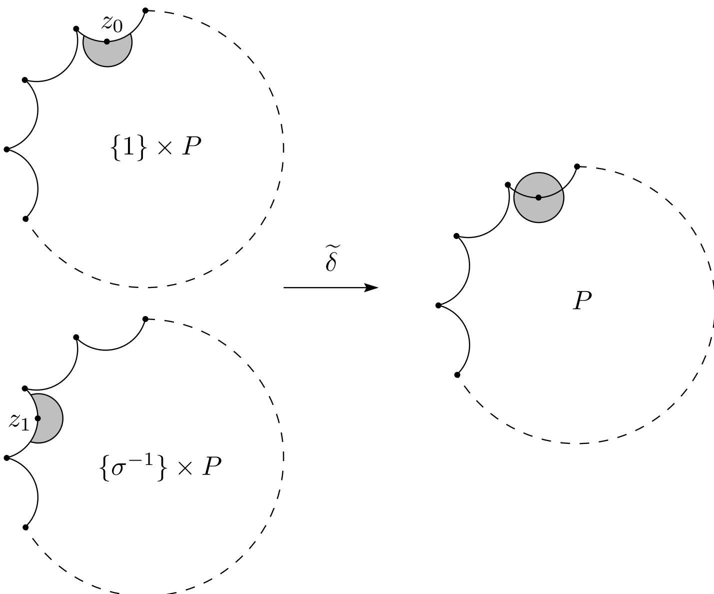


FIGURE 12.10. Hyperbolic neighborhood of an edge point.

Case II: *$z_0 \in \partial P$, but z_0 is not a vertex.* Let e_0 denote the edge containing z_0. By our choice of ε, $\overline{B}_\varepsilon(z_0) \cap P$ contains the entire portion of $\overline{B}_\varepsilon(z_0)$ lying on one side of e_0 (Figure 12.10). There is one edge pairing transformation σ that takes e_0 to another edge e_1, and thus takes z_0 to $z_1 = \sigma(z_0) \in e_1$. As a Möbius transformation of $\mathbb{B}^2$, σ takes $\overline{B}_\varepsilon(z_0)$ homeomorphically onto $\overline{B}_\varepsilon(z_1)$. Since $\overline{B}_\varepsilon(z_0) \cap P$ and $\sigma^{-1}(\overline{B}_\varepsilon(z_1) \cap P)$ lie on opposite sides of e_0, $\overline{B}_\varepsilon(z_0) = (\overline{B}_\varepsilon(z_0) \cap P) \cup \sigma^{-1}(\overline{B}_\varepsilon(z_1) \cap P)$. Then

$$\delta(\overline{U}) = \widetilde{\delta}(\overline{W}) = (\overline{B}_\varepsilon(z_0) \cap P) \cup \sigma^{-1}(\overline{B}_\varepsilon(z_1) \cap P) = \overline{B}_\varepsilon(z_0).$$

The restriction of δ to $\overline{U}$ is one-to-one, takes U onto $B_\varepsilon(z_0)$, and as before is a homeomorphism by the closed map lemma.

Case III: *z_0 is a vertex of P.* Then $\delta(\overline{U}) = \widetilde{\delta}(\overline{W})$ is the union of the sets

$$\widetilde{\delta}(\{\sigma_1^{-1} \cdots \sigma_j^{-1}\} \times (\overline{B}_\varepsilon(z_j) \cap P)) = \sigma_1^{-1} \circ \cdots \circ \sigma_j^{-1}(\overline{B}_\varepsilon(z_j) \cap P),$$

where $z_1, \dots, z_{4n}$ are the vertices of P. To see what these sets are, look back at the proof of (12.5); from that analysis, it follows that $\sigma_1^{-1} \circ \cdots \circ \sigma_j^{-1}$ maps z_j to z_0 and maps $\overline{B}_\varepsilon(z_j) \cap P$ to the sector of $\overline{B}_\varepsilon(z_0)$ lying between the geodesics passing through z_0 at angles $-j\theta$ and $(-j+1)\theta$ (Figure 12.11). These sectors fit together to make up the entire closed ball $\overline{B}_\varepsilon(z_0)$, and δ maps $\overline{U}$ bijectively to $\overline{B}_\varepsilon(z_0)$. As above, it is a homeomorphism by the closed map lemma.

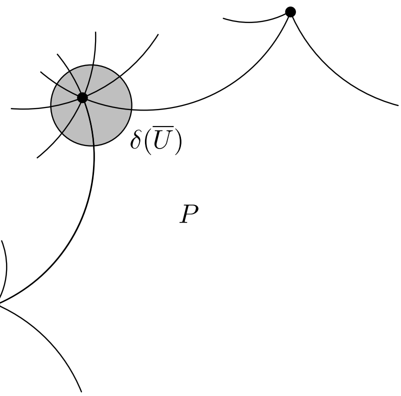


FIGURE 12.11. Hyperbolic neighborhood of a vertex point.

This completes the proof of the existence of hyperbolic neighborhoods and thus the proof that $\widetilde{p}\colon \widetilde{M} \to M$ is a covering map. To finish the proof of the theorem, we will show that $\delta\colon \widetilde{M} \to \mathbb{B}^2$ is also a covering map. Since $\mathbb{B}^2$ is simply connected, this implies that δ is a homeomorphism. The theorem follows from this, as we now show.

First, $\tau\colon G \to \Gamma$ is a group isomorphism: It is surjective because it takes generators of G to generators of Γ; and it is injective because if $\tau(g) = \mathrm{Id}$, then for any $x \in \widetilde{M}$ we have $\delta(g \cdot x) = \tau(g)\delta(x) = \delta(x)$, which implies $g \cdot x = x$ and therefore $g = 1$ because G acts freely. It follows that the action of Γ on $\mathbb{B}^2$ is equivalent to that of G on $\widetilde{M}$ under the homeomorphism δ, and the quotient map $\mathbb{B}^2 \to \mathbb{B}^2/\Gamma$ is equivalent to the covering map $\widetilde{p}\colon \widetilde{M} \to M$. Therefore, the action of Γ on $\mathbb{B}^2$ is free and proper, and the restriction of the covering map to P is $\widetilde{p} \circ \delta^{-1}|_P = p$. To see that Γ is a discrete subgroup of $\mathcal{M}$, suppose $\gamma_i \to \gamma$ in Γ. By continuity $\gamma_i z \to \gamma z$ for any $z \in \mathbb{B}^2$, and setting $g_i = \tau^{-1}(\gamma_i)$, $g = \tau^{-1}(\gamma)$, and $x = \delta^{-1}(z)$ we obtain $g_i \cdot x \to g \cdot x$. Since the g_i's are covering transformations, this can happen only if $g_i = g$ (and therefore $\gamma_i = \gamma$) for all sufficiently large i.

To show that δ is a covering, we need the following additional fact about regular hyperbolic neighborhoods: *There exists some $r > 0$ such that every point $x \in \widetilde{M}$ has a regular hyperbolic neighborhood U_x whose closure is mapped homeomorphically by δ onto $\overline{B}_r(\delta(x))$.* To prove this, let $K \subset \widetilde{M}$ denote the union of $\widetilde{P}$ together with its images $\sigma_i \cdot \widetilde{P}$ under the $4n$ edge pairing transformations σ_i. Since K is compact, so is its image $\delta(K) \subset \mathbb{B}^2$, and it is easy to see that $\delta(K)$ contains a neighborhood of P. As U ranges over regular hyperbolic neighborhoods of points in K, the sets $\delta(U)$ form an open cover of $\delta(K)$. Let c be a Lebesgue number for this cover, and choose $r < c$ small enough that for each $z \in P$ the hyperbolic ball $B_r(z)$ is contained in $\delta(K)$. This means that for every $z \in P$, there is a regular

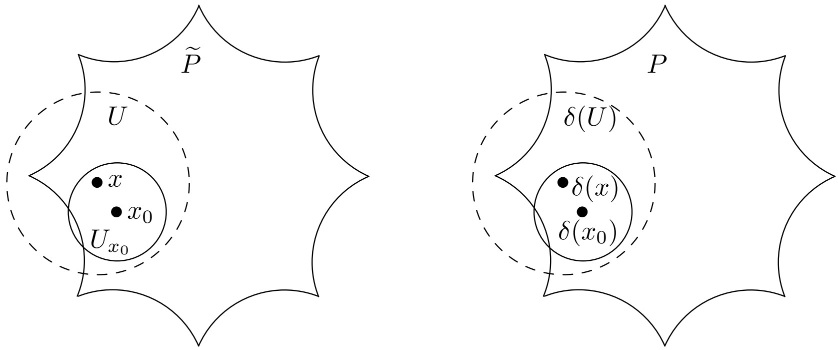


FIGURE 12.12. Finding regular hyperbolic balls of fixed radius.

hyperbolic neighborhood U of some point $x \in K$ such that $\overline{B}_r(z) \subset \delta(U)$. For each $x_0 \in \widetilde{P}$, choose a regular hyperbolic neighborhood U of some $x \in K$ such that $\overline{B}_r(\delta(x_0))$ is contained in $\delta(U)$ (Figure 12.12), and let $U_{x_0} = (\delta|_U)^{-1}(B_r(\delta(x_0))$; then $\delta\colon \overline{U}_{x_0} \to \overline{B}_r(\delta(x_0))$ is the restriction of a homeomorphism and hence is itself a homeomorphism. Since δ is injective on $\widetilde{P}$ and $\delta(x_0) \in \delta(U_{x_0})$, U_{x_0} is the desired neighborhood of x_0. For any other $x \in \widetilde{M}$, there is some $g \in G$ such that $g \cdot x \in \widetilde{P}$, so we can set $U_x = g^{-1} \cdot U_{g \cdot x}$.

We can now prove that δ is a covering map. First we need to show that it is surjective. If it were not, the image set $\delta(\widetilde{M})$ would have a boundary point $z_0 \in \mathbb{B}^2$. There is some point $z \in \delta(\widetilde{M})$ whose distance from z_0 is less than $r/2$. But then $z = \delta(x)$ for some $x \in \widetilde{M}$, and $\delta(U_x) = B_r(z)$, which is a neighborhood of z_0. This contradicts the assumption that z_0 is a boundary point of the image.

For any $z_0 \in \mathbb{B}^2$, we will show that $B_{r/2}(z_0)$ is evenly covered. Let V be a component of $\delta^{-1}(B_{r/2}(z_0))$ in $\widetilde{M}$. Since $\widetilde{M}$ is locally path connected, V is open. We need to show that $\delta\colon V \to B_{r/2}(z_0)$ is a homeomorphism. Choose $x \in V$, set $z = \delta(x)$, and let $\sigma = (\delta|_{U_x})^{-1}\colon B_r(z) \to U_x$.

Now, $\sigma(B_{r/2}(z_0))$ is a connected subset of $\delta^{-1}(B_{r/2}(z_0))$ that contains a point x in common with V, so it must be contained in V. This implies, for any $z' \in B_{r/2}(z_0)$, that $\delta(\sigma(z')) = z'$, so $\delta\colon V \to B_{r/2}(z_0)$ is surjective.

On the other hand, $\partial B_r(z)$ is disjoint from $B_{r/2}(z_0)$ by the triangle inequality. Since δ takes ∂U_x to $\partial B_r(z)$, it follows that $\partial U_x \cap V = \varnothing$. Now, $V \cap U_x$ is open in $\widetilde{M}$ and therefore open in V, and $V \cap U_x = V \cap \overline{U}_x$ is closed in V. Since V is connected, $V \cap U_x$ is all of V, which means that $V \subset U_x$. Thus $\delta|_V$ is the restriction of a homeomorphism, so it is injective and open, and therefore $\delta\colon V \to B_{r/2}(z_0)$ is a homeomorphism. $\square$

Corollary 12.18. *Let M be a compact surface. The universal covering space of M is homeomorphic to*

(*a*) $\mathbb{S}^2$ *if* $M \approx \mathbb{S}^2$ *or* $\mathbb{P}^2$;

(*b*) $\mathbb{R}^2$ *if* $M \approx \mathbb{T}^2$ *or* $\mathbb{P}^2 \# \mathbb{P}^2$;

(*c*) $\mathbb{B}^2$ *if M is any other surface.*

Proof. This was proved for all the orientable surfaces and $\mathbb{P}^2$ in this chapter. If M is a connected sum of $n \geq 2$ projective planes, then by Problem 11-8 M has a two-sheeted covering by the orientable surface N of genus $n-1$. If $\widetilde{M}$ is the universal covering space of M, then $\widetilde{M}$ also covers N by Corollary 12.6(a), so M and N have the same universal covering space. □

Note that $\mathbb{R}^2$ and $\mathbb{B}^2$ are homeomorphic, so up to topological equivalence there are only two simply connected 2-manifolds that cover compact surfaces. It is useful, however, to distinguish the two cases because of the different character of their covering transformations. For example, the covering transformations for the torus are all translations of the plane that preserve the Euclidean metric, while for the higher genus orientable surfaces they are Möbius transformations.

The Classification Theorem

In this section we assemble the results of this chapter to come up with a complete classification of coverings of a given space. The idea is that every covering of X is itself covered by the universal covering space, and intermediate coverings can be built from the universal covering as quotients by suitable group actions.

Theorem 12.19 (Classification of Coverings). *Let X be a connected, locally simply connected, locally compact Hausdorff space (for example, any connected manifold), and let $q \in X$ be any base point. There is a one-to-one correspondence between isomorphism classes of coverings of X and conjugacy classes of subgroups of $\pi_1(X,q)$. The correspondence associates each covering $p' \colon X' \to X$ with the conjugacy class of its induced subgroup.*

Proof. The covering isomorphism theorem shows that there is at most one isomorphism class of coverings corresponding to any conjugacy class of subgroups, so all we need to show is that there is at least one. Let $H \subset \pi_1(X,q)$ be any subgroup in the given conjugacy class. Let $p \colon \widetilde{X} \to X$ be the universal covering of X, and choose a base point $\widetilde{q} \in p^{-1}(q)$. Then the simply connected case of the covering group structure theorem (Corollary 11.32) shows that $\pi_1(X,q)$ is isomorphic to the covering group $\mathcal{C}_p(\widetilde{X})$, under the map $\alpha\colon \pi_1(X,q) \to \mathcal{C}_p(\widetilde{X})$ that sends $[f]$ to the unique covering transformation φ taking $\widetilde{q}$ to $\widetilde{q}\cdot[f]$. Let $\widetilde{H} = \alpha(H) \subset \mathcal{C}_p(\widetilde{X})$.

Since $\mathcal{C}_p(\widetilde{X})$ acts freely and properly on $\widetilde{X}$, it follows easily that $\widetilde{H}$ does too. So let X' denote the quotient space $\widetilde{X}/\widetilde{H}$ and $\pi\colon \widetilde{X} \to X'$ the quotient map; by Theorem 12.11, π is a normal covering map. Moreover, $p\colon \widetilde{X} \to X$ is constant on the fibers of π (since they are contained in the fibers of p), so p descends to a continuous map $p'\colon X' \to X$ such that the following diagram commutes:

$$\begin{tikzcd} \widetilde{X} \arrow[dd, "p"'] \arrow[dr, "\pi"] & \\ & X' \arrow[dl, "p'"] \\ X. & \end{tikzcd}$$

We have to show that p' is a covering map. Let $q_1 \in X$ be arbitrary, let U be a neighborhood of q_1 that is evenly covered by p, and let U' be any component of $p'^{-1}(U)$. To show that p' is a covering map, it suffices to show that U' is mapped homeomorphically onto U by p'.

Because X' is locally path connected, U' is open and closed in $p'^{-1}(U)$. Thus $\pi^{-1}(U')$ is open and closed in $\pi^{-1}(p'^{-1}(U)) = p^{-1}(U)$, which implies that it is a union of components of $p^{-1}(U)$. If $\widetilde{U}$ is any such component, the following diagram commutes:

$$\begin{tikzcd} \widetilde{U} \arrow[dd, "p"'] \arrow[dr, "\pi"] & \\ & U' \arrow[dl, "p'"] \\ U. & \end{tikzcd} \tag{12.8}$$

In this diagram, $p = p' \circ \pi$ is a homeomorphism, so π is injective on $\widetilde{U}$. If $\pi(\widetilde{U}) \neq U'$, then

$$\pi(\pi^{-1}(U')) = \bigcup_{\varphi\in\widetilde{H}} \pi(\varphi(\widetilde{U})) = \pi(\widetilde{U}) \neq U,$$

which contradicts the fact that $\pi\colon \widetilde{X} \to X'$ is surjective. Thus $\pi\colon \widetilde{U} \to U'$ is bijective, and because it is an open map, it is a homeomorphism. Since p and π are homeomorphisms in (12.8), so is p'.

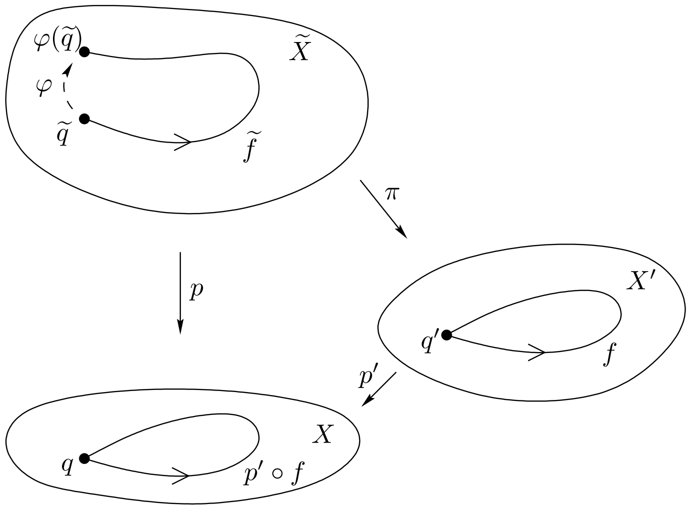


FIGURE 12.13. Proof of the classification theorem.

The last step is to show that $p'_*\pi_1(X', q') = H$ for some $q' \in X'$ such that $p'(q') = q$. Let $q' = \pi(\widetilde{q})$ and consider the following diagram:

$$\begin{tikzcd}
\pi_1(X', q') \arrow[r, "\alpha'"] \arrow[d, "p'_*"'] & \mathcal{C}_\pi(\widetilde{X}) \arrow[d, "\iota"] \\
\pi_1(X, q) \arrow[r, "\alpha"'] & \mathcal{C}_p(\widetilde{X}),
\end{tikzcd}$$

where α' and α represent the isomorphisms given by the covering group structure theorem, and ι is inclusion.

If this diagram commutes, we are done, because then

$$\begin{aligned}
p'_*\pi_1(X', q') &= \alpha^{-1} \circ \iota \circ \alpha'(\pi_1(X', q')) \\
&= \alpha^{-1} \circ \iota(\mathcal{C}_\pi(\widetilde{X})) \\
&= \alpha^{-1}(\widetilde{H}) = H.
\end{aligned}$$

To see that it commutes, let $[f] \in \pi_1(X', q')$ be arbitrary, and let $\varphi = \alpha'[f]$, so φ takes $\widetilde{q}$ to $\widetilde{q} \cdot [f] = \widetilde{f}(1)$, where $\widetilde{f}$ is the lift of f to a path in $\widetilde{X}$ starting at $\widetilde{q}$ (Figure 12.13). Then $\iota \circ \alpha'[f] = \varphi$, thought of as an element of $\mathcal{C}_p(\widetilde{X})$. On the other hand, $\alpha \circ p'_*[f] = \alpha[p' \circ f]$ is the transformation $\psi \in \mathcal{C}_p(\widetilde{X})$ taking $\widetilde{q}$ to $\widetilde{p' \circ f}(1)$. Now, $p \circ \widetilde{f} = p' \circ \pi \circ \widetilde{f} = p' \circ f$, so $\widetilde{f}$ is the lift of $p' \circ f$ starting at $\widetilde{q}$, which implies that $\widetilde{p' \circ f}(1) = \widetilde{f}(1)$. Thus $\varphi = \psi$ and the diagram commutes. $\square$

We end with a pair of interesting and representative examples.

Example 12.20 (Coverings of Lens Spaces). By the preceding theorem, the coverings of the lens space $L(n, m)$ are in one-to-one correspondence with subgroups of $\mathbb{Z}/\langle n\rangle$. (Since $\mathbb{Z}/\langle n\rangle$ is abelian, each conjugacy class contains precisely one subgroup.) Since every subgroup of a cyclic group is cyclic (Exercise A.27), the only possibilities for subgroups $G \subset \pi_1(L(n, m))$ are cyclic groups of order p where p is a factor of n. In each such case, a covering of $L(n, m)$ is obtained by restricting the action of $\mathbb{Z}/\langle n\rangle$ on $\mathbb{S}^3$ to G, and mapping the resulting quotient space down to $L(n, m)$ by sending each G-equivalence class to its $\mathbb{Z}/\langle n\rangle$-equivalence class. If $n = pq$ for positive integers p and q, let $G \subset \mathbb{Z}/\langle n\rangle$ be the cyclic subgroup of order p generated by (the coset of) q. It is easy to check from the definitions that $\mathbb{S}^3/G = L(p, m)$, and we obtain a q-sheeted covering $L(p, m) \to L(n, m)$. These are the only coverings of the lens spaces up to isomorphism.

Our last application will be to classify all the coverings of the torus up to isomorphism.

Proposition 12.21 (Classification of Torus Coverings). *Every covering of $\mathbb{T}^2$ is isomorphic to precisely one of the following:*

(*a*) *the universal covering $E : \mathbb{R}^2 \to \mathbb{T}^2$;*

(*b*) *the coverings $p : \mathbb{S}^1 \times \mathbb{R} \to \mathbb{T}^2$ by $p(z, y) = (z^a \varepsilon(y)^b, z^b \varepsilon(y)^{-a})$, where (a, b) are integers with $a \geq 0$ and $b > 0$ if $a = 0$;*

(*c*) *the coverings $p : \mathbb{T}^2 \to \mathbb{T}^2$ by $p(z, w) = (z^a w^b, w^c)$, where (a, b, c) are integers with $0 \leq b < a$ and $c > 0$.*

Proof. Note that all of these maps are coverings: the universal cover by Example 11.3; the maps in part (b) by Problem 12-7; and those in part (c) by Example 12.16.

Let us use $q = (1, 1) \in \mathbb{T}^2$ as base point, and represent $\pi_1(\mathbb{T}^2, q)$ as the product group $\langle\beta\rangle \times \langle\gamma\rangle$, where β and γ are the path classes of the standard generator of $\pi_1(\mathbb{S}^1, 1)$ in the first and second factors, respectively. Then the map $(m, n) \mapsto \beta^m\gamma^n$ is an isomorphism of $\mathbb{Z}^2$ with $\pi_1(\mathbb{T}^2, q)$.

The classification theorem says that isomorphism classes of coverings of $\mathbb{T}^2$ are in one-to-one correspondence with subgroups of $\pi_1(\mathbb{T}^2, q)$ under the correspondence that matches a covering $p : X \to \mathbb{T}^2$ with the subgroup induced by p. So we begin by showing that each subgroup of $\mathbb{Z}^2$ is one and only one of the following:

(i) the trivial subgroup;

(ii) infinite cyclic subgroups generated by (a, b) satisfying the conditions of (b) above;

(iii) subgroups of the form $\mathbb{Z}\langle(a,0),(b,c)\rangle$, where (a,b,c) satisfy the conditions of (c) above.

To prove this, let G be an arbitrary subgroup of $\mathbb{Z}^2$. Because $\mathbb{Z}^2$ is free abelian of rank 2, G is free abelian of rank at most 2 by Proposition 9.13. Thus there are three mutually exclusive cases, in which G has rank 0, 1, or 2. Clearly, the trivial subgroup has rank 0; we will show that the rank 1 and 2 cases correspond to (ii) and (iii), respectively.

If G has rank 1, it is cyclic. In this case there are two elements (a,b) and $(-a,-b)$ that generate G, and exactly one of these satisfies the conditions of (b). Thus (i) corresponds to the rank 1 case.

It remains to show that when G has rank 2 there are unique integers (a,b,c) satisfying the conditions in (c) such that $\{(a,0),(b,c)\}$ forms a basis for G. The subgroup $G_1 = G \cap (\mathbb{Z}\times\{0\})$ is not trivial: If $\{(m,n),(i,j)\}$ is any basis for G, then $j(m,n) - n(i,j)$ is an element of G in $\mathbb{Z}\times\{0\}$, which is not $(0,0)$ because of the independence of (m,n) and (i,j). Since $\mathbb{Z}\times\{0\}$ is cyclic, so is G_1. Let $(a,0)$ be a generator of G_1; replacing it by its negative if necessary, we may assume $a > 0$.

Since G has rank 2, it is not contained in G_1. As in the proof of Proposition 9.13, there is a basis for G of the form $\{(a,0),(b,c)\}$, where c is a generator of the image of G under the projection $\pi_2\colon \mathbb{Z}^2 \to \mathbb{Z}$. Replacing (b,c) by its negative if necessary, we may assume $c > 0$. Subtracting a multiple of $(a,0)$ from (b,c) (which still yields a basis), we may assume $0 \le b < a$. Thus we have found (a,b,c) satisfying the conditions in (c) such that $(a,0)$ and (b,c) are a basis for G.

Finally, we need to show that two such triples (a,b,c) and (a',b',c') that determine the same subgroup are identical. Since each basis can be expressed in terms the other, there is an integer matrix M such that

$$\begin{pmatrix} a & b \\ 0 & c \end{pmatrix} M = \begin{pmatrix} a' & b' \\ 0 & c' \end{pmatrix}.$$

Examining the lower left entry in this equation shows that M is also upper triangular. Since M has an inverse that also has integer entries, its determinant must be ± 1; and then the above equation shows that $\det M = 1$ (recall that a, c, a', and c' are all positive). Since M is upper triangular, its determinant is the product of its (integer) diagonal entries, so these must be both $+1$ or both -1; and then the fact that a and a' are both positive forces both diagonal entries to be 1, so $a = a'$ and $c = c'$. The upper right entry of the matrix equation then becomes $ak + b = b'$ (where k is the upper right entry of M). Since both b and b' satisfy $0 \le b < a$, this forces $k = 0$, so M is the identity.

To complete the proof, we need to check that the subgroups of $\pi_1(\mathbb{T}^2, q)$ induced by the covering maps (a), (b), (c) are exactly those corresponding to (i), (ii), (iii), respectively.

Case (a) is obvious, since the fundamental group of $\mathbb{R}^2$ is trivial.

For (b), note that the fundamental group of $\mathbb{S}^1 \times \mathbb{R}$ is infinite cyclic, generated by the path class of the loop $c(t) = (\alpha(t), 0)$. The image of this loop under p is $p \circ c(t) = (\alpha(t)^a, \alpha(t)^b)$, which represents the element $\beta^a \gamma^b \in \pi_1(\mathbb{T}^2, q)$. Under our isomorphism with $\mathbb{Z}^2$, this corresponds to (a, b) and generates the infinite cyclic group described in (ii).

For (c), it is easy to check that p carries the generators β and γ of $\pi_1(\mathbb{T}^2, q)$ to β^a and $\beta^b \gamma^c$. Under our isomorphism with $\mathbb{Z}^2$, the subgroup generated by these elements is exactly the one described in (iii). $\square$

Problems

12-1. Let E be the following subset of $\mathbb{R}^3 \times \mathbb{R}^3$:

$$E = \{(x, y) \in \mathbb{R}^3 \times \mathbb{R}^3 : x \neq y\}.$$

Define an equivalence relation in E by setting $(x, y) \sim (y, x)$ for all $(x, y) \in E$. Compute the fundamental group of $E/{\sim}$.

12-2. Let $M = \mathbb{T}^2 \# \mathbb{T}^2$.

(a) Show that the fundamental group of M has a subgroup of index 2.

(b) Prove that there exists a manifold $\widetilde{M}$ and a two-sheeted covering map $p \colon \widetilde{M} \to M$.

12-3. Consider the map $f \colon \mathbb{S}^1 \to \mathbb{T}^2$ given by

$$f(z) = (z^2, 1).$$

For which coverings $p \colon \widetilde{X} \to \mathbb{T}^2$ can f be lifted to $\widetilde{X}$?

12-4. Consider the action of $\mathbb{Z}$ on $\mathbb{R}^m \smallsetminus \{0\}$ defined by $n \cdot x = 2^n x$.

(a) Show that $\mathbb{Z}$ acts continuously, freely, and properly.

(b) Show that the orbit space $(\mathbb{R}^m \smallsetminus \{0\})/\mathbb{Z}$ is homeomorphic to $\mathbb{S}^{m-1} \times \mathbb{S}^1$.

(c) If $m \geq 3$, show that the universal covering space of $\mathbb{S}^{m-1} \times \mathbb{S}^1$ is homeomorphic to $\mathbb{R}^m \smallsetminus \{0\}$.

12-5. Identify a group Γ of homeomorphisms of the plane, generated by translations and reflections, such that $\mathbb{R}^2/\Gamma$ is homeomorphic to the Klein bottle.

12-6. Show that any continuous action of a finite group on a manifold is proper.

12-7. For any integers a, b, c, d such that $ad - bc \neq 0$, show that the map $p : \mathbb{S}^1 \times \mathbb{R} \to \mathbb{T}^2$ given by $p(z, y) = (z^a \varepsilon(y)^b, z^c \varepsilon(y)^d)$ is a covering map. [Hint: Using a commutative diagram similar to (12.2), show that p is an open map and a continuous homomorphism with discrete kernel.]

12-8. Prove the triangle inequality for the hyperbolic metric as follows. Show that it suffices to assume that one of the points is the origin, and use the identity $\cosh^2 x - \sinh^2 x = 1$ to show that $\sinh d(z,0) = 2|z|/(1-|z|^2)$, and therefore by the Euclidean triangle inequality,

$$\begin{aligned}\cosh d(z_1,z_2) &\le \cosh d(z_1,0)\cosh d(z_2,0) + \sinh d(z_1,0)\sinh d(z_2,0)\\ &= \cosh(d(z_1,0)+d(0,z_2)).\end{aligned}$$

12-9. Let G be a connected, locally simply connected, locally compact Hausdorff topological group, and let $\widetilde{G}$ be its universal covering space. Show that $\widetilde{G}$ has a unique group structure such that it is a topological group and such that the covering map $p\colon \widetilde{G} \to G$ is a homomorphism with discrete kernel.

13
Homology

In addition to the fundamental group and the higher homotopy groups, there are other groups that can be attached to a topological space in a way that is topologically invariant. To motivate them, let us look again at the fundamental group. Using the device of circle representatives as described in Chapter 7, we can think of the fundamental group of a space X as equivalence classes of maps from the circle into X modulo those that extend to the disk. Roughly, the idea of homology theory is to divide out by a somewhat larger equivalence relation, consisting of those maps that extend to a map into X from *any* surface whose boundary is the circle.

To see how this can lead to different results, let $X = \mathbb{T}^2 \# \mathbb{T}^2$ be the two-holed torus, and consider the loop f in X pictured in Figure 13.1. (It goes once around the boundary of the disk that is removed to form the connected sum.) In terms of our standard generators for $\pi_1(X)$, this loop is path homotopic to either $\alpha_1\beta_1\alpha_1^{-1}\beta_1^{-1}$ or $\beta_2\alpha_2\beta_2^{-1}\alpha_2^{-1}$, so it is not null

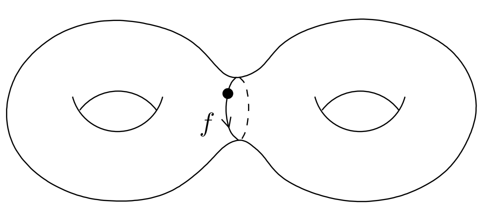


FIGURE 13.1. A loop that extends to a surface map.

homotopic, and its circle representative has no extension to a map from the closed disk into X. However, it is easy to see that the circle representative *does* extend to a map from $\mathbb{T}^2$ minus a disk into X—for example, the inclusion map of the left half of X is such an extension.

It turns out that a more satisfactory theory results if instead of considering loops modulo those that extend to maps from a 2-manifold with boundary, we consider instead formal sums of maps from a 1-simplex modulo those that are "boundaries" of sums of maps from a 2-simplex. Getting the definitions correct requires some care, and it is easy to lose sight of the geometric meaning among the technical details, but it will help if you keep the above example in mind throughout the discussion. The reward is a theory that extends easily to higher dimensions, is computationally tractable, and will allow us to prove a number of significant facts about manifolds that are much more difficult or even impossible to prove using homotopy groups alone.

We begin the chapter by defining a sequence of abelian groups attached to each topological space, called its singular homology groups, which formalize the intuitive discussion above. It follows immediately from the definition that these groups are topological invariants, and with a bit more work we show they are also homotopy invariants. Next we prove that there is a simple relationship between the first homology group $H_1(X)$ and the fundamental group, namely that $H_1(X)$ is naturally isomorphic to the abelianization of $\pi_1(X)$. Then we introduce one of the main tools for computing homology groups, the Mayer–Vietoris theorem, which is a homology analogue of the Seifert–Van Kampen theorem. Using these tools, we compute the homology groups of most of the spaces we have studied so far. We then describe some applications of homology: to the topological invariance of the dimension of a manifold, the existence of vector fields on spheres, and (using a different homology theory called simplicial homology) the topological invariance of the Euler characteristic of a polyhedron. In the final section we give a brief introduction to cohomology.

Singular Homology Groups

We begin with some definitions. For any integer $p \geq 0$, let $\Delta_p \subset \mathbb{R}^p$ denote the Euclidean simplex $\langle e_0, e_1, \ldots, e_p \rangle$, where $e_0 = 0$ and, for $1 \leq i \leq p$, $e_i = (0, \ldots, 1, \ldots, 0)$ is the vector with a 1 in the ith place and zeros elsewhere. We call Δ_p the *standard p-simplex*. If X is a topological space, a *singular p-simplex* in X is a continuous map $\sigma \colon \Delta_p \to X$. For example, a singular 0-simplex is just a map from the one-point space Δ_0 into X, which we may identify with a point in X; and a singular 1-simplex is a map from $\Delta_1 = [0,1] \subset \mathbb{R}$ into X, which is just a path in X. (A map is generally called "singular" if it fails to have some desirable property such as continuity or differentiability. In this case, the term singular is meant to reflect the fact that σ need not be an embedding, so its image may not look at all like a simplex.)

Let $C_p(X)$ denote the free abelian group generated by the set of all singular p-simplices in X. An element of $C_p(X)$, which can be written as a formal linear combination of singular simplices with integer coefficients, is called a *singular p-chain* in X, and the group $C_p(X)$ is called the *singular chain group* in dimension p.

There are some special singular simplices in Euclidean spaces that we will use frequently. Let $K \subset \mathbb{R}^n$ be a convex subset. For any $p+1$ points $v_0, \dots, v_p \in K$ (not necessarily in general position or even distinct), let $\alpha(v_0, \dots, v_p)\colon \Delta_p \to \mathbb{R}^n$ denote the restriction of the unique affine map that takes e_i to v_i for $i = 0, \dots, p$. By convexity, the image lies in K, so this is a singular p-simplex in K, called an *affine singular simplex.* A singular chain in which every singular simplex that appears is affine will be called an *affine chain.*

The point of homology theory is to use singular chains to detect "holes." The intuition is that any chain that closes up on itself (like a closed path) but is not equal to the "boundary value" of a chain of one higher dimension must surround a hole in X. To this end, we define a homomorphism from p-chains to $(p-1)$-chains that precisely captures the notion of boundary values.

For each $i = 0, \dots, p$, let $F_{i,p}\colon \Delta_{p-1} \to \Delta_p$ be the affine singular simplex

$$F_{i,p} = \alpha(e_0, \dots, \widehat{e_i}, \dots, e_p),$$

where the hat indicates that e_i is to be omitted. More specifically, $F_{i,p}$ is the affine map that sends

$$\begin{array}{ccc} e_0 & \mapsto & e_0 \\ \cdots & & \cdots \\ e_{i-1} & \mapsto & e_{i-1} \\ e_i & \mapsto & e_{i+1} \\ \cdots & & \cdots \\ e_{p-1} & \mapsto & e_p \end{array}$$

and therefore maps Δ_{p-1} homeomorphically onto the boundary face of Δ_p opposite the vertex e_i. We call $F_{i,p}$ the ith *face map* in dimension p.

For any singular simplex $\sigma\colon \Delta_p \to X$, define a $(p-1)$-chain $\partial\sigma$ called the *boundary* of σ by

$$\partial\sigma = \sum_{i=0}^{p} (-1)^i \sigma \circ F_{i,p}.$$

By the characteristic property of free abelian groups, this extends uniquely to a homomorphism $\partial\colon C_p(X) \to C_{p-1}(X)$, called the *boundary operator.*

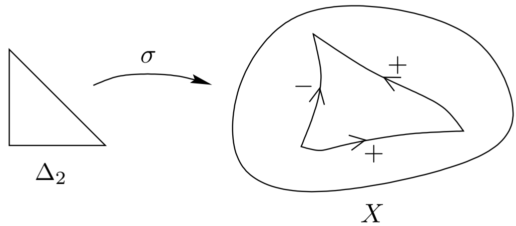


FIGURE 13.2. The boundary of a singular 2-simplex.

We sometimes indicate which chain group the boundary operator is acting on by a subscript, as in $\partial_p \colon C_p(X) \to C_{p-1}(X)$. The boundary of any 0-chain is defined to be zero.

A p-chain c is called a *cycle* if $\partial c = 0$, and it is called a *boundary* if there exists a $(p+1)$-chain b such that $c = \partial b$. The set $Z_p(X)$ of p-cycles is a subgroup of $C_p(X)$, because it is the kernel of the homomorphism ∂_p. Similarly, the set $B_p(X)$ of p-boundaries is also a subgroup (the image of ∂_{p+1}).

It might help clarify what is going on to work out some simple examples. When thinking about these examples, you should note the similarity between the formula for $\partial\sigma$ and the induced orientation on the boundary faces of a simplex, discussed in Chapter 5.

A singular 1-simplex is just a path $\sigma \colon I \to X$, and $\partial\sigma$ is the formal difference $\sigma(1) - \sigma(0)$. Therefore, a 1-cycle is a formal sum of paths with the property that the set of initial points counted with multiplicities is exactly the same as the set of terminal points with multiplicities. A typical example is a sum of paths $\sum_{i=1}^{k} \sigma_i$ such that $\sigma_i(1) = \sigma_{i+1}(0)$ and $\sigma_k(1) = \sigma_1(0)$. Apart from notation, this is pretty much the same thing as a product of paths (in the sense in which we used the term in Chapter 7) such that the last path ends where the first one starts (hence the term "cycle"). The only real difference is that chains do not keep track of the order in which the paths appear.

The boundary of a singular 2-simplex $\sigma \colon \Delta_2 \to X$ is a sum of three paths with signs (Figure 13.2). Think of this as a cycle in X that traverses the boundary values of σ in the counterclockwise direction. (Intuitively, you can think of a path with a negative sign as representing the same path going in the opposite direction; although they are not really the same, we will see below that they differ by a boundary, so they are equivalent from the point of view of homology.)

The most important feature of the singular boundary map is that "the boundary of a boundary is zero," as the next lemma shows.

Lemma 13.1. *For any singular chain c, $\partial(\partial c) = 0$.*

Proof. Since each chain group $C_p(X)$ is generated by singular simplices, it suffices to show this in the case in which $c = \sigma$ is a singular p-simplex.

First we note that the face maps satisfy the commutation relation

$$F_{i,p} \circ F_{j,p-1} = F_{j,p} \circ F_{i-1,p-1} \qquad \text{when } i > j, \tag{13.1}$$

as can be seen immediately by observing that the vertices of Δ_{p-2} are mapped according to the following chart:

$$\begin{array}{ccccc}
 & F_{j,p-1} & & F_{i,p} & \\
e_0 & \mapsto & e_0 & \mapsto & e_0 \\
\cdots & & \cdots & & \cdots \\
e_{j-1} & \mapsto & e_{j-1} & \mapsto & e_{j-1} \\
e_j & \mapsto & e_{j+1} & \mapsto & e_{j+1} \\
\cdots & & \cdots & & \cdots \\
e_{i-2} & \mapsto & e_{i-1} & \mapsto & e_{i-1} \\
e_{i-1} & \mapsto & e_i & \mapsto & e_{i+1} \\
\cdots & & \cdots & & \cdots \\
e_{p-2} & \mapsto & e_{p-1} & \mapsto & e_p.
\end{array}
\qquad
\begin{array}{ccccc}
 & F_{i-1,p-1} & & F_{j,p} & \\
e_0 & \mapsto & e_0 & \mapsto & e_0 \\
\cdots & & \cdots & & \cdots \\
e_{j-1} & \mapsto & e_{j-1} & \mapsto & e_{j-1} \\
e_j & \mapsto & e_j & \mapsto & e_{j+1} \\
\cdots & & \cdots & & \cdots \\
e_{i-2} & \mapsto & e_{i-2} & \mapsto & e_{i-1} \\
e_{i-1} & \mapsto & e_i & \mapsto & e_{i+1} \\
\cdots & & \cdots & & \cdots \\
e_{p-2} & \mapsto & e_{p-1} & \mapsto & e_p.
\end{array}$$

In other words, both compositions are equal to the affine simplex $\alpha(e_0, \dots, \widehat{e}_j, \dots, \widehat{e}_i, \dots, e_p)$. Using this, we compute

$$\begin{aligned}
\partial(\partial\sigma) &= \sum_{j=0}^{p-1}\sum_{i=0}^{p}(-1)^{i+j}\sigma \circ F_{i,p} \circ F_{j,p-1} \\
&= \sum_{0 \le j < i \le p} (-1)^{i+j}\sigma \circ F_{i,p} \circ F_{j,p-1} \\
&\qquad + \sum_{0 \le i \le j \le p-1} (-1)^{i+j}\sigma \circ F_{i,p} \circ F_{j,p-1}.
\end{aligned}$$

Making the substitutions $i = j'$, $j = i' - 1$ into the second sum and using (13.1), we see that the sums cancel term by term. □

Because of the preceding lemma, the group $B_p(X)$ of p-boundaries is a subgroup of the group $Z_p(X)$ of p-cycles. The pth *singular homology group* of X is defined to be the quotient group

$$H_p(X) = Z_p(X)/B_p(X) = \operatorname{Ker}\partial_p / \operatorname{Im}\partial_{p+1}.$$

It is zero if and only if every p-cycle is the boundary of some $(p+1)$-chain, which you should interpret intuitively as meaning that there are no p-dimensional "holes" in X. The equivalence class of a p-cycle c in $H_p(X)$ is denoted by $[c]$, and is called its *homology class*. If two p-cycles determine the same homology class (i.e., if they differ by a boundary), they are said to be *homologous*.

The significance of the homology groups derives from the fact that they are topological invariants. The proof is a very easy consequence of the fact that continuous maps induce homology homomorphisms. We begin by defining homomorphisms on the chain groups.

Given a continuous map $f\colon X \to Y$, let $f_\#\colon C_p(X) \to C_p(Y)$ be the homomorphism defined by setting $f_\#\sigma = f \circ \sigma$ for each singular p-simplex σ. The key fact is that $f_\#$ commutes with the boundary operators:

$$f_\#(\partial\sigma) = \sum_{i=0}^{p} (-1)^i f \circ \sigma \circ F_{i,p} = \partial(f_\#\sigma).$$

Because of this, $f_\#$ maps $Z_p(X)$ to $Z_p(Y)$ and $B_p(X)$ to $B_p(Y)$, and therefore passes to the quotient to define a homomorphism $f_*\colon H_p(X) \to H_p(Y)$, called the homomorphism *induced* by f.

Proposition 13.2 (Functorial Properties of Homology). *Let X, Y, and Z be topological spaces.*

(*a*) *The homomorphism $(\mathrm{Id}_X)_*\colon H_p(X) \to H_p(X)$ induced by the identity map of X is the identity of $H_p(X)$.*

(*b*) *If $f\colon X \to Y$ and $g\colon Y \to Z$ are continuous maps, then $(g \circ f)_* = g_* \circ f_*\colon H_p(X) \to H_p(Z)$.*

Thus the pth singular homology group defines a covariant functor from the category of topological spaces to the category of abelian groups.

Proof. It is trivial to check that both properties hold already for $f_\#$. □

Corollary 13.3 (Topological Invariance of Singular Homology). *If $f\colon X \to Y$ is a homeomorphism, then $f_*\colon H_p(X) \to H_p(Y)$ is an isomorphism.*

Exact Sequences and Chain Complexes

It is useful to look at the construction we just did in a somewhat more algebraic way. A sequence of abelian groups and homomorphisms

$$\cdots \to G_{p+1} \xrightarrow{\alpha_{p+1}} G_p \xrightarrow{\alpha_p} G_{p-1} \to \cdots$$

is said to be *exact* if $\operatorname{Im}\alpha_{p+1} = \operatorname{Ker}\alpha_p$ for all p. For example, a 5-term exact sequence of the form

$$0 \to A \xrightarrow{\alpha} B \xrightarrow{\beta} C \to 0$$

is called a *short exact sequence.* (The maps on the ends are the zero homomorphisms.) Because the image of the zero homomorphism is $\{0\}$, exactness at A means that α is injective, and similarly exactness at C means that β is surjective. Exactness at B means that $\operatorname{Ker}\beta = \alpha(A)$, and the first isomorphism theorem then tells us that $C \cong B/\alpha(A)$. A short exact sequence is thus a graphic summary of the first isomorphism theorem.

More generally, a sequence of abelian groups and homomorphisms

$$\cdots \to C_{p+1} \xrightarrow{\partial_{p+1}} C_p \xrightarrow{\partial_p} C_{p-1} \to \cdots$$

is called a *chain complex* if the composition of any two consecutive homomorphisms is the zero map: $\partial_p \circ \partial_{p+1} = 0$. This is equivalent to the requirement that $\operatorname{Im}\partial_{p+1} \subset \operatorname{Ker}\partial_p$. (The homomorphisms ∂_p are often called "boundary operators" by analogy with the case of singular homology.) We will denote such a chain complex by C_*, with the boundary maps being understood from the context. In many applications (such as the singular chain groups), C_p is defined only for $p \geq 0$, but it is sometimes convenient to extend this to all p by defining C_p to be the trivial group and the associated homomorphisms to be zero for $p < 0$.

The pth *homology group* of the chain complex C_* is

$$H_p(C_*) = \operatorname{Ker}\partial_p / \operatorname{Im}\partial_{p+1}.$$

Clearly, the chain complex is exact if and only if $H_p(C_*) = 0$ for all p; thus the homology groups provide a precise quantitative measurement of how the complex fails to be exact.

Now suppose C_* and D_* are chain complexes. A *chain map* $F\colon C_* \to D_*$ is a collection of homomorphisms $F\colon C_p \to D_p$ (we could distinguish them with subscripts, but there is no need) such that $\partial_p \circ F = F \circ \partial_p$ for all p:

$$\begin{tikzcd}
\cdots \arrow[r] & C_p \arrow[r, "\partial_p"] \arrow[d, "F"'] & C_{p-1} \arrow[r] \arrow[d, "F"] & \cdots \\
\cdots \arrow[r] & D_p \arrow[r, "\partial_p"'] & D_{p-1} \arrow[r] & \cdots .
\end{tikzcd}$$

For example, the homomorphisms $f_\#\colon C_p(X) \to C_p(Y)$ constructed above from a continuous map f define a chain map from the singular chain complex of X to that of Y. Any chain map takes $\operatorname{Ker}\partial$ to $\operatorname{Ker}\partial$ and $\operatorname{Im}\partial$ to $\operatorname{Im}\partial$, and therefore induces a homology homomorphism $F_*\colon H_p(C_*) \to H_p(D_*)$ for each p.

The study of exact sequences, chain complexes, and homology is part of the subject known as *homological algebra*. It began as a branch of topology, but has acquired a life of its own as a branch of algebra. We will return to these ideas briefly later in this chapter.

Elementary Computations

Although the definition of the singular homology groups may seem less intuitive than that of the fundamental group and the higher homotopy groups, the homology groups offer a number of advantages. For example, they are all abelian, which circumvents some of the thorny computational problems that beset the fundamental group. Also, there is no need to choose a base point, so unlike the homotopy groups, homology groups give us information about *all* the path components of a space, as the following lemma shows.

Lemma 13.4. *Let X be a space, let $\{X_\alpha\}_{\alpha\in A}$ be the set of path components of X, and let $\iota_\alpha\colon X_\alpha \hookrightarrow X$ be inclusion. Then for each $p \geq 0$ the maps $(\iota_\alpha)_*\colon H_p(X_\alpha) \to H_p(X)$ induce an isomorphism*

$$\bigoplus_{\alpha\in A} H_p(X_\alpha) \to H_p(X).$$

Proof. Since the image of any singular simplex must lie entirely in one path component, it is clear that the chain maps $(\iota_\alpha)_\#\colon C_p(X_\alpha) \to C_p(X)$ already induce isomorphisms

$$\bigoplus_{\alpha\in A} C_p(X_\alpha) \to C_p(X).$$

The result for homology follows easily from this. □

As in the case of the fundamental group, the definition of the homology groups does not give us much insight into how to compute them in general, because it involves taking quotients of huge groups by huge subgroups. There are, however, two simple cases that we can compute directly right now: the zero-dimensional homology groups of all spaces and all the homology groups of a one-point space. In the rest of this chapter we will develop some powerful tools for computing the rest of the homology groups.

Proposition 13.5 (Zero-Dimensional Homology). *For any topological space X, $H_0(X)$ is a free abelian group with basis consisting of an arbitrary point in each path component.*

Proof. It suffices to show that $H_0(X)$ is the infinite cyclic group generated by the class of any point when X is path connected, for then in the general case Lemma 13.4 guarantees that $H_0(X)$ is the direct sum of infinite cyclic groups, one for each path component.

A singular 0-chain is a formal linear combination of points in X with integer coefficients: $c = \sum_{i=1}^m n_i x_i$. Because the boundary operator is the zero map in dimension 0, every 0-chain is a cycle.

Assume that X is path connected, and define a map $\varepsilon\colon C_0(X) \to \mathbb{Z}$ by

$$\varepsilon\left(\sum_{i=1}^{m} n_i x_i\right) = \sum_{i=1}^{m} n_i.$$

It is clear that ε is a surjective homomorphism. We will show that $\operatorname{Ker}\varepsilon = B_0(X)$, from which it follows by the first isomorphism theorem that ε induces an isomorphism $H_0(X) \to \mathbb{Z}$. Since ε takes any single point to 1, the result follows.

If σ is a singular 1-simplex, then $\partial\sigma = \sigma(1) - \sigma(0)$, so $\varepsilon(\partial\sigma) = 1 - 1 = 0$. Therefore, $B_0(X) \subset \operatorname{Ker}\varepsilon$.

To show that $\operatorname{Ker}\varepsilon \subset B_0(X)$, choose any point $x_0 \in X$, and for each $x \in X$ let $\alpha(x)$ be a path from x_0 to x. This is a singular 1-simplex whose boundary is the 0-chain $x - x_0$. Thus, for an arbitrary 0-chain $c = \sum_i n_i x_i$ we compute

$$\partial\left(\sum_i n_i \alpha(x_i)\right) = \sum_i n_i x_i - \sum_i n_i x_0 = c - \varepsilon(c) x_0.$$

In particular, if $\varepsilon(c) = 0$, then $c \in B_0(X)$. □

Proposition 13.6 (Homology of a One-Point Space). *Let $*$ be a one-point space. The singular homology groups of $*$ are*

$$H_p(*) \cong \begin{cases} \mathbb{Z} & \textit{if } p = 0, \\ 0 & \textit{if } p > 0. \end{cases} \tag{13.2}$$

Proof. The case $p = 0$ follows from the preceding proposition, so we concentrate on $p > 0$. There is exactly one singular simplex in each dimension, namely the constant map $\sigma_p\colon \Delta_p \to *$, so each chain group $C_p(*)$ is the infinite cyclic group generated by σ_p. For $p > 0$, the boundary of σ_p is the alternating sum

$$\partial\sigma_p = \sum_{i=0}^{p} (-1)^i \sigma_p \circ F_{i,p} = \sum_{i=0}^{p} (-1)^i \sigma_{p-1} = \begin{cases} 0 & \text{if } p \text{ is odd}, \\ \sigma_{p-1} & \text{if } p \text{ is even}. \end{cases}$$

Thus $\partial\colon C_p(*) \to C_{p-1}(*)$ is an isomorphism when p is even and positive, and the zero map when p is odd:

$$\cdots \xrightarrow{\cong} C_3(*) \xrightarrow{0} C_2(*) \xrightarrow{\cong} C_1(*) \xrightarrow{0} C_0(*) \to 0.$$

It follows that for $p > 0$,

$$Z_p(*) = \begin{cases} C_p(*) & \text{if } p \text{ is odd}, \\ 0 & \text{if } p \text{ is even}; \end{cases}$$

$$B_p(*) = \begin{cases} C_p(*) & \text{if } p \text{ is odd}, \\ 0 & \text{if } p \text{ is even}. \end{cases}$$

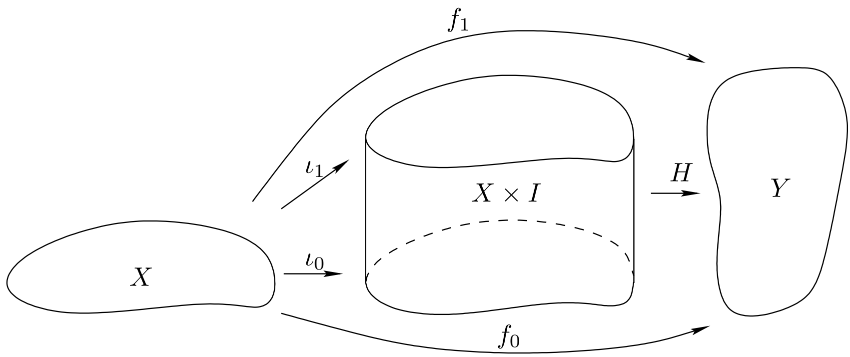


FIGURE 13.3. The setup for Theorem 13.7.

Taking quotients, we find that $H_p(*) = 0$. □

Homotopy Invariance

Just like the fundamental group, the singular homology groups are also homotopy invariant. The proof, as in the case of the fundamental group, depends on the fact that homotopic maps induce the same homology homomorphism.

Theorem 13.7. *If $f_0, f_1 \colon X \to Y$ are homotopic maps, then for each $p \geq 0$ the induced homomorphisms $(f_0)_*, (f_1)_* \colon H_p(X) \to H_p(Y)$ are equal.*

Before proving this theorem, we state its most important corollary.

Corollary 13.8 (Homotopy Invariance of Singular Homology). *If $f \colon X \to Y$ is a homotopy equivalence, then for each $p \geq 0$, $f_* \colon H_p(X) \to H_p(Y)$ is an isomorphism.*

Exercise 13.1. Prove Corollary 13.8.

Proof of Theorem 13.7. We begin by considering the special case in which $Y = X \times I$ and $f_i = \iota_i$, where $\iota_0, \iota_1 \colon X \to X \times I$ are the maps

$$\iota_0(x) = (x, 0), \qquad \iota_1(x) = (x, 1).$$

(See Figure 13.3.) Clearly, $\iota_0 \simeq \iota_1$. We will show below that $(\iota_0)_* = (\iota_1)_*$. As it turns out, this immediately implies the general case as follows. Suppose $f_0, f_1 \colon X \to Y$ are continuous maps and $H \colon X \times I \to Y$ is a homotopy from f_0 to f_1 (Figure 13.3). Then since $H \circ \iota_i = f_i$, we have

$$(f_0)_* = (H \circ \iota_0)_* = H_* \circ (\iota_0)_* = H_* \circ (\iota_1)_* = (H \circ \iota_1)_* = (f_1)_*.$$

To prove $(\iota_0)_* = (\iota_1)_*$, it would suffice to show that $(\iota_0)_\# c$ and $(\iota_1)_\# c$ differ by a boundary for each chain c. In fact, a little experimentation will probably convince you that this is usually false. But in fact all we need is that they differ by a boundary *when c is a cycle*. So we might try to define a map $h\colon Z_p(X) \to C_{p+1}(X \times I)$ such that

$$\partial h(c) = (\iota_1)_\# c - (\iota_0)_\# c. \tag{13.3}$$

It turns out to be hard to define such a thing for cycles only. Instead, we will define $h(c)$ for *all* p-chains c, and show that it satisfies a formula that implies (13.3) when c is a cycle.

For each $p \geq 0$, we will define a homomorphism $h\colon C_p(X) \to C_{p+1}(X \times I)$ that satisfies the following identity:

$$h \circ \partial + \partial \circ h = (\iota_1)_\# - (\iota_0)_\#. \tag{13.4}$$

From (13.4) it follows immediately that $(\iota_1)_\# c - (\iota_0)_\# c = \partial h(c)$ whenever $\partial c = 0$, and therefore $(\iota_1)_*[c] = (\iota_0)_*[c]$.

The construction of h is basically a "triangulated" version of the obvious homotopy from ι_0 to ι_1. Consider the convex set $\Delta_p \times I \subset \mathbb{R}^{p+1} = \mathbb{R}^p \times \mathbb{R}$. Note that $\Delta_p \times \{0\}$ and $\Delta_p \times \{1\}$ are Euclidean p-simplices in $\mathbb{R}^{p+1}$. Let us denote the vertices of $\Delta_p \times \{0\}$ by $E_i = (e_i, 0)$ and those of $\Delta_p \times \{1\}$ by $E'_i = (e_i, 1)$. For $0 \leq i \leq p$, let $G_{i,p}\colon \Delta_p \to \Delta_p \times I$ be the following affine singular $(p+1)$-simplex in $\mathbb{R}^{p+1}$:

$$G_{i,p} = \alpha(E_0, \dots, E_i, E'_i, \dots, E'_p).$$

Then define $h\colon C_p(X) \to C_{p+1}(X \times I)$ by

$$h(\sigma) = \sum_{i=0}^{p} (-1)^i (\sigma \times \mathrm{Id}) \circ G_{i,p}.$$

Note that $G_{i,p}$ takes its values in $\Delta \times I$ and $\sigma \times \mathrm{Id}$ is a map from $\Delta \times I$ to $X \times I$, so this does indeed define a $(p+1)$-chain in $X \times I$.

To get an idea of what this means geometrically, consider the case $p = 2$. The three simplices $\langle E_0, E'_0, E'_1, E'_2 \rangle$, $\langle E_0, E_1, E'_1, E'_2 \rangle$, and $\langle E_0, E_1, E_2, E'_2 \rangle$ give a triangulation of $\Delta_2 \times I$ (see Figure 13.4). In the special case in which σ is the identity map of Δ_2, $h(\sigma)$ is a sum of affine singular simplices mapping Δ_3 homeomorphically onto each one of these 3-simplices, with signs chosen to correspond to the natural orientation on each simplex. In the general case, $h(\sigma)$ is this singular chain followed by the map $\sigma \times \mathrm{Id}$, and thus is a chain in $X \times I$ whose image is the product set $\sigma(\Delta_2) \times I$.

Now we need to prove that h satisfies (13.4). For this purpose, we will need some relations between the affine simplices $G_{i,p}$ and the face maps $F_{j,p}$. First, if $1 \leq j \leq p$, note that $G_{j,p}$ and $G_{j-1,p}$ agree on all the vertices

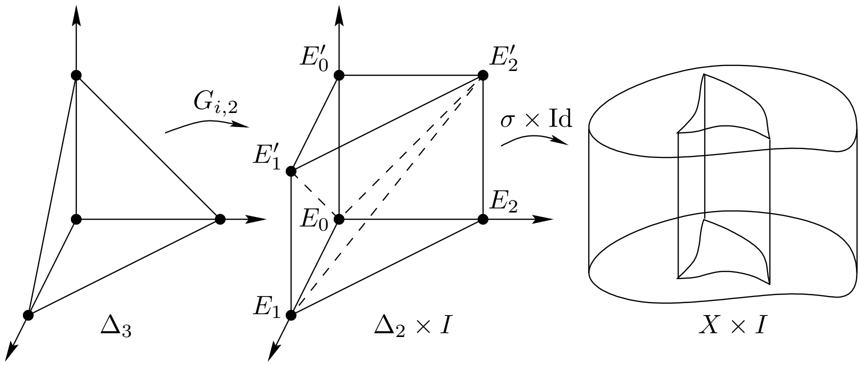


FIGURE 13.4. The operator h in dimension 2.

of Δ_p except e_j. Because $F_{j,p+1}$ skips e_j, the compositions $G_{j,p} \circ F_{j,p+1}$ and $G_{j-1,p} \circ F_{j,p+1}$ are equal. In fact, it is straightforward to check that

$$G_{j,p} \circ F_{j,p+1} = G_{j-1,p} \circ F_{j,p+1} = \alpha(E_0, \dots, E_{j-1}, E_j', \dots, E_p'). \tag{13.5}$$

Similarly, by following what each map does to basis elements as we did in the proof of Lemma 13.1, one can compute that

$$(F_{j,p} \times \mathrm{Id}) \circ G_{i,p-1} = \begin{cases} G_{i+1,p} \circ F_{j,p+1} & \text{if } i \geq j, \\ G_{i,p} \circ F_{j+1,p+1} & \text{if } i < j. \end{cases} \tag{13.6}$$

Let σ be an arbitrary singular p-simplex in X. Using (13.6), we compute

$$\begin{aligned} h(\partial\sigma) &= h \sum_{j=0}^{p} (-1)^j \sigma \circ F_{j,p} \\ &= \sum_{i=0}^{p-1} \sum_{j=0}^{p} (-1)^{i+j} \big((\sigma \circ F_{j,p}) \times \mathrm{Id}\big) \circ G_{i,p-1} \\ &= \sum_{i=0}^{p-1} \sum_{j=0}^{p} (-1)^{i+j} (\sigma \times \mathrm{Id}) \circ (F_{j,p} \times \mathrm{Id}) \circ G_{i,p-1} \\ &= \sum_{0 \leq j \leq i \leq p-1} (-1)^{i+j} (\sigma \times \mathrm{Id}) \circ G_{i+1,p} \circ F_{j,p+1} \\ &\quad + \sum_{0 \leq i < j \leq p} (-1)^{i+j} (\sigma \times \mathrm{Id}) \circ G_{i,p} \circ F_{j+1,p+1}. \end{aligned} \tag{13.7}$$

On the other hand,

$$\begin{aligned}\partial h(\sigma) &= \partial \sum_{i=0}^{p} (-1)^i (\sigma \times \mathrm{Id}) \circ G_{i,p} \\ &= \sum_{j=0}^{p+1} \sum_{i=0}^{p} (-1)^{i+j} (\sigma \times \mathrm{Id}) \circ G_{i,p} \circ F_{j,p+1}.\end{aligned}$$

Separating the terms where $i < j - 1$, $i = j - 1$, $i = j$, and $i > j$, this becomes

$$\begin{aligned}\partial h(\sigma) = &\sum_{0 \le i < j-1 < j \le p+1} (-1)^{i+j} (\sigma \times \mathrm{Id}) \circ G_{i,p} \circ F_{j,p+1} \\ &- \sum_{1 \le j \le p+1} (\sigma \times \mathrm{Id}) \circ G_{j-1,p} \circ F_{j,p+1} \\ &+ \sum_{0 \le j \le p} (\sigma \times \mathrm{Id}) \circ G_{j,p} \circ F_{j,p+1} \\ &+ \sum_{0 \le j < i \le p} (-1)^{i+j} (\sigma \times \mathrm{Id}) \circ G_{i,p} \circ F_{j,p+1}.\end{aligned}$$

Making the index substitutions $j = j' + 1$ in the first sum and $i = i' + 1$ in the last, we see that these two sums exactly cancel those in (13.7). By virtue of (13.5), all the terms in the middle two sums cancel except those where $j = 0$ and $j = p + 1$. These two terms yield

$$\begin{aligned}h(\partial\sigma) + \partial h(\sigma) &= -(\sigma \times \mathrm{Id}) \circ \alpha(E_0, \dots, E_p) + (\sigma \times \mathrm{Id}) \circ \alpha(E'_0, \dots, E'_p) \\ &= -(\iota_0)_\# \sigma + (\iota_1)_\# \sigma.\end{aligned}$$

This completes the proof. □

As an immediate application, we can conclude that contractible spaces have trivial homology in all dimensions greater than zero. (It is infinite cyclic in dimension zero by Proposition 13.5.)

Corollary 13.9. *Suppose X is a contractible space. Then $H_p(X) = 0$ for all $p > 0$.*

There is an abstract algebraic version of what we just did. Suppose $F, G\colon C_* \to D_*$ are chain maps. A collection of homomorphisms $h\colon C_p \to D_{p+1}$ is called a *chain homotopy* from F to G if the following identity is satisfied on each group C_p:

$$h \circ \partial + \partial \circ h = G - F.$$

If there exists such a map, F and G are said to be *chain homotopic.*

Exercise 13.2. If $F, G\colon C_* \to D_*$ are chain homotopic chain maps, show that $F_* = G_*\colon H_p(C_*) \to H_p(D_*)$ for all p.

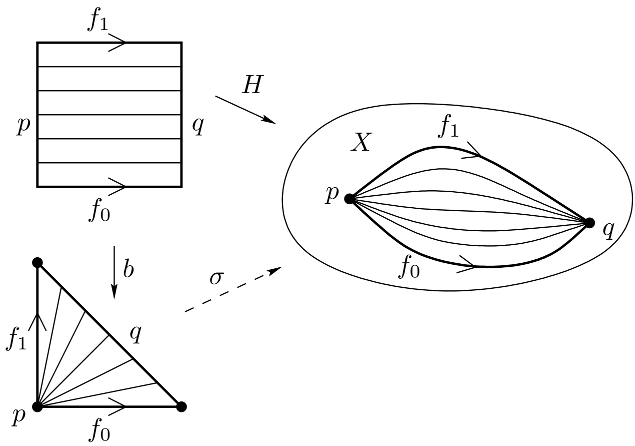


FIGURE 13.5. Path homotopic paths differ by a boundary.

Homology and the Fundamental Group

In this section we show that there is a simple relationship between the first homology group of a path connected space and its fundamental group: The former is just the abelianization of the latter. This will enable us to compute the first homology groups of all the spaces whose fundamental groups we know.

We begin by defining a map from the fundamental group to the first homology group. Let X be a space and q any point in X. A loop f based at q is also a singular 1-simplex. In fact, it is a cycle, since $\partial f = f(1) - f(0) = 0$. Therefore, any loop determines a 1-homology class. The following lemma shows that the resulting class depends only on the path homotopy class of f.

Lemma 13.10. *Suppose f_0 and f_1 are paths in X, and $f_0 \sim f_1$. Then, considered as a singular chain, $f_0 - f_1$ is a boundary.*

Proof. We must show there is a singular 2-chain whose boundary is the 1-chain $f_0 - f_1$. Let $H: f_0 \sim f_1$, and let $b: I \times I \to \Delta_2$ be the map

$$b(x, y) = (x - xy, xy), \tag{13.8}$$

which maps the square onto the triangle by sending each horizontal line segment linearly to a radial line segment (Figure 13.5). Then b is a quotient map by the closed map lemma, and identifies the left-hand edge of the square to the origin. Since H respects the identifications made by b, it passes to the quotient to yield a continuous map $\sigma: \Delta_2 \to X$, i.e., a singular 2-simplex. From the definition of the boundary operator, $\partial\sigma = c_q - f_1 + f_0$,

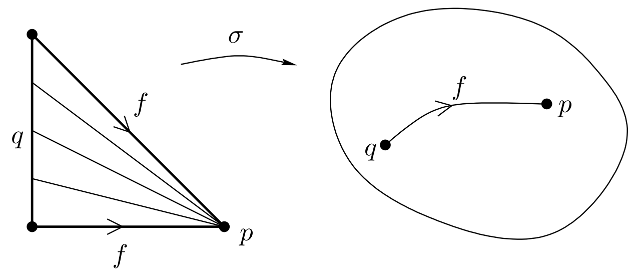


FIGURE 13.6. Proof that $[f^{-1}]_H = -[f]_H$.

where $q = f_0(1)$. Since c_q is the boundary of the constant 2-simplex that maps Δ_2 to q, it follows that $f_0 - f_1$ is a boundary. □

In this section, because we will be dealing with various equivalence relations on paths, we adopt the following notation. For any path in X (not necessarily a loop), we let $[f]_\pi$ denote its equivalence class modulo path homotopy. In particular, if f is a loop based at q, then $[f]_\pi$ is its path class in $\pi_1(X,q)$. Similarly, if c is any 1-chain we let $[c]_H$ denote its equivalence class modulo $B_1(X)$, so if c is a cycle (a loop for example), then $[c]_H$ is an element of $H_1(X)$. Define a map $\gamma\colon \pi_1(X,q) \to H_1(X)$, which we will call the *Poincaré homomorphism*, by

$$\gamma([f]_\pi) = [f]_H.$$

By Lemma 13.10, γ is well-defined.

Theorem 13.11. *Let X be a path connected space and $q \in X$. Then $\gamma\colon \pi_1(X,q) \to H_1(X)$ is a surjective homomorphism whose kernel is the commutator subgroup of $\pi_1(X,q)$. Consequently, $H_1(X)$ is isomorphic to the abelianization of $\pi_1(X,q)$.*

Proof. We begin by showing that $[f^{-1}]_H = -[f]_H$ for any path f in X. To see this, define a singular 2-simplex $\sigma\colon \Delta_2 \to X$ by $\sigma(x,y) = f(x)$ (Figure 13.6). Then $\partial\sigma = f^{-1} - c_q + f$, where $q = f(0)$. Since c_q is a boundary, it follows that the 1-chains f^{-1} and $-f$ differ by a boundary.

Next we show that γ is a homomorphism. Somewhat more generally, we will show that $[f \cdot g]_H = [f]_H + [g]_H$ for any two paths f, g such that $f(1) = g(0)$. When applied to loops f and g based at q, this implies that γ is a homomorphism.

Given such paths f and g, define a singular 2-simplex $\sigma\colon \Delta_2 \to X$ by

$$\sigma(x,y) = \begin{cases} f(y - x + 1) & \text{if } y \le x, \\ g(y - x) & \text{if } y \ge x. \end{cases}$$

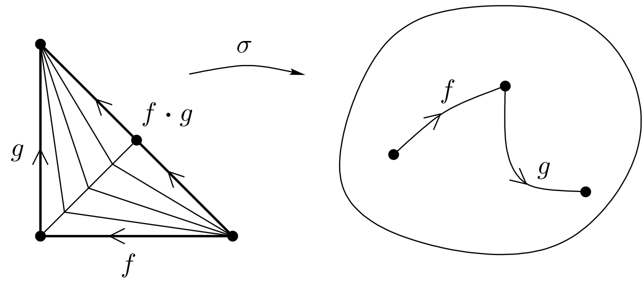


FIGURE 13.7. Proof that γ is a homomorphism.

(See Figure 13.7.) This is constant on each line segment $y - x =$ constant, and is continuous by the gluing lemma. It is easy to check that its boundary is the 1-chain $(f \cdot g) - g + f^{-1}$, from which it follows that

$$[f \cdot g]_H = [g]_H - [f^{-1}]_H = [g]_H + [f]_H.$$

Thus γ is a homomorphism.

Next we need to show that γ is surjective. For each point $x \in X$, let $\alpha(x)$ be a specific path from q to x, with $\alpha(q)$ chosen to be the constant path c_q. Since each path $\alpha(x)$ is in particular a 1-chain, the map $x \mapsto \alpha(x)$ extends uniquely to a group homomorphism $\alpha\colon C_0(X) \to C_1(X)$. For any path σ in X, define a loop $\widetilde{\sigma}$ based at q by

$$\widetilde{\sigma} = \alpha(\sigma(0)) \cdot \sigma \cdot \alpha(\sigma(1))^{-1}.$$

Observe that

$$\begin{aligned} \gamma([\widetilde{\sigma}]_\pi) &= [\alpha(\sigma(0)) \cdot \sigma \cdot \alpha(\sigma(1))^{-1}]_H \\ &= [\alpha(\sigma(0))]_H + [\sigma]_H - [\alpha(\sigma(1))]_H \\ &= [\sigma]_H - [\alpha(\partial\sigma)]_H. \end{aligned} \tag{13.9}$$

Now suppose $c = \sum_{i=1}^m n_i\sigma_i$ is an arbitrary 1-chain. Let f be the loop

$$f = (\widetilde{\sigma}_1)^{n_1} \cdot \cdots \cdot (\widetilde{\sigma}_m)^{n_m}.$$

From (13.9) and the fact that γ is a homomorphism it follows that

$$\begin{aligned} \gamma([f]_\pi) &= \sum_{i=1}^m n_i\big([\sigma_i]_H - [\alpha(\partial\sigma_i)]_H\big) \\ &= [c]_H - [\alpha(\partial c)]_H. \end{aligned}$$

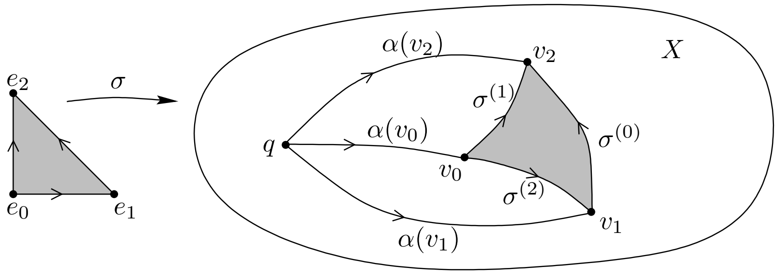


FIGURE 13.8. Proof that $\operatorname{Ker}\gamma$ is the commutator subgroup.

In particular, if c is a cycle, then $\gamma([f]_\pi) = [c]_H$, which shows that γ is surjective.

Because $H_1(X)$ is an abelian group, $\operatorname{Ker}\gamma$ clearly contains the commutator subgroup $[\pi_1(X,q), \pi_1(X,q)]$. All that remains is to show that the commutator subgroup is the entire kernel.

Let Π denote the abelianized fundamental group of X, and for any loop f based at q let $[f]_\Pi$ denote the equivalence class of $[f]_\pi$ in Π. Because the product in Π is induced by path multiplication, we will indicate it with a dot and write it multiplicatively even though Π is abelian. For any singular 1-simplex σ, let $\beta(\sigma) = [\widetilde{\sigma}]_\Pi \in \Pi$. Because Π is abelian, this extends uniquely to a homomorphism $\beta\colon C_1(X) \to \Pi$. We will show that β takes all 1-boundaries to the identity element of Π.

Let σ be an arbitrary singular 2-simplex. Write $v_i = \sigma(e_i)$ and $\sigma^{(i)} = \sigma \circ F_{i,2}$, so that $\partial\sigma = \sigma^{(0)} - \sigma^{(1)} + \sigma^{(2)}$ (see Figure 13.8). Note that the loop $\sigma^{(0)} \cdot (\sigma^{(1)})^{-1} \cdot \sigma^{(2)}$ is path homotopic to the constant loop c_{v_1}. (This can be seen either by identifying Δ_2 with the closed disk via a homeomorphism and noting that σ provides an extension of the circle representative of $\sigma^{(0)} \cdot (\sigma^{(1)})^{-1} \cdot \sigma^{(2)}$ to the disk; or by applying Lemma 7.12 to the composition $\sigma \circ b$, where $b\colon I \times I \to \Delta_2$ is given by (13.8).) We compute

$$\begin{aligned}
\beta(\partial\sigma) &= [\widetilde{\sigma}^{(0)}]_\Pi \cdot \left([\widetilde{\sigma}^{(1)}]_\Pi\right)^{-1} \cdot [\widetilde{\sigma}^{(2)}]_\Pi \\
&= [\widetilde{\sigma}^{(0)} \cdot (\widetilde{\sigma}^{(1)})^{-1} \cdot \widetilde{\sigma}^{(2)}]_\Pi \\
&= [\alpha(v_1) \cdot \sigma^{(0)} \cdot \alpha(v_2)^{-1} \cdot \alpha(v_2) \cdot (\sigma^{(1)})^{-1} \cdot \alpha(v_0)^{-1} \\
&\qquad \cdot \alpha(v_0) \cdot \sigma^{(2)} \cdot \alpha(v_1)^{-1}]_\Pi \\
&= [\alpha(v_1) \cdot \sigma^{(0)} \cdot (\sigma^{(1)})^{-1} \cdot \sigma^{(2)} \cdot \alpha(v_1)^{-1}]_\Pi \\
&= [\alpha(v_1) \cdot c_{v_1} \cdot \alpha(v_1)^{-1}]_\Pi \\
&= [c_q]_\Pi,
\end{aligned}$$

which proves that $B_1(X) \subset \operatorname{Ker}\beta$.

Now suppose f is a loop such that $[f]_\pi \in \operatorname{Ker}\gamma$. This means that $[f]_H = 0$, or equivalently that the singular 1-chain f is a boundary. On the one hand, because f is a loop based at q, $\beta(f) = [\widetilde{f}]_\Pi = [f]_\Pi$. On the other hand, since β takes boundaries to the identity element of Π, it follows that $[f]_\Pi = 1$, or equivalently that $[f]_\pi$ is in the commutator subgroup. □

Corollary 13.12. *The following spaces have the indicated first homology groups.*

$$\begin{aligned} H_1(\mathbb{S}^1) &\cong \mathbb{Z}; \\ H_1(\mathbb{S}^n) &= 0 \quad \textit{if } n \geq 2; \\ H_1(\underbrace{\mathbb{T}^2 \# \cdots \# \mathbb{T}^2}_{n}) &\cong \mathbb{Z}^{2n}; \\ H_1(\underbrace{\mathbb{P}^2 \# \cdots \# \mathbb{P}^2}_{n}) &\cong \mathbb{Z}^{n-1} \times \mathbb{Z}/\langle 2 \rangle. \end{aligned}$$

The Poincaré homomorphism $\gamma\colon \pi_1(X,q) \to H_1(X)$ can be generalized easily to a homomorphism from $\pi_k(X,q)$ to $H_k(X)$ for any k, called the *Hurewicz homomorphism*. The relationship between the higher homotopy and homology groups is not so simple however, except in one important special case: The *Hurewicz theorem*, proved by Witold Hurewicz in 1934, says that if X is path connected and $\pi_j(X,q)$ is trivial for $1 \leq j < k$, then $H_j(X)$ is trivial for the same values of j and the Hurewicz homomorphism is an isomorphism between $\pi_k(X,q)$ and $H_k(X)$. For a proof, see [Spa89] or [Whi78].

The Mayer–Vietoris Theorem

Our main tool for computing higher-dimensional homology groups will be a result analogous to the Seifert–Van Kampen theorem, in that it gives a recipe for computing the homology groups of a space that is the union of two open sets in terms of the homology of the two open sets and that of their intersection.

Statement of the Theorem

The setup for the theorem is similar to that of the Seifert–Van Kampen theorem: We are given a space X and two open subsets $U, V \subset X$ whose union is X. (In this case, there is no requirement that any of the spaces be path connected.) There are four inclusion maps

$$\begin{tikzcd} & U \arrow[dr, "k"] & \\ U \cap V \arrow[ur, "i"] \arrow[dr, "j"'] & & X, \\ & V \arrow[ur, "l"'] & \end{tikzcd}$$

all of which induce homology homomorphisms.

Theorem 13.13 (Mayer–Vietoris). *Let X be a topological space, and let U, V be open subsets of X whose union is X. Then for each p there is a homomorphism $\partial_* \colon H_p(X) \to H_{p-1}(U \cap V)$ such that the following sequence is exact:*

$$\cdots \xrightarrow{\partial_*} H_p(U \cap V) \xrightarrow{i_* \oplus j_*} H_p(U) \oplus H_p(V) \xrightarrow{k_* - l_*} H_p(X) \\ \xrightarrow{\partial_*} H_{p-1}(U \cap V) \xrightarrow{i_* \oplus j_*} \cdots . \tag{13.10}$$

The sequence (13.10) is called the *Mayer–Vietoris sequence* of the triple (X, U, V), and ∂_* is called the *connecting homomorphism*. The other maps are the obvious ones: $(i_* \oplus j_*)[c] = (i_*[c], j_*[c])$ and $(k_* - l_*)([c], [c']) = k_*[c] - l_*[c']$.

Before proving the theorem, let us apply it to an example to show how it can be used to compute homology groups.

Proposition 13.14 (Homology Groups of Spheres). *For $n \geq 1$, $\mathbb{S}^n$ has the following singular homology groups:*

$$H_p(\mathbb{S}^n) \cong \begin{cases} \mathbb{Z} & \text{if } p = 0, \\ 0 & \text{if } 0 < p < n, \\ \mathbb{Z} & \text{if } p = n, \\ 0 & \text{if } p > n. \end{cases}$$

Proof. We use the Mayer–Vietoris sequence as follows. Let N and S denote the north and south poles, and let $U = \mathbb{S}^n \smallsetminus \{N\}$, $V = \mathbb{S}^n \smallsetminus \{S\}$ as in the proof that $\mathbb{S}^n$ is simply connected (Theorem 8.7). Part of the Mayer–Vietoris sequence reads

$$H_p(U) \oplus H_p(V) \to H_p(\mathbb{S}^n) \xrightarrow{\partial_*} H_{p-1}(U \cap V) \to H_{p-1}(U) \oplus H_{p-1}(V).$$

Because U and V are contractible, when $p > 1$ this sequence reduces to

$$0 \to H_p(\mathbb{S}^n) \xrightarrow{\partial_*} H_{p-1}(U \cap V) \to 0,$$

from which it follows that ∂_* is an isomorphism. Thus, since $U \cap V$ is homotopy equivalent to $\mathbb{S}^{n-1}$,

$$H_p(\mathbb{S}^n) \cong H_{p-1}(U \cap V) \cong H_{p-1}(\mathbb{S}^{n-1}) \quad \text{for } p > 1,\ n \geq 1. \tag{13.11}$$

We will prove the proposition by induction on n. In the case $n = 1$, $H_0(\mathbb{S}^1) \cong H_1(\mathbb{S}^1) \cong \mathbb{Z}$ by Proposition 13.5 and Corollary 13.12. For $p > 1$, (13.11) shows that $H_p(\mathbb{S}^1) \cong H_{p-1}(\mathbb{S}^0)$. Since each component of $\mathbb{S}^0$ is a one-point space, $H_{p-1}(\mathbb{S}^0)$ is the trivial group by Proposition 13.6 and Lemma 13.4.

Now let $n > 1$, and suppose the result is true for $\mathbb{S}^{n-1}$. The cases $p = 0$ and $p = 1$ are again taken care of by Proposition 13.5 and Corollary 13.12. For $p > 1$, (13.11) and the inductive hypothesis give

$$H_p(\mathbb{S}^n) \cong H_{p-1}(\mathbb{S}^{n-1}) \cong \begin{cases} 0 & \text{if } p < n, \\ \mathbb{Z} & \text{if } p = n, \\ 0 & \text{if } p > n. \end{cases}$$

□

Proof of the Theorem

To prove the Mayer–Vietoris theorem, we need to introduce a few more basic concepts from homological algebra.

Suppose C_*, D_*, and E_* are chain complexes. A sequence of chain maps

$$\cdots \to C_* \xrightarrow{F} D_* \xrightarrow{G} E_* \to \cdots$$

is said to be *exact* if each of the sequences

$$\cdots \to C_p \xrightarrow{F} D_p \xrightarrow{G} E_p \to \cdots$$

is exact.

The following lemma is a standard result in homological algebra. The proof, which is easier to do than it is to read, uses a technique commonly called "diagram chasing." The best way to understand it is probably to read the first paragraph or two to get an idea of how the arguments go, and then sit down with pencil and paper and carry out the rest yourself.

Lemma 13.15 (The Zigzag Lemma). *Let*

$$0 \to C_* \xrightarrow{F} D_* \xrightarrow{G} E_* \to 0$$

be a short exact sequence of chain maps. Then for each p there is a connecting homomorphism $\partial_* \colon H_p(E_*) \to H_{p-1}(C_*)$ such that the following sequence is exact:

$$\cdots \xrightarrow{\partial_*} H_p(C_*) \xrightarrow{F_*} H_p(D_*) \xrightarrow{G_*} H_p(E_*) \xrightarrow{\partial_*} H_{p-1}(C_*) \xrightarrow{F_*} \cdots. \tag{13.12}$$

The sequence (13.12) is called the *long exact homology sequence* associated with the given short exact sequence of chain maps.

Proof. Consider the diagram

$$\begin{tikzcd}
0 \arrow[r] & C_{p+1} \arrow[r, "F"] \arrow[d, "\partial"] & D_{p+1} \arrow[r, "G"] \arrow[d, "\partial"] & E_{p+1} \arrow[r] \arrow[d, "\partial"] & 0 \\
0 \arrow[r] & C_p \arrow[r, "F"] \arrow[d, "\partial"] & D_p \arrow[r, "G"] \arrow[d, "\partial"] & E_p \arrow[r] \arrow[d, "\partial"] & 0 \\
0 \arrow[r] & C_{p-1} \arrow[r, "F"] \arrow[d, "\partial"] & D_{p-1} \arrow[r, "G"] \arrow[d, "\partial"] & E_{p-1} \arrow[r] \arrow[d, "\partial"] & 0 \\
0 \arrow[r] & C_{p-2} \arrow[r, "F"] & D_{p-2} \arrow[r, "G"] & E_{p-2} \arrow[r] & 0.
\end{tikzcd}$$

The hypothesis is that this diagram commutes and the horizontal rows are exact.

We will use brackets to denote the homology class of a cycle in any of these groups, so for example if $d_p \in D_p$ satisfies $\partial d_p = 0$, then $[d_p] \in H_p(D_*)$. To define the connecting homomorphism ∂_*, let $[e_p] \in H_p(E_*)$ be arbitrary. This means that $e_p \in E_p$ and $\partial e_p = 0$. Surjectivity of $G\colon D_p \to E_p$ means that there is an element $d_p \in D_p$ such that $Gd_p = e_p$, and then commutativity of the diagram means that $G\partial d_p = \partial G d_p = \partial e_p = 0$, so $\partial d_p \in \operatorname{Ker} G$. By exactness at D_{p-1} there is an element $c_{p-1} \in C_{p-1}$ such that $Fc_{p-1} = \partial d_p$. Now, $F\partial c_{p-1} = \partial F c_{p-1} = \partial\partial d_p = 0$, and since F is injective, $\partial c_{p-1} = 0$. Therefore, c_{p-1} represents a homology class in $H_{p-1}(C_*)$.

We wish to set $\partial_*[e_p] = [c_{p-1}]$. To do so, we have to make sure the homology class of c_{p-1} does not depend on any of the choices we made along the way. Another set of choices will be of the form $e'_p \in E_p$ such that $e_p - e'_p = \partial e_{p+1}$, $d'_p \in D_p$ such that $Gd'_p = e'_p$, and $c'_{p-1} \in C_{p-1}$ such that $Fc'_{p-1} = \partial d'_p$. Because G is surjective, there exists $d_{p+1} \in D_{p+1}$ such that $Gd_{p+1} = e_{p+1}$. Then $G\partial d_{p+1} = \partial G d_{p+1} = \partial e_{p+1} = e_p - e'_p$, so $G(d_p - d'_p) = e_p - e'_p = G\partial d_{p+1}$. Since $d_p - d'_p - \partial d_{p+1} \in \operatorname{Ker} G$, there exists $c_p \in C_p$ such that $Fc_p = d_p - d'_p - \partial d_{p+1}$. Now $F\partial c_p = \partial F c_p = \partial(d_p - d'_p - \partial d_{p+1}) = \partial d_p - \partial d'_p = Fc_{p-1} - Fc'_{p-1}$. Since F is injective, this implies $\partial c_p = c_{p-1} - c'_{p-1}$, or $[c_{p-1}] = [c'_{p-1}]$. To summarize, we have defined $\partial_*[e_p] = [c_{p-1}]$, provided that there exists $d_p \in D_p$ such that

$$Gd_p = e_p; \qquad Fc_{p-1} = \partial d_p.$$

To prove that ∂_* is a homomorphism, just note that if $\partial_*[e_p] = [c_{p-1}]$ and $\partial_*[e'_p] = [c'_{p-1}]$, there exist $d_p, d'_p \in D_p$ such that $Gd_p = e_p$, $Gd'_p = e'_p$, $Fc_{p-1} = \partial d_p$, $Fc'_{p-1} = \partial d'_p$. It follows immediately that $G(d_p + d'_p) = e_p + e'_p$

and $F(c_{p-1} + c'_{p-1}) = \partial(d_p + d'_p)$, and so $\partial_*[e_p + e'_p] = [c_{p-1} + c'_{p-1}] = \partial_*[e_p] + \partial_*[e'_p]$.

Now we have to prove exactness of (13.12). Let us start at $H_p(C_*)$. Suppose $[c_p] = \partial_*[e_{p+1}]$. Then looking back at the definition of ∂_*, there is some d_{p+1} such that $Fc_p = \partial d_{p+1}$, so $F_*[c_p] = [Fc_p] = [\partial d_{p+1}] = 0$; thus $\operatorname{Im}\partial_* \subset \operatorname{Ker} F_*$. Conversely, if $F_*[c_p] = [Fc_p] = 0$, there is some $d_{p+1} \in D_{p+1}$ such that $Fc_p = \partial d_{p+1}$, and then $\partial G d_{p+1} = G\partial d_{p+1} = GFc_p = 0$. In particular, this means $e_{p+1} = Gd_{p+1}$ represents a homology class in $H_{p+1}(E_*)$, and threading through the definition of ∂_* we find that $\partial_*[e_{p+1}] = [c_p]$. Thus $\operatorname{Ker} F_* \subset \operatorname{Im}\partial_*$.

Next we prove exactness at $H_p(D_*)$. From $GF = 0$ it follows immediately that $G_*F_* = 0$, so $\operatorname{Im} F_* \subset \operatorname{Ker} G_*$. If $G_*[d_p] = [Gd_p] = 0$, there exists $e_{p+1} \in E_{p+1}$ such that $\partial e_{p+1} = Gd_p$. By surjectivity of G, there is some $d_{p+1} \in D_{p+1}$ such that $Gd_{p+1} = e_{p+1}$, and then $G\partial d_{p+1} = \partial G d_{p+1} = \partial e_{p+1} = Gd_p$. Thus $d_p - \partial d_{p+1} \in \operatorname{Ker} G = \operatorname{Im} F$, so there is $c_p \in C_p$ with $Fc_p = d_p - \partial d_{p+1}$. Moreover, $F\partial c_p = \partial F c_p = \partial(d_p - \partial d_{p+1}) = \partial d_p = 0$, so $\partial c_p = 0$ by injectivity of F. Thus c_p represents a homology class in $H_p(C_*)$, and $F_*[c_p] = [Fc_p] = [d_p - \partial d_{p+1}] = [d_p]$. This proves that $\operatorname{Ker} G_* \subset \operatorname{Im} F_*$.

Finally, we prove exactness at $H_p(E_*)$. Suppose $[e_p] \in \operatorname{Im} G_*$. This means that $[e_p] = G_*[d_p]$ for some $d_p \in D_p$ with $\partial d_p = 0$, so $e_p = Gd_p + \partial e_{p+1}$. Replacing e_p with $e_p - \partial e_{p+1}$, we may assume $Gd_p = e_p$. Then by definition $\partial_*[e_p] = [c_{p-1}]$, where $c_{p-1} \in C_{p-1}$ is chosen so that $Fc_{p-1} = \partial d_p$. But in this case $\partial d_p = 0$, so we may take $c_{p-1} = 0$ and therefore $\partial_*[e_p] = 0$. Conversely, suppose $\partial_*[e_p] = 0$. This means that there exists $d_p \in D_p$ such that $Gd_p = e_p$ and $c_{p-1} \in C_{p-1}$ such that $Fc_{p-1} = \partial d_p$, and c_{p-1} is a boundary. Writing $c_{p-1} = \partial c_p$, we find that $\partial F c_p = F\partial c_p = Fc_{p-1} = \partial d_p$. Thus $d_p - Fc_p$ represents a homology class, and $G_*[d_p - Fc_p] = [Gd_p - GFc_p] = [e_p - 0] = [e_p]$. Therefore, $\operatorname{Ker}\partial_* \subset \operatorname{Im} G_*$, and the proof is complete. □

The connecting homomorphism in the long exact homology sequence satisfies an important naturality property, which we will use later in this chapter.

Proposition 13.16 (Naturality of Connecting Homomorphisms). *Suppose*

$$\begin{tikzcd}
0 \arrow[r] & C_* \arrow[r, "F"] \arrow[d, "\kappa"] & D_* \arrow[r, "G"] \arrow[d, "\delta"] & E_* \arrow[r] \arrow[d, "\varepsilon"] & 0 \\
0 \arrow[r] & C'_* \arrow[r, "F'"] & D'_* \arrow[r, "G'"] & E'_* \arrow[r] & 0
\end{tikzcd} \tag{13.13}$$

is a commutative diagram of chain maps in which the horizontal rows are exact. Then the following diagram commutes for each p:

$$\begin{CD}
H_p(E_*) @>{\partial_*}>> H_{p-1}(C_*) \\
@V{\varepsilon_*}VV @VV{\kappa_*}V \\
H_p(E'_*) @>>{\partial_*}> H_{p-1}(C'_*).
\end{CD}$$

Proof. Let $[e_p] \in H_p(E_*)$ be arbitrary. Then $\partial_*[e_p] = [c_{p-1}]$, where $Fc_{p-1} = \partial d_p$ for some d_p such that $Gd_p = e_p$. Then by commutativity of (13.13),

$$\begin{aligned}
F'(\kappa c_{p-1}) &= \delta F c_{p-1} = \delta \partial d_p = \partial(\delta d_p); \\
G'(\delta d_p) &= \varepsilon G d_p = \varepsilon e_p.
\end{aligned}$$

By definition, this means that

$$\partial_* \varepsilon_* [e_p] = \partial_* [\varepsilon e_p] = [\kappa c_{p-1}] = \kappa_* [c_{p-1}] = \kappa_* \partial_* [e_p],$$

which was to be proved. □

While we are on the subject, here is another algebraic result whose proof is a routine diagram chase.

Lemma 13.17 (The Five Lemma). *Suppose the horizontal rows are exact in the following commutative diagram of abelian groups and homomorphisms:*

$$\begin{CD}
A_1 @>{\alpha_1}>> A_2 @>{\alpha_2}>> A_3 @>{\alpha_3}>> A_4 @>{\alpha_4}>> A_5 \\
@VV{f_1}V @VV{f_2}V @VV{f_3}V @VV{f_4}V @VV{f_5}V \\
B_1 @>{\beta_1}>> B_2 @>{\beta_2}>> B_3 @>{\beta_3}>> B_4 @>{\beta_4}>> B_5.
\end{CD}$$

If f_1, f_2, f_4, and f_5 are isomorphisms, then f_3 is also.

Proof. We will prove that f_3 is surjective, and leave the proof of injectivity to you. Suppose $b_3 \in B_3$ is arbitrary. By surjectivity of f_4, there exists $a_4 \in A_4$ such that $f_4 a_4 = \beta_3 b_3$. By commutativity and exactness, $f_5 \alpha_4 a_4 = \beta_4 f_4 a_4 = \beta_4 \beta_3 b_3 = 0$. Since f_5 is injective, this means that $\alpha_4 a_4 = 0$, and by exactness again there exists $a_3 \in A_3$ such that $a_4 = \alpha_3 a_3$. Substituting, we obtain $\beta_3 b_3 = f_4 a_4 = f_4 \alpha_3 a_3 = \beta_3 f_3 a_3$, which implies $b_3 - f_3 a_3 \in \operatorname{Ker} \beta_3 = \operatorname{Im} \beta_2$. Thus there exists $b_2 \in B_2$ such that $\beta_2 b_2 = b_3 - f_3 a_3$, and by surjectivity of f_2 there exists $a_2 \in A_2$ such that $b_2 = f_2 a_2$. Summarizing, we have $b_3 - f_3 a_3 = \beta_2 b_2 = \beta_2 f_2 a_2 = f_3 \alpha_2 a_2$, which shows that $b_3 \in \operatorname{Im} f_3$. □

Exercise 13.3. Finish the proof of the five lemma by showing that f_3 is injective.

Proof of the Mayer–Vietoris theorem. Let X, U, and V be as in the statement of the theorem. Consider the three chain complexes $C_*(U \cap V)$, $C_*(U) \oplus C_*(V)$, and $C_*(X)$. (The boundary operator in the second complex is $\partial(c, c') = (\partial c, \partial c')$.) We are interested in the following sequence of maps:

$$0 \to C_p(U \cap V) \xrightarrow{i_\# \oplus j_\#} C_p(U) \oplus C_p(V) \xrightarrow{k_\# - l_\#} C_p(X).$$

Because the chain maps $i_\#, j_\#, k_\#, l_\#$ are all induced by inclusion, their action is simply to consider a chain in one space as a chain in a bigger space. It is easy to check that $i_\# \oplus j_\#$ and $k_\# - l_\#$ are chain maps and that this sequence is exact, as far as it goes. For example, if c and c' are chains in U and V, respectively, such that $k_\# c - l_\# c' = 0$, this means that they are equal when thought of as chains in X. For this to be the case, the two chains must be identical, and the image of each singular simplex in each chain must actually lie in $U \cap V$. Thus c is actually a chain in $U \cap V$, and $(c, c') = (i_\# \oplus j_\#)(c)$. The rest of the conditions for exactness are similar.

Unfortunately, however, $k_\# - l_\#$ is not surjective. It is not hard to see why: The image of this map is the set of all p-chains in X that can be written as a sum of a chain in U plus a chain in V. Any singular p-simplex whose image is not contained in either U or V therefore defines a chain that is not in the image. Thus we cannot apply the zigzag lemma directly to this sequence.

Instead, we use the following subterfuge: Let $\mathcal{U}$ denote the open cover of X consisting of the sets U and V, and for each p let $C_p^{\mathcal{U}}(X)$ denote the subgroup of $C_p(X)$ generated by singular simplices whose images lie either entirely in U or entirely in V. The boundary operator carries $C_p^{\mathcal{U}}(X)$ into $C_{p-1}^{\mathcal{U}}(X)$, so we get a new chain complex $C_*^{\mathcal{U}}(X)$. Clearly, the following sequence is exact:

$$0 \to C_*(U \cap V) \xrightarrow{i_\# \oplus j_\#} C_*(U) \oplus C_*(V) \xrightarrow{k_\# - l_\#} C_*^{\mathcal{U}}(X) \to 0.$$

The zigzag lemma then yields the following long exact homology sequence:

$$\begin{aligned} \cdots \xrightarrow{\partial_*} H_p(U \cap V) \xrightarrow{i_* \oplus j_*} H_p(U) \oplus H_p(V) &\xrightarrow{k_* - l_*} H_p^{\mathcal{U}}(X) \\ &\xrightarrow{\partial_*} H_{p-1}(U \cap V) \xrightarrow{i_* \oplus j_*} \cdots, \end{aligned} \tag{13.14}$$

where $H_p^{\mathcal{U}}(X)$ is the pth homology group of the complex $C_*^{\mathcal{U}}(X)$. This is almost what we are looking for. The final step is to invoke Proposition 13.18 below, which shows that inclusion $C_*^{\mathcal{U}}(X) \hookrightarrow C_*(X)$ induces a homology isomorphism $H_p^{\mathcal{U}}(X) \cong H_p(X)$. Making this substitution into (13.14), we obtain the Mayer–Vietoris sequence. □

The missing step in the above proof is the fact that the singular homology of X can be detected by looking only at singular simplices that lie either in U or in V. More generally, suppose $\mathcal{U}$ is any open cover of X. A singular chain c is said to be $\mathcal{U}$-*small* if every singular simplex that appears in c has image lying entirely in one of the open sets of $\mathcal{U}$. Let $C_p^{\mathcal{U}}(X)$ denote the subgroup of $C_p(X)$ consisting of $\mathcal{U}$-small chains, and let $H_p^{\mathcal{U}}(X)$ denote the homology of the complex $C_*^{\mathcal{U}}(X)$.

Proposition 13.18. *If $\mathcal{U}$ is any open cover of X, the inclusion map $C_*^{\mathcal{U}}(X) \to C_*(X)$ induces a homology isomorphism $H_p^{\mathcal{U}}(X) \cong H_p(X)$ for all p.*

The idea of the proof is simple, although the technical details are somewhat involved. If $\sigma\colon \Delta_p \to X$ is any singular p-simplex, the plan is to show that there is a homologous p-chain obtained by subdividing Δ_p and restricting σ to each of the p-simplices of the subdivision. If we subdivide sufficiently finely, we can ensure that each of the resulting simplices will be $\mathcal{U}$-small. The tricky part is to do this in a systematic way that allows us to keep track of the boundary operators. Before the formal proof, let us lay some groundwork.

Recall the barycentric subdivision of a Euclidean simplicial complex introduced in Chapter 5. It is obtained by replacing each simplex with its barycenter together with cones on appropriate lower-dimensional simplices from the barycenter. To define a subdivision operator in singular homology, we begin by extending the cone construction to affine singular simplices.

If $\sigma = \alpha(v_0, \dots, v_p)$ is an affine singular p-simplex in some convex set $K \subset \mathbb{R}^m$ and v is any point in K, we define an affine singular $(p+1)$-simplex $v * \sigma$ called the *cone* on σ from v by

$$v * \sigma = v * \alpha(v_0, \dots, v_p) = \alpha(v, v_0, \dots, v_p).$$

In other words, $v * \sigma\colon \Delta_{p+1} \to K$ is the unique affine simplex that sends e_0 to v and whose 0th face map is equal to σ. We extend this operator to affine chains by linearity: $v * \left(\sum_i n_i \sigma_i\right) = \sum_i n_i (v * \sigma_i)$. (It is not defined for arbitrary singular chains.)

Lemma 13.19. *If c is an affine chain, then*

$$\partial(v * c) = c - v * \partial c. \tag{13.15}$$

Proof. For an affine simplex $\sigma = \alpha(v_0, \dots, v_p)$, this is just a computation:

$$\begin{aligned}
\partial(v * \sigma) &= \partial\alpha(v, v_0, \dots, v_p)\\
&= \sum_{i=0}^{p+1} (-1)^i \alpha(v, v_0, \dots, v_p) \circ F_{i,p}\\
&= \alpha(v_0, \dots, v_p) + \sum_{i=0}^{p} (-1)^{i+1} \alpha(v, v_0, \dots, \widehat{v}_i, \dots, v_p)\\
&= \sigma - v * \partial\sigma.
\end{aligned}$$

The general case follows by linearity. $\square$

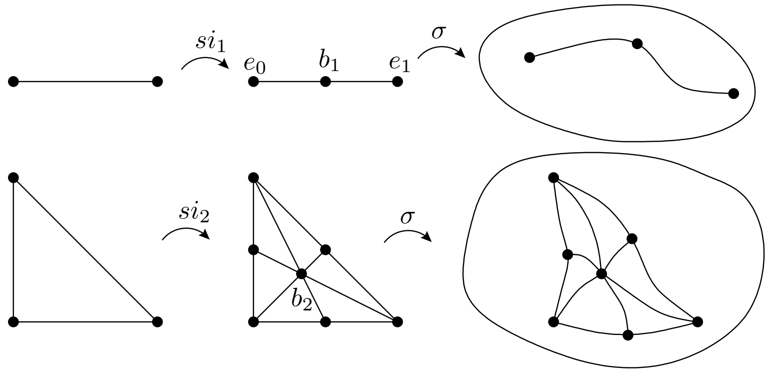


FIGURE 13.9. Singular subdivisions in dimensions 1 and 2.

Now we define an operator s taking p-chains to p-chains, called the *singular subdivision operator.* We define it first for affine chains in $\mathbb{R}^n$, by induction on p. For $p = 0$, simply set $s = \mathrm{Id}$. For $p > 0$, set

$$sc = b_p * s\partial c,$$

where b_p is the barycenter of Δ_p. For example, if $i_p \colon \Delta_p \to \Delta_p$ is the identity map, thought of as an affine singular p-simplex in Δ_p, si_p is a sum of affine simplices mapping homeomorphically onto the simplices of the barycentric subdivision of Δ_p.

For a singular p-simplex σ in any space X, note that $\sigma = \sigma_\# i_p$, where $\sigma_\# \colon C_p(\Delta_p) \to C_p(X)$ is the chain map obtained from the continuous map $\sigma \colon \Delta_p \to X$. We define $s\sigma = \sigma_\#(si_p)$, and extend by linearity to all of $C_p(X)$. Low-dimensional examples are pictured in Figure 13.9. We can iterate s to obtain operators $s^2 = s \circ s$ and more generally $s^k = s \circ s^{k-1}$.

Lemma 13.20. *The singular subdivision operator has the following properties.*

(*a*) $s \circ f_\# = f_\# \circ s$ *for any continuous map* f.

(*b*) $\partial \circ s = s \circ \partial$.

(*c*) *Given an open cover* $\mathcal{U}$ *of* X *and any* $c \in C_p(X)$, *there exists* m *such that* $s^m c \in C_p^{\mathcal{U}}(X)$.

Proof. The first identity follows immediately from the definition of s:

$$s(f_\#\sigma) = s(f \circ \sigma) = (f \circ \sigma)_\#(si_p) = f_\#\sigma_\#(si_p) = f_\#(s\sigma).$$

The second is proved by induction on p. For $p = 0$ it is trivial, and for $p > 0$ we use part (a), (13.15), and the inductive hypothesis to compute

$$\begin{aligned}
\partial s\sigma &= \partial\sigma_{\#}(b_p * s\partial i_p)\\
&= \sigma_{\#}\partial(b_p * s\partial i_p)\\
&= \sigma_{\#}(s\partial i_p - b_p * \partial s\partial i_p)\\
&= s\sigma_{\#}\partial i_p - \sigma_{\#}b_p * (s\partial\partial i_p)\\
&= s\partial\sigma_{\#}i_p - 0\\
&= s\partial\sigma.
\end{aligned}$$

To prove (c), define the *mesh* of an affine chain c in $\mathbb{R}^n$ to be the maximum of the diameters of the images of the affine simplices that appear in c. By Lemma 5.18, by choosing m large enough, we can make the mesh of $s^m i_p$ arbitrarily small.

If σ is any singular simplex in X, by the Lebesgue number lemma there exists $\delta > 0$ such that σ maps any subset of Δ_p of diameter less than δ into one of the sets of $\mathcal{U}$. In particular, if c is an affine chain in Δ_p whose mesh is less than δ, then $\sigma_{\#}c \in C_p^{\mathcal{U}}(X)$. Choose δ to be the minimum of the Lebesgue numbers for all the singular simplices appearing in c, and choose m large enough that $s^m i_p$ has mesh less than δ. Then $s^m\sigma = \sigma_{\#}(s^m i_p) \in C_p^{\mathcal{U}}(X)$ as desired. □

With the machinery we have set up, it is now an easy matter to prove Proposition 13.18.

Proof of Proposition 13.18. The crux of the proof is the construction of a chain homotopy between s and the identity map of $C_p(X)$. Recall that this is a homomorphism $h\colon C_p(X) \to C_{p+1}(X)$ satisfying

$$\partial \circ h + h \circ \partial = \mathrm{Id} - s. \tag{13.16}$$

We define h by induction on p. For $p = 0$, h is the zero homomorphism. For $p > 0$, if σ is a singular p-simplex in any space, define

$$h\sigma = \sigma_{\#}b_p * (i_p - si_p - h\partial i_p).$$

As with s, it is a trivial consequence of the definition that $h \circ f_{\#} = f_{\#} \circ h$ for any continuous map f. Observe also that if σ is a $\mathcal{U}$-small simplex, then $h\sigma$ is a $\mathcal{U}$-small chain, so h also maps $C_p^{\mathcal{U}}(X)$ to $C_{p+1}^{\mathcal{U}}(X)$.

The chain homotopy identity (13.16) is proved by induction on p. For $p = 0$ it is immediate because $h = \partial = 0$ and $s = \mathrm{Id}$. Suppose it holds for $(p-1)$-chains in all spaces. If σ is a singular p-simplex, then

$$\begin{aligned}
\partial h\sigma &= \partial\sigma_{\#}b_p * (i_p - si_p - h\partial i_p)\\
&= \sigma_{\#}\partial b_p * (i_p - si_p - h\partial i_p)\\
&= \sigma_{\#}(i_p - si_p - h\partial i_p) - \sigma_{\#}b_p * (\partial i_p - \partial si_p - \partial h\partial i_p).
\end{aligned}$$

The expression inside the second set of parentheses is equal to $\partial i_p - s\partial i_p - \partial h\partial i_p - h\partial\partial i_p$, which is zero by the inductive hypothesis because ∂i_p is a $(p-1)$-chain. Therefore,

$$\partial h\sigma = \sigma_{\#} i_p - s\sigma_{\#} i_p - h\partial\sigma_{\#} i_p = \sigma - s\sigma - h\partial\sigma,$$

which was to be proved.

Now if c is any singular cycle in X, (13.16) shows that

$$c - sc = \partial hc + h\partial c = \partial hc,$$

so sc differs from c by a boundary. If $c \in C_p^{\mathfrak{U}}(X)$, the difference is the boundary of a chain in $C_{p+1}^{\mathfrak{U}}(X)$. By induction the same is true for $s^m c$ for any positive integer m. Moreover, $s^m c$ is a cycle because s commutes with ∂.

The inclusion map $\iota\colon C_p^{\mathfrak{U}}(X) \hookrightarrow C_p(X)$ is clearly a chain map, and so induces a homology homomorphism $\iota_*\colon H_p^{\mathfrak{U}}(X) \to H_p(X)$. It is surjective because for any $[c] \in H_p(X)$ we can choose m large enough that $s^m c \in C_p^{\mathfrak{U}}(X)$, and the argument above shows that c is homologous to $s^m c$. To prove injectivity, suppose $[c] \in H_p^{\mathfrak{U}}(X)$ satisfies $\iota_*[c] = 0$. This means that there is a $(p+1)$-chain $b \in C_{p+1}(X)$ such that $c = \partial b$. Choose m large enough that $s^m b \in C_{p+1}^{\mathfrak{U}}(X)$. Then $\partial s^m b = s^m \partial b = s^m c$, which differs from c by the boundary of a chain in $C_{p+1}^{\mathfrak{U}}(X)$. Thus c represents the zero element of $H_p^{\mathfrak{U}}(X)$. □

Applications

In this section we give a sampling of the numerous significant applications of homology theory to the study of manifolds. These applications are based on the fact that the homology groups give us a simple way to distinguish topologically between spheres of different dimensions and between homotopy classes of maps of spheres, something that the fundamental group could not do. Several more such applications are outlined in the problems.

Invariance of Dimension

The dimension of a manifold is part of its definition: An n-dimensional manifold is one that admits local homeomorphisms to open subsets of $\mathbb{R}^n$. It seems intuitively obvious that dimension ought to be a topological invariant: A manifold of dimension n ought not to be homeomorphic to one of some other dimension. This is true, but the proof is decidedly nontrivial. A proof for $n = 1$ was outlined in Problem 4-1 using the fact that $\mathbb{R}^n \smallsetminus \{0\}$ is connected when $n > 1$. Similarly, Problem 8-5 suggested a proof for $n = 2$ using the fact that $\mathbb{R}^n \smallsetminus \{0\}$ is simply connected when $n > 2$. But

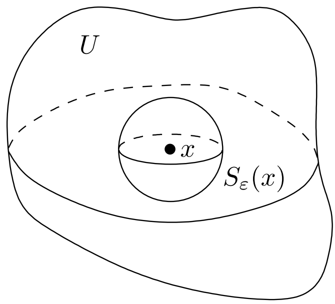


FIGURE 13.10. Proof of Lemma 13.21.

neither connectedness nor simple connectedness can distinguish $\mathbb{R}^n \smallsetminus \{0\}$ from $\mathbb{R}^m \smallsetminus \{0\}$ when both m and n are larger than 2. Homology can.

Lemma 13.21. *Let U be an open subset of $\mathbb{R}^n$, $n \geq 2$. If x is any point in U, then $H^{n-1}(U \smallsetminus \{x\}) \neq 0$.*

Proof. Choose $\varepsilon > 0$ small enough that the sphere $S_\varepsilon(x)$ of radius ε about x is contained in U (Figure 13.10). Consider the following commutative diagram of inclusion maps:

$$\begin{tikzcd} S_\varepsilon(x) \arrow[d, "i"'] \arrow[dr, "k"] & \\ U \smallsetminus \{x\} \arrow[r, "j"'] & \mathbb{R}^n \smallsetminus \{x\}. \end{tikzcd}$$

These induce homology homomorphisms

$$\begin{tikzcd} H_{n-1}(S_\varepsilon(x)) \arrow[d, "i_*"'] \arrow[dr, "k_*"] & \\ H_{n-1}(U \smallsetminus \{x\}) \arrow[r, "j_*"'] & H_{n-1}(\mathbb{R}^n \smallsetminus \{x\}). \end{tikzcd}$$

Because k is a homotopy equivalence, k_* is an isomorphism. This implies that i_* is injective (and j_* is surjective). Since $H_{n-1}(S_\varepsilon(x))$ is not trivial, neither is $H_{n-1}(U \smallsetminus \{x\})$. $\square$

Theorem 13.22 (Invariance of Dimension). *If $m \neq n$, a topological space cannot be both an n-manifold and an m-manifold.*

Proof. The zero-dimensional case is easy to dispose of, because a 0-manifold is a discrete space, and points in a positive-dimensional manifold are not open sets. Suppose M is both an m-manifold and an n-manifold, and assume that $n > m \geq 1$. Any $x \in M$ has a neighborhood $U \subset M$ that is homeomorphic to $\mathbb{R}^n$. Because an open subset of a manifold is again a manifold, U is also an m-manifold, so x has a neighborhood $V \subset U$ that is homeomorphic to $\mathbb{R}^m$. On the one hand, because V is homeomorphic to an open subset in $\mathbb{R}^n$, Lemma 13.21 implies $H^{n-1}(V \smallsetminus \{x\}) \neq 0$. On the other hand, $V \smallsetminus \{x\} \approx \mathbb{R}^m \smallsetminus \{0\} \simeq \mathbb{S}^{m-1}$, so $H^{n-1}(V \smallsetminus \{x\}) = 0$. □

Degree Theory for Spheres

In Problem 8-7, we defined the degree of a continuous map $f\colon \mathbb{S}^1 \to \mathbb{S}^1$. Homology theory allows us to extend this definition to higher-dimensional spheres. Suppose $n \geq 1$. Because $H_n(\mathbb{S}^n)$ is infinite cyclic, if $f\colon \mathbb{S}^n \to \mathbb{S}^n$ is any continuous map, $f_*\colon H_n(\mathbb{S}^n) \to H_n(\mathbb{S}^n)$ is multiplication by a unique integer (Exercise A.28), called the *degree* of f and denoted by $\deg f$.

Proposition 13.23. *Suppose $n \geq 1$ and $f, g\colon \mathbb{S}^n \to \mathbb{S}^n$ are continuous maps.*

(*a*) *If $f \simeq g$, then* $\deg f = \deg g$.

(*b*) $\deg(g \circ f) = (\deg g)(\deg f)$.

(*c*) *The identity map has degree* 1.

Proof. Part (a) follows from the fact that homotopic maps induce the same homology homomorphism; part (b) from the fact that $(g \circ f)_* = g_* \circ f_*$; and part (c) from the fact that the identity map induces the identity homomorphism on homology. □

For a map $f\colon \mathbb{S}^1 \to \mathbb{S}^1$ the definition of $\deg f$ given in Problem 8-7 was the unique integer k such that the homomorphism $(\rho \circ f)_*\colon \pi_1(\mathbb{S}^1, 1) \to \pi_1(\mathbb{S}^1, 1)$ is given by $\gamma \mapsto \gamma^k$, where ρ is the rotation taking $f(1)$ to 1. For the moment, let us call that definition the *homotopic degree* of f, and the degree we have defined in this chapter its *homological degree*.

Lemma 13.24. *The homological degree and the homotopic degree of a continuous map $f\colon \mathbb{S}^1 \to \mathbb{S}^1$ are equal.*

Proof. By examining the definition of the Poincaré homomorphism γ, it is easy to see that the following diagram commutes:

$$\begin{tikzcd}
\pi_1(\mathbb{S}^1, 1) \arrow[r, "(\rho \circ f)_*"] \arrow[d, "\gamma"'] & \pi_1(\mathbb{S}^1, 1) \arrow[d, "\gamma"] \\
H_1(\mathbb{S}^1) \arrow[r, "(\rho \circ f)_*"'] & H_1(\mathbb{S}^1).
\end{tikzcd}$$

It follows that the homotopic degree of f is equal to the homological degree of $\rho \circ f$. Since the rotation ρ is homotopic to the identity map, it has homological degree 1, so the homological degree of $\rho \circ f$ is equal to that of f. □

Some of the most important applications of degree theory come from considering the antipodal map $A\colon \mathbb{S}^n \to \mathbb{S}^n$ given by $A(x) = -x$.

Lemma 13.25. *For each $n \geq 1$, the antipodal map $A\colon \mathbb{S}^n \to \mathbb{S}^n$ has degree $(-1)^{n+1}$.*

Proof. Consider the *reflection maps* $R_i\colon \mathbb{S}^n \to \mathbb{S}^n$ given by

$$R_i(x_1, \ldots, x_i, \ldots, x_{n+1}) = (x_1, \ldots, -x_i, \ldots, x_{n+1}).$$

We will prove by induction on n that R_i has degree -1; because the antipodal map is equal to the $(n+1)$-fold composition $R_1 \circ \cdots \circ R_{n+1}$, it follows that A has degree $(-1)^{n+1}$.

Note that if $\deg R_i = -1$ for one value of i the same is true for all of them, because R_i can be obtained from R_j by conjugating with the linear transformation that interchanges x_i and x_j.

For $n = 1$, $R_2(z) = \overline{z}$ in complex notation, which has degree -1 by Problem 8-7. So suppose $n > 1$, and assume that the claim is true for reflections in dimension $n-1$.

Recall that in the course of proving Proposition 13.14 we showed that $H_n(\mathbb{S}^n) \cong H_{n-1}(\mathbb{S}^{n-1})$. In fact, we will refine that argument to show that there is an isomorphism between these groups such that the following diagram commutes:

$$\begin{tikzcd} H_n(\mathbb{S}^n) \arrow[r] \arrow[d, "R_{1*}"'] & H_{n-1}(\mathbb{S}^{n-1}) \arrow[d, "R_{1*}"] \\ H_n(\mathbb{S}^n) \arrow[r] & H_{n-1}(\mathbb{S}^{n-1}). \end{tikzcd} \tag{13.17}$$

From this it follows immediately by induction that R_1 has degree -1 on $\mathbb{S}^n$.

To prove (13.17), let $\mathcal{U} = \{U, V\}$ be the covering of $\mathbb{S}^n$ by contractible open sets used in the proof of Proposition 13.14 (the complements of the north and south poles). Note that R_1 preserves the sets U and V, and therefore induces chain maps that make the following diagram commute:

$$\begin{tikzcd} 0 \arrow[r] & C_*(U \cap V) \arrow[r] \arrow[d, "R_{1\#}"] & C_*(U) \oplus C_*(V) \arrow[r] \arrow[d, "R_{1\#} \oplus R_{1\#}"] & C_*^{\mathcal{U}}(\mathbb{S}^n) \arrow[r] \arrow[d, "R_{1\#}"] & 0 \\ 0 \arrow[r] & C_*(U \cap V) \arrow[r] & C_*(U) \oplus C_*(V) \arrow[r] & C_*^{\mathcal{U}}(\mathbb{S}^n) \arrow[r] & 0. \end{tikzcd}$$

Therefore, by the naturality property of ∂_*, the following diagram also commutes:

$$\begin{tikzcd}
H_n(\mathbb{S}^n) \arrow[r, "\partial_*"] \arrow[d, "R_{1*}"] & H_{n-1}(U \cap V) \arrow[d, "R_{1*}"] & H_{n-1}(\mathbb{S}^{n-1}) \arrow[l, "\iota_*"'] \arrow[d, "R_{1*}"] \\
H_n(\mathbb{S}^n) \arrow[r, "\partial_*"'] & H_{n-1}(U \cap V) & H_{n-1}(\mathbb{S}^{n-1}), \arrow[l, "\iota_*"]
\end{tikzcd}$$

where $\mathbb{S}^{n-1} = \mathbb{S}^n \cap \{x : x_{n+1} = 0\}$ is the equatorial $(n-1)$-sphere and $\iota\colon \mathbb{S}^{n-1} \to U \cap V$ is inclusion. The horizontal maps are isomorphisms: ι_* because ι is a homotopy equivalence, and ∂_* by the argument in the proof of Proposition 13.14. Composing the horizontal isomorphisms and eliminating the middle column, we obtain (13.17). $\square$

Proposition 13.26. *The antipodal map $A\colon \mathbb{S}^n \to \mathbb{S}^n$ is homotopic to the identity map if and only if n is odd.*

Proof. If $n = 2k-1$ is odd, an explicit homotopy $H\colon \mathrm{Id} \simeq A$ is given by

$$\begin{aligned} H(x,t) = ((\cos \pi t)x_1 + (\sin \pi t)x_2, (\cos \pi t)x_2 - (\sin \pi t)x_1, \\ \dots, (\cos \pi t)x_{2k-1} + (\sin \pi t)x_{2k}, (\cos \pi t)x_{2k} - (\sin \pi t)x_{2k-1}). \end{aligned}$$

If $n = 0$, A interchanges the two points of $\mathbb{S}^0$, and so is clearly not homotopic to the identity. When n is even and positive, A has degree -1, while the identity map has degree 1, so they are not homotopic. $\square$

A *vector field* on $\mathbb{S}^n$ is a continuous map $V\colon \mathbb{S}^n \to \mathbb{R}^{n+1}$ such that for each $x \in \mathbb{S}^n$, $V(x)$ is tangent to $\mathbb{S}^n$ at x, or in other words the Euclidean dot product $V(x) \cdot x = 0$. The following theorem is popularly known as the "hairy ball theorem" because in the two-dimensional case it implies that you cannot comb the hair on a hairy billiard ball without introducing a discontinuity somewhere.

Theorem 13.27 (The Hairy Ball Theorem). *There exists a nowhere vanishing vector field on $\mathbb{S}^n$ if and only if n is odd.*

Proof. Suppose there exists such a vector field V. By replacing V with $V/|V|$, we can assume $|V(x)| = 1$ everywhere. We use V to construct a homotopy between the identity map and the antipodal map as follows:

$$H(x,t) = (\cos \pi t)x + (\sin \pi t)V(x).$$

Direct computation, using the facts that $|x|^2 = |V(x)|^2 = 1$ and $x \cdot V(x) = 0$, shows that H takes its values in $\mathbb{S}^n$. Since $H(x,0) = x$ and $H(x,1) = -x$, H is the desired homotopy. By Proposition 13.26, n must be odd.

Conversely, when $n = 2k-1$ is odd, the following explicit vector field is easily checked to be tangent to the sphere and nowhere vanishing:

$$V(x_1, \dots, x_{2k}) = (x_2, -x_1, x_4, -x_3, \dots, x_{2k}, -x_{2k-1}).$$

□

The Homology of a Simplicial Complex

Because both simplicial complexes and singular homology are built out of simplices, it is reasonable to expect that the homology of a simplicial complex should be computable from the combinatorial structure of the complex. In this section we define another kind of homology group associated with a simplicial complex, called simplicial homology groups, and prove that they are isomorphic to the singular homology groups. As an application, we prove the topological invariance of the Euler characteristic.

Let $\mathcal{K}$ be a finite simplicial complex. We define the pth *simplicial chain group* of $\mathcal{K}$, denoted by $C_p^\Delta(\mathcal{K})$, to be the free abelian group on the set of p-simplices in $\mathcal{K}$.

To define the boundary operator, we choose a total ordering of the vertices $(v_1, \dots, v_N)$, and for a p-simplex $\langle v_{k_0}, \dots, v_{k_p}\rangle$ we set

$$\partial\langle v_{k_0}, \dots, v_{k_p}\rangle = \sum_{i=0}^{p} (-1)^i \langle v_{k_0}, \dots, \widehat{v}_{k_i}, \dots, v_{k_p}\rangle \quad \text{if } k_0 < \dots < k_p.$$

This extends uniquely to a homomorphism $\partial\colon C_p^\Delta(\mathcal{K}) \to C_{p-1}^\Delta(\mathcal{K})$.

Lemma 13.28. *For any simplicial p-chain c, $\partial(\partial c) = 0$.*

Proof. It suffices to show this when c is a p-simplex $\sigma = \langle v_{k_0}, \dots, v_{k_p}\rangle$, in which case (assuming that the vertices appear in increasing order),

$$\begin{aligned}
\partial(\partial\sigma) &= \partial \sum_{i=0}^{p} (-1)^i \langle v_{k_0}, \dots, \widehat{v}_{k_i}, \dots, v_{k_p}\rangle \\
&= \sum_{0 \le j < i \le p} (-1)^{i+j} \langle v_{k_0}, \dots, \widehat{v}_{k_j}, \dots, \widehat{v}_{k_i}, \dots, v_{k_p}\rangle \\
&\quad + \sum_{0 \le i \le j \le p-1} (-1)^{i+j} \langle v_{k_0}, \dots, \widehat{v}_{k_i}, \dots, \widehat{v}_{k_{j+1}}, \dots, v_{k_p}\rangle.
\end{aligned}$$

After $j = i' - 1$ and $i = j'$ are substituted in the second sum, these two sums cancel each other term by term. □

We define

$$Z_p^{\Delta}(\mathcal{K}) = \operatorname{Ker} \partial \colon C_p^{\Delta}(\mathcal{K}) \to C_{p-1}^{\Delta}(\mathcal{K}),$$
$$B_p^{\Delta}(\mathcal{K}) = \operatorname{Im} \partial \colon C_{p+1}^{\Delta}(\mathcal{K}) \to C_p^{\Delta}(\mathcal{K}),$$

the groups of *simplicial cycles* and *simplicial boundaries*, respectively. The preceding lemma shows that $B_p^{\Delta}(\mathcal{K})$ is a subgroup of $Z_p^{\Delta}(\mathcal{K})$, so we may define the pth *simplicial homology group* of $\mathcal{K}$ to be the quotient

$$H_p^{\Delta}(\mathcal{K}) = Z_p^{\Delta}(\mathcal{K}) / B_p^{\Delta}(\mathcal{K}).$$

Because the simplicial chain groups of a finite complex are all finitely generated, simplicial homology can in principle be computed directly from the combinatorial structure of a complex. In practice this is not usually efficient, at least without a computer, because triangulations of even very simple spaces typically have a large number of simplices. We will see below that the simplicial homology groups are isomorphic to the singular ones. However, there is one case in which simplicial homology is not hard to compute directly.

Lemma 13.29. *Let $\mathcal{K}$ be a complex consisting of a single n-simplex and its faces. Then $H_0^{\Delta}(\mathcal{K})$ is the infinite cyclic group generated by the homology class of any vertex, and $H_p^{\Delta}(\mathcal{K})$ is trivial for $p > 0$.*

Proof. We assume that an ordering $(v_0, \dots, v_p)$ has been chosen for the vertices of $\mathcal{K}$. Define a homomorphism $h \colon C_p^{\Delta}(\mathcal{K}) \to C_{p+1}^{\Delta}(\mathcal{K})$ by setting, for any p-simplex $\tau = \langle v_{k_0}, \dots, v_{k_p} \rangle \in \mathcal{K}$,

$$h\tau = \begin{cases} \langle v_0, v_{k_0}, \dots, v_{k_p} \rangle & \text{if } k_0 \neq 0, \\ 0 & \text{if } k_0 = 0, \end{cases}$$

and extending h to a homomorphism.

When $p > 0$, a straightforward computation shows that $\partial \circ h + h \circ \partial = \operatorname{Id}$. Thus if c is any p-cycle, $c = \partial h c$, which shows that $H_p^{\Delta}(\mathcal{K}) = 0$.

For $p = 0$, define a homomorphism $\varepsilon \colon C_0^{\Delta}(\mathcal{K}) \to \mathbb{Z}$ by $\varepsilon(\sum_i n_i \langle v_i \rangle) = \sum_i n_i$ as in the proof of Proposition 13.5. Because $\varepsilon\langle v_0 \rangle = 1$, ε is surjective. Another computation shows that $\partial h c = c - \varepsilon(c)\langle v_0 \rangle$ for any 0-chain c, so any chain in $\operatorname{Ker} \varepsilon$ is a boundary. Conversely, any boundary is in $\operatorname{Ker} \varepsilon$ because $\varepsilon \partial \langle v_i, v_j \rangle = \varepsilon(\langle v_j \rangle - \langle v_i \rangle) = 0$ (assuming $i < j$). This shows that ε descends to an isomorphism from $H_0^{\Delta}(\mathcal{K})$ to $\mathbb{Z}$, so $H_0^{\Delta}(\mathcal{K})$ is the infinite cyclic group generated by the class of $\langle v_0 \rangle$. □

It should be noted that there are several alternative ways of defining simplicial homology groups. The one most commonly used is to define the simplicial chain group as the free abelian group on the set of *oriented* simplices, with the convention that $\sigma' = -\sigma$ if σ' is the same simplex as

σ with the opposite orientation. Another possible chain group is the free abelian group on all *ordered* simplices, considering different vertex orderings of the same simplex as distinct generators. Both of these definitions have the advantage that, unlike our definition, they do not depend on a choice of ordering of the vertices; this is important if one wishes to define homomorphisms induced by simplicial maps (which may not preserve the vertex ordering) and prove functorial properties such as topological and homotopy invariance. If our goal were to develop an entire theory of simplicial homology groups, we would have to use one of these definitions. But our aim is more modest: We wish only to show that simplicial homology gives an alternative way of computing the singular homology groups, so we use a definition that is technically somewhat simpler, and confine our attention to the properties needed for this purpose.

The main result we need is an analogue of the Mayer–Vietoris theorem for simplicial homology. Fortunately, its proof is much easier than in the singular case.

The setup for this theorem is slightly different from that of its singular cousin. In this case, instead of considering open subsets, we suppose $\mathcal{K}'$ and $\mathcal{K}''$ are subcomplexes of $\mathcal{K}$, and let $\mathcal{L} = \mathcal{K}' \cap \mathcal{K}''$ (which is also a subcomplex). As in the singular case, we have inclusion maps

$$\begin{tikzcd} & \mathcal{K}' \arrow[dr, "k"] & \\ \mathcal{L} \arrow[ur, "i"] \arrow[dr, "j"'] & & \mathcal{K}. \\ & \mathcal{K}'' \arrow[ur, "l"'] & \end{tikzcd}$$

Each of these induces an inclusion map on simplicial chains, which is a chain map, provided that we choose the vertex orderings in $\mathcal{K}'$, $\mathcal{K}''$, and $\mathcal{L}$ to be the restrictions of the ordering we chose for $\mathcal{K}$. Therefore, all four maps induce homology homomorphisms as well.

Theorem 13.30 (Simplicial Mayer–Vietoris Theorem). *Let $\mathcal{K}$ be a finite simplicial complex, with subcomplexes $\mathcal{K}', \mathcal{K}''$ whose union is $\mathcal{K}$, and let $\mathcal{L} = \mathcal{K}' \cap \mathcal{K}''$. For each p there is a connecting homomorphism $\partial_*\colon H_p^\Delta(\mathcal{K}) \to H_{p-1}^\Delta(\mathcal{L})$ such that the following sequence is exact:*

$$\cdots \xrightarrow{\partial_*} H_p^\Delta(\mathcal{L}) \xrightarrow{i_* \oplus j_*} H_p^\Delta(\mathcal{K}') \oplus H_p(\mathcal{K}'') \xrightarrow{k_* - l_*} H_p^\Delta(\mathcal{K})$$
$$\xrightarrow{\partial_*} H_{p-1}^\Delta(\mathcal{L}) \xrightarrow{i_* \oplus j_*} \cdots .$$

Proof. The sequence of chain maps

$$0 \to C_p^\Delta(\mathcal{L}) \xrightarrow{i_\# \oplus j_\#} C_p^\Delta(\mathcal{K}') \oplus C_p^\Delta(\mathcal{K}'') \xrightarrow{k_\# - l_\#} C_p^\Delta(\mathcal{K}) \to 0$$

is easily seen to be exact in this case. The existence of the connecting homomorphism and the exactness of the Mayer–Vietoris sequence then follow immediately from the zigzag lemma. □

To analyze the relationship between simplicial and singular homology, we define a map from the simplicial chain complex of $\mathcal{K}$ to the singular chain complex of its geometric realization as follows. For any p-simplex $\sigma = \langle v_{k_0}, \dots, v_{k_p} \rangle \in \mathcal{K}$, let $\alpha(\sigma)$ denote the affine singular p-simplex $\alpha(v_{k_0}, \dots, v_{k_p})$ in $|\mathcal{K}|$ (with the vertices in increasing order). This extends uniquely to a homomorphism $\alpha\colon C_p^{\Delta}(\mathcal{K}) \to C_p(|\mathcal{K}|)$. To see that it is a chain map, just compute

$$\begin{aligned}\partial\alpha(v_{k_0}, \dots, v_{k_p}) &= \sum_{i=0}^{p} (-1)^i \alpha(v_{k_0}, \dots, v_{k_p}) \circ F_{i,p} \\ &= \sum_{i=0}^{p} (-1)^i \alpha(v_{k_0}, \dots, \widehat{v}_{k_i}, \dots, v_{k_p}) \\ &= \alpha(\partial\langle v_{k_0}, \dots, v_{k_p} \rangle).\end{aligned}$$

Therefore, α induces a homology homomorphism $\alpha_*\colon H_p^{\Delta}(\mathcal{K}) \to H_p(|\mathcal{K}|)$.

Theorem 13.31. *For any finite complex $\mathcal{K}$, the map $\alpha_*\colon H_p^{\Delta}(\mathcal{K}) \to H_p(|\mathcal{K}|)$ is an isomorphism for all p.*

Proof. We prove the theorem by induction on the dimension of $\mathcal{K}$. If $\dim \mathcal{K} = 0$, then $\mathcal{K}$ is just a finite set of vertices. In this case, $C_0^{\Delta}(\mathcal{K})$ is the free abelian group on the set of vertices, and all the other simplicial chain groups are trivial. Therefore, all the boundary operators are zero, and $H_p^{\Delta}(\mathcal{K}) \cong C_p^{\Delta}(\mathcal{K})$, which is isomorphic to the corresponding singular group. The map $\alpha_*\colon H_0^{\Delta}(\mathcal{K}) \to H_0(|\mathcal{K}|)$ takes each generator $[\langle v \rangle]$ to a generator $[\alpha(v)]$, so the theorem is proved in this case.

Now suppose the theorem is true for complexes of dimension $n-1$, and let $\mathcal{K}$ have dimension n. We proceed by induction on the number of n-simplices in $\mathcal{K}$. When there are no n-simplices, $\mathcal{K}$ is $(n-1)$-dimensional, so the theorem is true in that case.

Suppose $\mathcal{K}'$ is the subcomplex of $\mathcal{K}$ obtained by deleting a single n-simplex σ. (It is a subcomplex because $\mathcal{K}$ has no simplices of dimension greater than n.) Let $\mathcal{K}''$ denote the subcomplex consisting of σ and all its faces, and $\mathcal{L} = \mathcal{K}' \cap \mathcal{K}''$. We will prove the inductive step by comparing the Mayer–Vietoris sequence of $\mathcal{K}$ (in simplicial homology) with that of $|\mathcal{K}|$ (in singular homology).

To set up the sequence in singular homology, let V be a neighborhood of $|\sigma|$ that admits a strong deformation retraction onto $|\sigma|$ (such a neighborhood exists by Problem 7-6), and let $U = |\mathcal{K}| \smallsetminus \{x\}$ for some point $x \in \operatorname{Int}|\sigma|$. Clearly, $|\mathcal{K}'| \subset U$, $|\mathcal{K}''| \subset V$, and $|\mathcal{L}| \subset U \cap V$. All of these inclusions are homotopy equivalences: Our choice of V guarantees that it admits a strong deformation retraction onto $|\mathcal{K}''|$, and it is easy to construct a strong deformation of U onto $|\mathcal{K}'|$ that deforms $\sigma \setminus \{x\}$ onto its boundary while leaving $|\mathcal{K}'|$ fixed. Gluing these maps together, we obtain a strong deformation retraction of $U \cap V$ onto $\mathcal{L}$.

Restricting $\alpha\colon C_p^\Delta(\mathcal{K}) \to C_p(|\mathcal{K}|)$ to the chain groups of the various subcomplexes yields the following diagram of chain maps, which is obviously commutative:

$$\begin{tikzcd}
0 \arrow[r] & C_*^\Delta(\mathcal{L}) \arrow[r, "i_\# \oplus j_\#"] \arrow[d, "\alpha"] & C_*^\Delta(\mathcal{K}') \oplus C_*^\Delta(\mathcal{K}'') \arrow[r, "k_\# - l_\#"] \arrow[d, "\alpha \oplus \alpha"] & C_*^\Delta(\mathcal{K}) \arrow[r] \arrow[d, "\alpha"] & 0 \\
0 \arrow[r] & C_*(U \cap V) \arrow[r, "i_\# \oplus j_\#"] & C_*(U) \oplus C_*(V) \arrow[r, "k_\# - l_\#"] & C_*^{\mathcal{U}}(|\mathcal{K}|) \arrow[r] & 0.
\end{tikzcd}$$

Therefore, using Proposition 13.16 we see that the following diagram commutes and has exact rows:

$$\begin{tikzcd}
H_p^\Delta(\mathcal{L}) \arrow[r] \arrow[d] & H_p^\Delta(\mathcal{K}') \oplus H_p^\Delta(\mathcal{K}'') \arrow[r] \arrow[d] & H_p^\Delta(\mathcal{K}) \arrow[r] \arrow[d] & {} \\
H_p(U \cap V) \arrow[r] & H_p(U) \oplus H_p(V) \arrow[r] & H_p^{\mathcal{U}}(|\mathcal{K}|) \arrow[r] & {}
\end{tikzcd}$$

$$\begin{tikzcd}
H_{p-1}^\Delta(\mathcal{L}) \arrow[r] \arrow[d] & H_{p-1}^\Delta(\mathcal{K}') \oplus H_{p-1}^\Delta(\mathcal{K}'') \arrow[d] \\
H_{p-1}(U \cap V) \arrow[r] & H_{p-1}(U) \oplus H_{p-1}(V).
\end{tikzcd}$$

With $U \cap V$, U, and V in these groups replaced by their homotopy equivalent spaces $|\mathcal{L}|$, $|\mathcal{K}'|$, and $|\mathcal{K}''|$, the diagram still commutes because the homotopy equivalences are all inclusion maps. Thus we finally arrive at a diagram in which all the vertical homomorphisms except the center one are isomorphisms. (For $\mathcal{K}'$ and $\mathcal{L}$ this follows from the inductive hypothesis, and for $\mathcal{K}''$ it follows from Lemma 13.29.) Therefore, by the five lemma, the middle arrow is also an isomorphism. Finally, replacing $H_p^{\mathcal{U}}$ by H_p, we obtain the result. □

Topological Invariance of the Euler Characteristic

Our most significant application of simplicial homology is the following theorem, which generalizes Corollary 10.15 and Problem 10-9.

Theorem 13.32. *The Euler characteristic of a finite simplicial complex $\mathcal{K}$ is given by the formula*

$$\chi(\mathcal{K}) = \sum_p (-1)^p \operatorname{rank} H_p(|\mathcal{K}|).$$

Therefore, the Euler characteristic is a topological invariant of $|\mathcal{K}|$.

Proof. Let $n = \dim \mathcal{K}$. Recall from Chapter 5 that the Euler characteristic of $\mathcal{K}$ is defined as

$$\chi(\mathcal{K}) = \sum_{p=0}^{n} (-1)^p c_p,$$

where c_p is the number of p-simplices in $\mathcal{K}$. Note that c_p is also the rank of the simplicial chain group $C_p^\Delta(\mathcal{K})$.

Consider the following short exact sequences:

$$0 \to B_p^\Delta(\mathcal{K}) \hookrightarrow Z_p^\Delta(\mathcal{K}) \to H_p^\Delta(\mathcal{K}) \to 0,$$
$$0 \to Z_p^\Delta(\mathcal{K}) \hookrightarrow C_p^\Delta(\mathcal{K}) \xrightarrow{\partial} B_{p-1}^\Delta(\mathcal{K}) \to 0.$$

Let us write

$$b_p = \operatorname{rank} B_p^\Delta(\mathcal{K}), \quad z_p = \operatorname{rank} Z_p^\Delta(\mathcal{K}), \quad h_p = \operatorname{rank} H_p^\Delta(\mathcal{K}).$$

By Proposition 9.16, we have the following equalities:

$$\begin{aligned} z_p &= h_p + b_p, \\ c_p &= z_p + b_{p-1}. \end{aligned}$$

Therefore,

$$\begin{aligned} \chi(\mathcal{K}) &= \sum_{p=0}^{n} (-1)^p c_p \\ &= \sum_{p=0}^{n} (-1)^p (z_p + b_{p-1}) \\ &= \sum_{p=0}^{n} (-1)^p (h_p + b_p + b_{p-1}). \end{aligned}$$

Because $b_{-1} = b_n = 0$, the b_p and b_{p-1} terms above form a telescoping sum adding to zero. Because $h_p = \operatorname{rank} H_p(|\mathcal{K}|)$ by Theorem 13.31, this completes the proof. □

For any topological space X, the integer $\beta_p(X) = \operatorname{rank} H_p(X)$ (if it is finite) is called the pth *Betti number* of X. We define the *Euler characteristic* of X by

$$\chi(X) = \sum_{p} (-1)^p \beta_p(X)$$

provided that each $\beta_p(X)$ is finite and $\beta_p(X) = 0$ for p sufficiently large. The preceding theorem then says that $\chi(\mathcal{K}) = \chi(|\mathcal{K}|)$ for a finite simplicial complex $\mathcal{K}$.

Cohomology

As Proposition 13.2 shows, the singular homology groups are covariant functors from the category of topological spaces to the category of abelian groups. For many applications, it turns out to be much more useful to have contravariant functors. We will not pursue any of these applications here, but content ourselves to note that one of the most important, the de Rham theory of differential forms, plays a central role in differential geometry.

To give you a view of what is to come, in this final section we introduce singular cohomology, which is essentially a contravariant version of singular homology. It does not give us any new information about topological spaces, but the information is organized in a different way, which is much more appropriate for some applications.

In Example 7.32 we observed that for any fixed abelian group G, there is a contravariant functor from the category of abelian groups to itself that sends each group X to the group $\operatorname{Hom}(X, G)$ of homomorphisms into G, and each homomorphism $f\colon X \to Y$ to the induced homomorphism $f^*\colon \operatorname{Hom}(Y, G) \to \operatorname{Hom}(X, G)$ given by $f^*(\varphi) = \varphi \circ f$. We apply this to the singular chain groups as follows. Given a topological space X and an abelian group G, for any integer $p \geq 0$ let $C^p(X; G)$ denote the group $\operatorname{Hom}(C_p(X), G)$. Elements of $C^p(X; G)$ are called p-dimensional *singular cochains* with coefficients in G (p-cochains for short).

The boundary operator $\partial\colon C_{p+1}(X) \to C_p(X)$ induces a homomorphism $\delta\colon C^p(X; G) \to C^{p+1}(X; G)$, called the *coboundary operator*, characterized by

$$(\delta\varphi)(c) = \varphi(\partial c).$$

It is immediate that $\delta \circ \delta = 0$, so we have a chain complex

$$\cdots \to C^{p-1}(X; G) \xrightarrow{\delta} C^p(X; G) \xrightarrow{\delta} C^{p+1}(X; G) \to \cdots .$$

(Actually, when the arrows go in the direction of increasing indices as in this case, it is customary to call it a *cochain complex.*) A p-cochain φ is called a *cocycle* if $\delta\varphi = 0$, and a *coboundary* if there exists $\psi \in C^{p-1}(X; G)$ such that $\delta\psi = \varphi$. The subgroups of $C^p(X; G)$ consisting of cocycles and coboundaries are denoted by $Z^p(X; G)$ and $B^p(X; G)$, respectively.

We define the pth *singular cohomology group* of X with coefficients in G to be the quotient

$$H^p(X; G) = Z^p(X; G)/B^p(X; G).$$

If $f\colon X \to Y$ is a continuous map, we obtain a map $f^{\#}\colon C^p(Y; G) \to C^p(X; G)$ (note the reversal of direction) by

$$(f^{\#}\varphi)(c) = \varphi(f_{\#}c).$$

This map commutes with the coboundary operators because

$$(f^{\#}\delta\varphi)(c) = \delta\varphi(f_{\#}c) = \varphi(\partial f_{\#}c) = \varphi(f_{\#}\partial c) = (f^{\#}\varphi)(\partial c) = (\delta f^{\#}\varphi)(c).$$

(A map that commutes with δ is called, predictably enough, a *cochain map.*) Therefore, $f^{\#}$ induces a cohomology homomorphism $f^*\colon H^p(Y;G) \to H^p(X;G)$ by $f^*[\varphi] = [f^{\#}\varphi]$.

Proposition 13.33. *The induced cohomology homomorphism satisfies the following properties.*

(*a*) *If $f\colon X \to Y$ and $g\colon Y \to Z$ are continuous, then $(g \circ f)^* = f^* \circ g^*$.*

(*b*) *The homomorphism induced by the identity map is the identity.*

Therefore, the assignment $X \mapsto H^p(X;G)$, $f \mapsto f^*$ defines a contravariant functor from the category of topological spaces to the category of abelian groups.

Corollary 13.34 (Topological Invariance of Cohomology). *If $f\colon X \to Y$ is a homeomorphism, then for any abelian group G and any integer $p \geq 0$, $f^*\colon H^p(Y;G) \to H^p(X;G)$ is an isomorphism.*

Exercise 13.4. Prove Proposition 13.33 and Corollary 13.34.

In a very specific sense, the singular cohomology groups express the same information as the homology groups, but in rearranged form. The precise statement is given by the *universal coefficient theorem*, which gives an exact sequence from which the cohomology groups with any coefficients can be computed from the singular homology groups. The statement and proof can be found in [Mun75] or [Spa89]. We will not go into the general case here, but we can easily handle one special case.

Let $\mathbb{F}$ be a field of characteristic zero, which just means that $\mathbb{F}$ is torsion free as an abelian group under addition. (In most applications $\mathbb{F}$ will be $\mathbb{R}$, $\mathbb{C}$, or $\mathbb{Q}$.) We can form the cohomology groups $H^p(X;\mathbb{F})$ as usual, just by regarding $\mathbb{F}$ as an abelian group; but in this case they have a bit more structure. The basic algebraic facts are expressed in the following lemma.

Lemma 13.35. *Let $\mathbb{F}$ be a field of characteristic zero.*

(*a*) *For any abelian group G, the set $\operatorname{Hom}(G,\mathbb{F})$ of group homomorphisms from G to $\mathbb{F}$ is a vector space over $\mathbb{F}$ with scalar multiplication defined pointwise: $(a\varphi)(g) = a(\varphi(g))$ for $a \in \mathbb{F}$.*

(*b*) *If $f\colon G_1 \to G_2$ is a group homomorphism, then the induced homomorphism $f^*\colon \operatorname{Hom}(G_2,\mathbb{F}) \to \operatorname{Hom}(G_1,\mathbb{F})$ is a linear transformation of vector spaces.*

(c) *If G is finitely generated, the dimension of* $\operatorname{Hom}(G,\mathbb{F})$ *is equal to the rank of G.*

Proof. The proofs of (a) and (b) are straightforward (and hold for any field, not just one of characteristic zero), and are left as an exercise. For (c), we proceed as follows. First suppose G is free abelian of rank n, and let $g_1,\dots,g_n$ be a basis for G (as an abelian group). For each i, define a homomorphism $\varphi_i\colon G\to\mathbb{F}$ by setting

$$\varphi_i(g_j)=\begin{cases}1 & \text{if } i=j,\\ 0 & \text{if } i\neq j.\end{cases}$$

If $\sum_i a_i\varphi_i$ is the zero homomorphism for some scalars $a_i\in\mathbb{F}$, applying this homomorphism to g_j shows that $a_j=0$, so the φ_i's are linearly independent. On the other hand, it is easy to see that an arbitrary $\varphi\in\operatorname{Hom}(G,\mathbb{F})$ can be written $\varphi=\sum_i a_i\varphi_i$ with $a_i=\varphi(g_i)$; thus the φ_i's are a basis for $\operatorname{Hom}(G,\mathbb{F})$, proving the result in this case.

In the general case, let $G_{\text{tor}}\subset G$ be the torsion subgroup of G. The surjective homomorphism $\pi\colon G\to G/G_{\text{tor}}$ induces a homomorphism $\pi^*\colon \operatorname{Hom}(G/G_{\text{tor}},\mathbb{F})\to\operatorname{Hom}(G,\mathbb{F})$. It follows easily from the surjectivity of π that π^* is injective. On the other hand, let $\varphi\in\operatorname{Hom}(G,\mathbb{F})$ be arbitrary. If $g\in G$ satisfies $kg=0$, then $\varphi(g)=\varphi(kg)/k=0$, so $G_{\text{tor}}\subset\operatorname{Ker}\varphi$ and φ descends to a homomorphism $\widetilde{\varphi}\in\operatorname{Hom}(G/G_{\text{tor}},\mathbb{F})$. Clearly, $\pi^*\widetilde{\varphi}=\varphi$, so π^* is an isomorphism. Because G/G_{tor} is free abelian, we have $\dim\operatorname{Hom}(G,\mathbb{F})=\dim\operatorname{Hom}(G/G_{\text{tor}},\mathbb{F})=\operatorname{rank}(G/G_{\text{tor}})=\operatorname{rank}G$. □

Exercise 13.5. Prove parts (a) and (b) of Lemma 13.35.

Applying this to $C^p(X;\mathbb{F})=\operatorname{Hom}(C_p(X),\mathbb{F})$, we see that the cochain groups are $\mathbb{F}$-vector spaces and the coboundary operators are linear maps. It follows that $Z^p(X;\mathbb{F})$ and $B^p(X;\mathbb{F})$ are vector spaces as is the quotient $H^p(X;\mathbb{F})=Z^p(X;\mathbb{F})/B^p(X;\mathbb{F})$. Moreover, for any continuous map $f\colon X\to Y$, the induced cohomology map $f^*\colon H^p(Y;\mathbb{F})\to H^p(X;\mathbb{F})$ is also a linear map.

The special feature of field coefficients that makes the cohomology groups easier to calculate is expressed in the following lemma.

Lemma 13.36 (Extension Lemma). *Let $\mathbb{F}$ be a field of characteristic zero. If G is an abelian group, any group homomorphism from a subgroup of G to $\mathbb{F}$ admits an extension to all of G.*

Proof. Suppose $H\subset G$ is a subgroup and $f\colon H\to\mathbb{F}$ is a homomorphism. Consider the set $\mathcal{F}$ of all pairs (H',f'), where H' is a subgroup of G containing H and $f'\colon H'\to\mathbb{F}$ is an extension of f. Define a partial ordering on $\mathcal{F}$ by declaring $(H',f')\le(H'',f'')$ if $H'\subset H''$ and $f''|_{H'}=f'$. Given any totally ordered subset $\mathcal{T}\subset\mathcal{F}$, define $\widetilde{H}$ to be the union of all the spaces H' such that $(H', f') \in \mathcal{T}$. There is a uniquely defined homomorphism $\widetilde{f}\colon \widetilde{H} \to \mathbb{F}$, defined by setting $\widetilde{f}(h) = f'(h)$ for any pair $(H', f') \in \mathcal{T}$ such that $h \in H'$. The pair $(\widetilde{H}, \widetilde{f})$ is easily seen to be an upper bound for $\mathcal{T}$. Thus by Zorn's lemma (Lemma A.3 in the Appendix), there exists a maximal element in $\mathcal{F}$; call it (H_0, f_0).

If $H_0 = G$, we are done. If not, we will show that f_0 can be extended to a larger subgroup containing H_0, which contradicts the maximality of H_0.

Suppose there is some element $g \in G \smallsetminus H_0$. Let H_g denote the subgroup

$$H_g = \{h + mg : h \in H_0, m \in \mathbb{Z}\}.$$

The quotient group H_g/H_0 is cyclic and generated by the coset of g. There are two cases.

If H_g/H_0 is infinite, then no multiple of g is in H_0, so every element of H_g can be written *uniquely* in the form $h + mg$ and we can define an extension f_0' of f_0 just by setting $f_0'(h+mg) = f_0(h)$. On the other hand, if H_g/H_0 is finite, let n be the order of this group. This means that $mg \in H_0$ if and only if m is a multiple of n. Let $k = f_0(ng)/n \in \mathbb{F}$, and define an extension f_0' of f_0 by letting

$$f_0'(h + mg) = f_0(h) + mk.$$

To show that this is well-defined, suppose $h + mg = h' + m'g$ for $h, h' \in H_0$ and $m, m' \in \mathbb{Z}$. Then $(m - m')g = h' - h \in H_0$, which implies $m - m' = jn$ for some integer j. We compute

$$\begin{aligned}(f_0(h) + mk) - (f_0(h') + m'k) &= f_0(h - h') + (m - m')k \\ &= f_0(-jng) + jnk = 0.\end{aligned}$$

Therefore, f_0' is an extension of f_0, which completes the proof. □

Now we come to the main result of this section, which gives explicit formulas for singular cohomology with coefficients in $\mathbb{F}$.

Theorem 13.37. *Let $\mathbb{F}$ be a field of characteristic zero. For any topological space X, the vector spaces $H^p(X;\mathbb{F})$ and $\operatorname{Hom}(H_p(X), \mathbb{F})$ are isomorphic; hence if $H_p(X)$ is finitely generated, then the dimension of $H^p(X;\mathbb{F})$ is equal to the rank of $H_p(X)$.*

Proof. An arbitrary cocycle $\varphi \in Z^p(X;\mathbb{F})$ defines a homomorphism $\widetilde{\varphi}\colon H_p(X) \to \mathbb{F}$ by

$$\widetilde{\varphi}[c] = \varphi(c).$$

Since $\varphi(\partial b) = \delta\varphi(b) = 0$, this is well-defined independently of the choice of representative c in its homology class. If $\varphi = \delta\eta$ is a coboundary, then $\widetilde{\varphi}[c] = \varphi(c) = \delta\eta(c) = \eta(\partial c) = 0$, so the homomorphism $\varphi \mapsto \widetilde{\varphi}$ contains the coboundary group $B^p(X;\mathbb{F})$ in its kernel. It therefore descends to a homomorphism $\beta\colon H^p(X;\mathbb{F}) \to \operatorname{Hom}(H_p(X),\mathbb{F})$, given by $\beta[\varphi] = \widetilde{\varphi}$. We will show that β is an isomorphism.

Let $f \in \operatorname{Hom}(H_p(X),\mathbb{F})$ be arbitrary. Letting $\pi\colon Z_p(X) \to H_p(X)$ denote the projection defining $H_p(X)$, we obtain a homomorphism $f \circ \pi\colon Z_p(X) \to \mathbb{F}$. By the extension lemma, this extends to a homomorphism $\varphi\colon C_p(X) \to \mathbb{F}$, i.e., a p-cochain. In fact, φ is a coboundary, because

$$(\delta\varphi)c = \varphi(\partial c) = f \circ \pi(\partial c) = f[\partial c] = 0.$$

Unwinding the definitions, we see that $f = \beta[\varphi]$, so β is surjective.

To show that it is injective, suppose $\beta[\varphi] = 0$. This means that $\varphi \in C^p(X;\mathbb{F})$ satisfies $\varphi(c) = 0$ for all cycles c, so $Z_p(X) \subset \operatorname{Ker}\varphi$. Therefore, φ descends to a homomorphism $\widetilde{\varphi}\colon C_p(X)/Z_p(X) \to \mathbb{F}$.

On the other hand, the surjective homomorphism $\partial\colon C_p(X) \to B_{p-1}(X)$ has kernel equal to $Z_p(X)$, and therefore induces an isomorphism $\widetilde{\partial}\colon C_p(X)/Z_p(X) \to B_{p-1}(X)$. Composition gives a homomorphism $\widetilde{\varphi} \circ \widetilde{\partial}^{-1}\colon B_{p-1}(X) \to \mathbb{F}$:

$$B_{p-1}(X) \xrightarrow{\widetilde{\partial}^{-1}} C_p(X)/Z_p(X) \xrightarrow{\widetilde{\varphi}} \mathbb{F}.$$

By the extension lemma, this extends to a homomorphism $\eta\colon C_{p-1}(X) \to \mathbb{F}$. If $c \in C_p(X)$ is arbitrary,

$$\eta(\partial c) = (\widetilde{\varphi} \circ \widetilde{\partial}^{-1})(\partial c) = \varphi(c),$$

which shows that $\varphi = \delta\eta$, and so $[\varphi] = 0$. Thus β is injective, completing the proof. $\square$

As a consequence of this theorem, the Euler characteristic of a space can also be computed in terms of its cohomology. The following corollary follows immediately from the theorem.

Corollary 13.38. *If X is a topological space such that $H_p(X)$ is finitely generated for all p and zero for p sufficiently large, then for any field $\mathbb{F}$ of characteristic zero,*

$$\chi(X) = \sum_p (-1)^p \dim H^p(X;\mathbb{F}).$$

Problems

13-1. Let $\mathbb{P}^n$ be the real projective space of dimension n.

(a) Show that $\mathbb{P}^n$ is homeomorphic to the quotient of $\overline{\mathbb{B}^n}$ by the relation that identifies antipodal points on the boundary sphere.

(b) Use (a) and the results of this chapter to compute the singular homology groups of $\mathbb{P}^2$ and $\mathbb{P}^3$.

13-2. Let $n \geq 1$. If $f\colon \mathbb{S}^n \to \mathbb{S}^n$ is a continuous map that has a continuous extension to a map $F\colon \overline{\mathbb{B}^{n+1}} \to \mathbb{S}^n$, show that f has degree zero.

13-3. Show that $\mathbb{S}^n$ is not a retract of $\overline{\mathbb{B}^{n+1}}$ for any n.

13-4. Prove the *Brouwer fixed point theorem*: Any continuous map $f\colon \overline{\mathbb{B}^n} \to \overline{\mathbb{B}^n}$ has a fixed point. [See Problem 8-9.]

13-5. Show that the dimension of a finite-dimensional simplicial complex $\mathcal{K}$ is a topological invariant of $|\mathcal{K}|$, and that any triangulation of an n-manifold has dimension n. [Be careful: We are not assuming that the complexes are finite.]

13-6. Prove that the singular homology groups of any compact polyhedron are finitely generated.

13-7. If M is a triangulable compact manifold, show that $H_p(M) = 0$ if $p > \dim M$.

13-8. An n-dimensional *pseudomanifold* is an n-dimensional simplicial complex in which every simplex is a face of some n-simplex, every $(n-1)$-simplex is a face of exactly two n-simplices, and for every pair of n-simplices σ, σ' there exists a finite sequence of n-simplices $\sigma = \sigma_1, \dots, \sigma_k = \sigma'$ such that σ_i and σ_{i+1} have an $(n-1)$-dimensional face in common. Show that the nth (singular or simplicial) homology group of an n-dimensional pseudomanifold is infinite cyclic if it is orientable and trivial if not. [It can be shown (see, e.g., [Mun75]) that every triangulated, connected, compact manifold is a pseudomanifold, and then this result characterizes the nth homology of triangulable compact n-manifolds. But this requires more machinery than we have developed.]

13-9. Suppose M is an n-manifold with boundary. Show that the set of boundary points and the set of interior points of M are disjoint.

13-10. Let X_1 and X_2 be spaces with nondegenerate base points q_1 and q_2. Show that $H_p(X_1 \vee X_2) \cong H_p(X_1) \oplus H_p(X_2)$ for all $p > 0$. [Hint: For $p = 1$, use Problem 10-15.]

13-11. A (covariant or contravariant) functor from the category of abelian groups to itself is said to be *exact* if it takes exact sequences to exact sequences. If $\mathbb{F}$ is a field of characteristic zero, show that the functor $G \mapsto \operatorname{Hom}(G, \mathbb{F})$, $f \mapsto f^*$ is exact.

13-12. If U and V are open subsets of the topological space X, prove that there is an exact Mayer–Vietoris sequence for cohomology with coefficients in a field $\mathbb{F}$ of characteristic zero:

$$\cdots \to H^{p-1}(U \cap V; \mathbb{F}) \to H^p(X; \mathbb{F}) \to H^p(U; \mathbb{F}) \oplus H^p(V; \mathbb{F}) \to H^p(U \cap V; \mathbb{F}) \to \cdots .$$

[Hint: Use Problem 13-11.]

13-13. An abelian group K is said to be *divisible* if for any $k \in K$ and nonzero $n \in \mathbb{Z}$, there exists $k' \in K$ such that $nk' = k$. It is said to be *injective* if for every group G, any homomorphism from a subgroup of G into K extends to all of G. Show that an abelian group K is injective if and only if it is divisible if and only if the functor $G \mapsto \operatorname{Hom}(G, K)$ is exact.

Appendix
Review of Prerequisites

The most important prerequisite for studying this book is a thorough grounding in advanced calculus. Since there are hundreds of books that treat this subject well, we will simply assume familiarity with it, and remind the reader of important facts when necessary. We also assume that the reader is familiar with the terminology and rules of ordinary logic.

The other prerequisites are a solid understanding of the basic properties of sets, metric spaces, and groups, at the level that you would find in most undergraduate courses in real analysis and abstract algebra.

In this appendix we briefly review some fundamental aspects of these three subjects. If you have not studied this material before, you cannot hope to learn it from scratch here. But this appendix can serve as a reminder of important concepts that you may have forgotten, as a way to standardize our notation and terminology, and as a source of references to books where you can look up more of the details to refresh your memory. You can use the exercises to test your knowledge, or to brush up on any aspects of the subject on which you feel your knowledge is shaky.

Set Theory

In this book, as in most modern mathematics, mathematical statements are couched in the language of set theory. We give here a brief descriptive summary of the parts of set theory that we will use, in the form that is commonly called "naive set theory." The word naive should be understood in the same sense in which it is used by Paul Halmos in his classic text *Naive Set Theory* [Hal74]: The axioms of set theory are to be viewed much as Euclid viewed his geometric axioms, as intuitively clear statements of fact from which reliable conclusions can be drawn.

One must be a bit careful with the axioms, to be sure, because it is possible to get into trouble by trying to construct sets too freely, as is illustrated by the famous paradox of Bertrand Russell described below. It is primarily for this reason that we take the trouble to enumerate the axioms at all. For more detail on the subject, in the same spirit as the treatment here, consult [Hal74] or [Dev93]. We leave it to the set theorists to explore the deep consequences of the axioms and the relationships among different axiom systems.

Basic Concepts

The word *set* is, mathematically, an undefined term. A set should be thought of as an assemblage of "mathematical objects," whatever they may be—things such as numbers, ordered pairs, functions, or other sets. The properties of sets, and the rules for manipulating them, are expressed in the axioms we list below. We sometimes use the words *collection* and *family* as synonyms for set.

The fundamental relationship involving sets, which we also leave mathematically undefined, is that of *membership*. Intuitively, if x is one of the objects in the set S, then we say that x is a *member* or an *element* of S, or x *belongs to* S, written $x \in S$. The essential characteristic of sets is that they are determined by their members. Formally, we define $S = T$ to mean $x \in S \iff x \in T$.

The set containing no elements is called the *empty set* and denoted by $\varnothing$. It is unique, because any two sets with no elements are equal by our definition of set equality, so we are justified in calling it "the" empty set. (We could postulate its existence as a separate axiom, but its existence will follow from our other axioms, as you will see below.) If S and T are sets such that every element of S is also an element of T, then S is a *subset* of T, written $S \subset T$. It is a *proper subset* if $S \subset T$ but $S \neq T$. The notation $T \supset S$ means $S \subset T$. Clearly, $S = T$ if and only if $S \subset T$ and $T \subset S$.

The axioms for sets describe precisely what sets can be asserted to exist, and what properties they have. Here is the first one.

- Specification axiom: Given a set S and a sentence $P(x)$ that is either true or false whenever x is any particular element of S, there is a set consisting of all those $x \in S$ for which $P(x)$ is true, denoted by $\{x \in S : P(x)\}$.

Note that one must start with a specific set before the specification axiom can be used. This requirement rules out forming sets out of self-contradictory specifications such as the one discovered by Bertrand Russell and now known as "Russell's paradox": The sentence $\mathcal{C} = \{X : X \notin X\}$ looks as if it might define a set, but it does not, because each statement $\mathcal{C} \in \mathcal{C}$ and $\mathcal{C} \notin \mathcal{C}$ implies its own negation. Similarly, the specification axiom implies that there does not exist a "set of all sets," for if there were such a set $\mathcal{S}$, we could use the specification axiom to define $\mathcal{C} = \{S \in \mathcal{S} : S \notin S\}$ and reach the same contradiction.

Still, there are times when we will need to speak of "all sets" or other similar aggregations, primarily in the context of category theory (see Chapter 7). For this purpose, we reserve the word *class* to refer to an aggregate of mathematical objects that may or may not constitute a set.

- POWER SET AXIOM: Given any set S, there is a set $\mathcal{P}(S)$, called the *power set* of S, whose elements are exactly the subsets of S.
- UNION AXIOM: Given any collection $\mathcal{C}$ of sets, there is a set called their *union* and denoted by $\bigcup \mathcal{C}$, with the property that $x \in \bigcup \mathcal{C}$ if and only if $x \in S$ for some $S \in \mathcal{C}$.

Given any nonempty collection $\mathcal{C}$ of sets, their *intersection*, denoted by $\bigcap \mathcal{C}$, is defined as the set

$$\bigcap \mathcal{C} = \{x \in \bigcup \mathcal{C} : x \in S \text{ for every } S \in \mathcal{C}\}.$$

Other notations for unions and intersections are

$$\bigcup_{S \in \mathcal{C}} S; \qquad S_1 \cup S_2 \cup \cdots ;$$

$$\bigcap_{S \in \mathcal{C}} S; \qquad S_1 \cap S_2 \cap \cdots .$$

Given any collection $\mathcal{C}$ of sets, if $A \cap B = \varnothing$ whenever $A, B \in \mathcal{C}$ and $A \neq B$, the sets in $\mathcal{C}$ are said to be *disjoint.*

If A and B are any sets, their *set difference* is defined to be the set

$$A \smallsetminus B = \{x \in A : x \notin B\},$$

which exists by the specification axiom. If $B \subset A$, the set difference $A \smallsetminus B$ is also called the *complement* of B in A.

When sets are defined by specification, it is common to abbreviate the notation in certain circumstances if it can be done unambiguously. For example, if the elements of a set can be named explicitly, the set is commonly specified simply by listing its elements, as in $\{a_1, a_2, \dots, a_k\}$. As long as each of the elements a_i is an element of some other set S_i, this is a legitimate use of our axioms and can be interpreted as $\{x \in S_1 \cup \cdots \cup S_k : x = a_1 \text{ or } x = a_2 \text{ or } \dots \text{ or } x = a_k\}$. Since the resulting set is the same regardless of what sets S_i the a_i's originally came from, there is no need to include them in the notation. A set $\{a\}$ with a unique element a is called a *singleton.*

Cartesian Products, Relations, and Functions

Another primitive concept that we will use without a formal definition is that of an *ordered pair*. Think of it as a pair of objects in a specific order, indicated by writing them in parentheses and separated by a comma, as in (a, b). The objects a and b are called the *components* of the ordered pair. The defining characteristic is that two ordered pairs are equal if and only if their first components are equal and their second components are equal: $(a, b) = (a', b') \iff a = a'$ and $b = b'$.

- CARTESIAN PRODUCT AXIOM: Given sets A and B, there exists a set $A \times B$, called their *Cartesian product*, whose members are precisely the ordered pairs (a, b) for every $a \in A$ and $b \in B$.

With these axioms we can define the most important constructions in mathematics: relations and functions. A *relation* between sets X and Y is a subset of $X \times Y$. If r is a relation, it is often convenient to use some notation such as $x \; ⓡ \; y$ to mean $(x, y) \in r$. For example, both "equals" and "less than" are relations in $\mathbb{R} \times \mathbb{R}$.

An important special case arises when we consider relations between a set S and itself, which we usually call a relation "on S." Let $\sim$ denote such a relation. It is said to be *reflexive* if $x \sim x$ for all $x \in S$, *symmetric* if $x \sim y$ implies $y \sim x$, and *transitive* if $x \sim y$ and $y \sim z$ imply $x \sim z$. A relation that is reflexive, symmetric, and transitive is called an *equivalence relation*. Given an equivalence relation $\sim$, for each $x \in S$ the *equivalence class* of x is defined to be the set

$$[x] = \{y \in S : y \sim x\}.$$

The set of equivalence classes is denoted by $S/{\sim}$.

Closely related to equivalence relations is the following notion: A *partition* of a set S is a collection $\mathcal{C}$ of disjoint nonempty subsets of S whose union is S. In this situation one also says that S is the *disjoint union* of the sets in $\mathcal{C}$.

Exercise A.1. Given an equivalence relation $\sim$ on a set S, show that the set $S/{\sim}$ of equivalence classes is a partition of S. Conversely, given a partition of S, show that there is a unique equivalence relation whose set of equivalence classes is exactly the original partition.

If r is any relation on a set S, the next exercise shows that there is a "smallest" equivalence relation $\sim$ such that $x \; ⓡ \; y \implies x \sim y$. It is called the *equivalence relation generated by* r.

Exercise A.2. Let $r \subset X \times X$ be any relation, and define $\sim$ to be the intersection of all equivalence relations in $X \times X$ that contain r.

(a) Show that $\sim$ is an equivalence relation.

(b) Show that $x \sim y$ if and only if one of the following is true: $x = y$, $x \circledR' y$, or there is a finite sequence of elements $z_1, \dots, z_n \in X$ such that $x \circledR' z_1 \circledR' \cdots \circledR' z_n \circledR' y$, where $x \circledR' y$ means "$x \circledR y$ or $y \circledR x$." (See below for the formal definition of a finite sequence.)

Another particularly important type of relation is a *partial ordering*: This is a relation $\le$ on a set X that is reflexive, transitive, and *antisymmetric*, which means that $x \le y$ and $y \le x$ together imply $x = y$. If in addition every pair $x, y \in X$ satisfy either $x \le y$ or $y \le x$, it is called a *total ordering* (or sometimes a *linear* or *simple ordering*). The notation $x < y$ is defined to mean $x \le y$ and $x \ne y$, and the notations $x > y$ and $x \ge y$ have the obvious meanings. If X is a set endowed with an ordering, one often says that X is a (totally or partially) ordered set, with the ordering being understood from the context.

The most common examples of totally ordered sets are number systems such as the real numbers or the integers (which we will introduce formally below). An important example of a partially ordered set is the set $\mathcal{P}(S)$ of subsets of a given set S, with the partial order relation defined by inclusion: $X \le Y$ if and only if $X \subset Y$. It is easy to see that any subset of a partially ordered set is itself partially ordered with (the restriction of) the same order relation, and if the original ordering is total, then the subset is also totally ordered.

If X is a partially ordered set and $S \subset X$ is any subset, an element $x \in X$ is said to be an *upper bound* for S if $x \ge s$ for every $s \in S$. If S has an upper bound, it is said to be *bounded above*. The terms *lower bound* and *bounded below* are defined similarly. An element $s \in S$ is said to be *maximal* if there is no $s' \in S$ such that $s' > s$, and it is the *largest element* of S if $s' \le s$ for every $s' \in S$. Minimal and smallest elements are defined similarly. Clearly, the largest element of S, if it exists, is unique and maximal. If S is totally ordered, a maximal element is automatically largest; but in a partially ordered set this may not be the case, because there may be elements that are neither larger nor smaller than s. A totally ordered set X is said to be *well-ordered* if every nonempty subset $S \subset X$ has a smallest element. For example, the set of natural numbers is well-ordered, but the integers and the real numbers are not.

A *function* from X to Y is a relation $f \subset X \times Y$ with the property that for every $x \in X$ there is a unique $y \in Y$ such that $(x, y) \in f$. This unique element of Y is denoted by $f(x)$. The sets X and Y are called the *domain* and *range* of f, respectively. The words *map* and *mapping* are used synonymously for function.

The notation $f\colon X \to Y$ means "f is a function from X to Y" (or, depending on how it is used in a sentence, "f, a function from X to Y," or "f, from X to Y"). The equation $y = f(x)$ is also sometimes written $f\colon x \mapsto y$ or, if the name of the function is not important, $x \mapsto y$. Note that the type of arrow ($\mapsto$) used to denote the action of a function on an element of its domain is different from the arrow ($\to$) used between the domain and range.

Given two functions $f\colon X \to Y$ and $g\colon Y \to Z$, their *composition* is the function $g \circ f\colon X \to Z$ defined by $(g \circ f)(x) = g(f(x))$.

For every set X, there exists a natural function $\mathrm{Id}_X\colon X \to X$ called the *identity map* of X, defined by $f(x) = x$ for all $x \in X$. If $S \subset X$ is a subset, there is a function $\iota_S\colon S \to X$ called the *inclusion map* of S, given by $\iota_S(x) = x$ for $x \in S$. We sometimes use the notation $\iota_S\colon S \hookrightarrow X$ to emphasize the fact that it is an inclusion map. If $f\colon X \to Y$ and S is a subset of X, there is a function $f|_S\colon S \to Y$ called the *restriction* of f to S, obtained by applying f only to elements of S. In terms of ordered pairs, $f|_S$ is just the subset of $S \times Y$ consisting of ordered pairs $(x, y) \in f$ such that $x \in S$. It is immediate that $f|_S = f \circ \iota_S$, and ι_S is just the restriction of Id_X to S. If $g\colon S \to Y$ is a map and $f\colon X \to Y$ is a map whose restriction to S is equal to g, we say that f is an *extension* of g.

Let $f\colon X \to Y$ be a function. If $S \subset X$, the *image of S under f* is the set

$$f(S) = \{y \in Y : y = f(x) \text{ for some } x \in S\}.$$

The set $f(X) \subset Y$, the image of the entire domain X, is just called the *image of f*. (Warning: In analysis it is common to use the word "range" to denote what we call the image of a function, and the word "codomain" to denote what we call its range.) If B is a subset of Y, the *inverse image* of B, denoted by $f^{-1}(B)$, is the set

$$f^{-1}(B) = \{x \in X : f(x) \in B\}.$$

If $B = \{b\}$ is a singleton, it is common to use the notation $f^{-1}(b)$ in place of the more accurate but more cumbersome $f^{-1}(\{b\})$.

Exercise A.3. Let $f\colon X \to Y$ be a map.

(a) If $A \subset B \subset Y$, then $f^{-1}(A) \subset f^{-1}(B)$.

(b) If $B \subset Y$, then $f^{-1}(Y \smallsetminus B) = X \smallsetminus f^{-1}(B)$.

(c) Give a counterexample to show that it is not generally true that $f(X \smallsetminus A) = Y \smallsetminus f(A)$ whenever $A \subset X$.

The function f is said to be *injective* or *one-to-one* if $f(x) = f(y)$ implies $x = y$. It is said to be *surjective* or to map X *onto* Y if $f(X) = Y$, or in other words if every $y \in Y$ is equal to $f(x)$ for some $x \in X$. A function that is both injective and surjective is said to be *bijective* or a *one-to-one correspondence*. A bijective map from a set X to itself is also called a *permutation* of X.

Given $f\colon X \to Y$, if there exists a map $g\colon Y \to X$ such that $f \circ g = \mathrm{Id}_Y$ and $g \circ f = \mathrm{Id}_X$, then g is said to be an *inverse* for f. Since inverses are unique (see the next exercise), the inverse map is denoted unambiguously by f^{-1} when it exists. More generally, if g satisfies only $g \circ f = \mathrm{Id}_X$, it is called a *left inverse* for f, and if $f \circ g = \mathrm{Id}_Y$, g is a *right inverse* for f.

Lemma A.1. *If $f\colon X \to Y$ is a function and $X \neq \varnothing$, then f has a left inverse if and only if it is injective, and a right inverse if and only if it is surjective.*

Proof. Suppose g is a left inverse for f. If $f(x) = f(x')$, applying g to both sides implies $x = x'$, so f is injective. Similarly, if g is a right inverse and $y \in Y$ is arbitrary, then $f(g(y)) = y$, so f is surjective.

Now suppose f is injective. Choose any $x_0 \in X$, and define $g\colon Y \to X$ by $g(y) = x$ if $y \in f(X)$ and $y = f(x)$, and $g(y) = x_0$ if $y \notin f(X)$. It is immediate that $g \circ f = \mathrm{Id}_X$. The proof that surjectivity implies the existence of a right inverse requires the axiom of choice, so we postpone it until the end of the section (Exercise A.8). □

Exercise A.4. Let f be a function.

(a) Show that f has an inverse if and only if it is bijective.

(b) If f has an inverse, show that it is unique.

Number Systems and Cardinality

So far, all the set-theoretic axioms we have introduced describe ways of obtaining new sets from already existing ones. Before the theory will have much content, we need to know that some sets exist. We take the set of real numbers as our starting point. The properties that characterize it are that it is an *ordered field* (a field in the algebraic sense, endowed with a total ordering in which $y < z \implies x + y < x + z$ and $x > 0,\ y > 0 \implies xy > 0$) that is *complete* (every nonempty subset with an upper bound has a least upper bound).

- EXISTENCE AXIOM: There exists a complete ordered field $\mathbb{R}$, called the *real numbers*.

Because this axiom guarantees the existence of at least one set, we now can assert the existence of the empty set, since $\{x \in \mathbb{R} : x \neq x\} = \varnothing$.

Exercise A.5. Show that the real numbers are unique, in the sense that any complete ordered field admits an order-preserving isomorphism with $\mathbb{R}$.

Let $S \subset \mathbb{R}$ be a nonempty subset with an upper bound. The least upper bound of S is also called the *supremum* of S, and is denoted by $\sup S$. Similarly, any nonempty set T with a lower bound has a greatest lower bound, also called its *infimum* and denoted by $\inf T$.

We will work extensively with the usual subsets of $\mathbb{R}$:

- The set of *natural numbers* $\mathbb{N}$ (the positive counting numbers), defined as the smallest subset of $\mathbb{R}$ containing 1 and containing $n+1$ whenever it contains n.

- The set of *integers* $\mathbb{Z} = \{n \in \mathbb{R} : n = 0 \text{ or } n \in \mathbb{N} \text{ or } -n \in \mathbb{N}\}$.

- The set of *rational numbers* $\mathbb{Q} = \{x \in \mathbb{R} : x = p/q \text{ for some } p, q \in \mathbb{Z}\}$.

We consider the set $\mathbb{C}$ of complex numbers to be simply $\mathbb{R} \times \mathbb{R}$, in which the real numbers are identified with the subset $\mathbb{R} \times \{0\} \subset \mathbb{C}$ and i stands for the imaginary unit $(0, 1)$. Multiplication and addition of complex numbers are defined by the usual rules with $i^2 = -1$; thus $x + iy$ is another notation for (x, y).

For any pair of integers $m \le n$, the notation $\{m, \dots, n\}$ means $\{k \in \mathbb{Z} : m \le k \le n\}$. For subsets of the real numbers, we use the following notation:

$$
\begin{aligned}
[a, b] &= \{x \in \mathbb{R} : a \le x \le b\},\\
(a, b) &= \{x \in \mathbb{R} : a < x < b\},\\
[a, b) &= \{x \in \mathbb{R} : a \le x < b\},\\
(a, b] &= \{x \in \mathbb{R} : a < x \le b\}.
\end{aligned}
$$

We also allow a or b or both to be replaced by either of the symbols ∞ or $-\infty$ in any of the above definitions in which it makes sense, with the obvious meanings. A nonempty subset $J \subset \mathbb{R}$ is called an *interval* if whenever $a, b \in J$, every c such that $a < c < b$ is also in J.

Exercise A.6. Show that an interval must be one of the nine types of sets $[a, b]$, (a, b), $[a, b)$, $(a, b]$, $(-\infty, b]$, $(-\infty, b)$, $[a, \infty)$, (a, ∞), or $(-\infty, \infty)$.

The natural numbers play a special role in set theory, as a yardstick for measuring sizes of sets. Two sets are said to have the same *cardinality* if there exists a one-to-one correspondence between them. A set is *finite* if it is empty or has the same cardinality as $\{1, \dots, n\}$ for some $n \in \mathbb{N}$ (in which case it is said to have cardinality n), and otherwise it is *infinite.* A set is *countably infinite* if it has the same cardinality as $\mathbb{N}$, *countable* if it is either finite or countably infinite, and *uncountable* otherwise. The sets $\mathbb{N}$, $\mathbb{Z}$, and $\mathbb{Q}$ are countable, but $\mathbb{R}$ and $\mathbb{C}$ are not.

Exercise A.7.

(a) Prove that the union of a countable collection of countable sets is countable.

(b) Prove that any subset of a countable set is countable.

Indexed Collections

Using what we have introduced so far, it is easy to extend the notion of ordered pair. Given a natural number n and a set S, an *ordered n-tuple* of elements of S is a function $x\colon \{1,\dots,n\} \to S$. It is customary to write x_i instead of $x(i)$ for the value of x at i, and the whole n-tuple can be written $(x_1,\dots,x_n)$ or $\{x_i : i = 1,\dots,n\}$ or $\{x_i\}_{i=1}^n$. Similarly, a *sequence* of elements of S is a function $x\colon \mathbb{N} \to S$, written $\{x_1, x_2, \dots\}$ or $\{x_i : i \in \mathbb{N}\}$ or $\{x_i\}_{i=1}^\infty$. An ordered n-tuple is sometimes called a *finite sequence.*

A *subsequence* of a sequence $\{x_i\}_{i\in\mathbb{N}}$ in a set S is a sequence of the form $\{x_{f(j)}\}_{j\in\mathbb{N}}$, where $f\colon \mathbb{N} \to \mathbb{N}$ is a function that is *strictly increasing*, meaning that $i < j$ implies $f(i) < f(j)$. We usually write i_j for $f(j)$.

We sometimes need to deal with collections of objects that are indexed, not by the natural numbers or subsets of them, but by arbitrary sets, potentially even uncountable ones. An *indexed collection* of elements of S is just a function from a set A (called the *index set*) to S, and in this context is denoted by $\{x_\alpha : \alpha \in A\}$ or $\{x_\alpha\}_{\alpha\in A}$. Occasionally, when the index set is understood or is irrelevant, we will omit it from the notation and simply write $\{x_\alpha\}$. If $\{X_\alpha\}_{\alpha\in A}$ is an indexed collection of sets, $\bigcup_{\alpha\in A} X_\alpha$ is just another notation for the union of the (unindexed) collection $\{X \in S : X = X_\alpha \text{ for some } \alpha \in A\}$, where S is the range of the indexing function. If the index set is finite, the union is usually written as $X_1 \cup \cdots \cup X_n$. A similar remark applies to the intersection $\bigcap_{\alpha\in A} X_\alpha$ or $X_1 \cap \cdots \cap X_n$.

Earlier we mentioned that given a set S and a partition of it, S is said to be the disjoint union of the sets in the partition. It sometimes happens that we are given a collection of sets, which may or may not be disjoint, but which we want to consider as disjoint subsets of a larger set. For example, we might want to form a set consisting of "five copies of $\mathbb{R}$," in which we consider the different copies to be disjoint from each other. We can accomplish this by the following trick. Suppose $\{X_\alpha\}_{\alpha\in A}$ is an indexed collection of nonempty sets. If we imagine "tagging" each element of X_α with its index α, we can make the sets X_α and X_β disjoint when $\alpha \neq \beta$, even if they were not disjoint to begin with. Formally, an element x with a tag α is just an ordered pair (x,α). Thus we define the *disjoint union* of the indexed collection, denoted by $\coprod_{\alpha\in A} X_\alpha$, to be the set

$$\coprod_{\alpha\in A} X_\alpha = \{(x,\alpha) : \alpha \in A \text{ and } x \in X_\alpha\}.$$

If the index set is finite, the disjoint union is usually written as $X_1 \amalg \cdots \amalg X_n$.

For each set X_α, there is a natural injective map $\iota_\alpha\colon X_\alpha \to \coprod_{\alpha\in A} X_\alpha$, given by $\iota_\alpha(x) = (x,\alpha)$. The images of these maps are disjoint from each other, so we can identify each set X_α with its image under ι_α. In practice, we think of each X_α as a subset of the disjoint union and think of the injection ι_α as an inclusion map. With this convention, this usage of the term disjoint union is consistent with our previous one.

The definition of Cartesian product now extends easily from two sets to arbitrarily many. If $(X_1, \dots, X_n)$ is an ordered n-tuple of sets, their Cartesian product $X_1 \times \cdots \times X_n$ is just the set of all ordered n-tuples $(x_1, \dots, x_n)$ such that $x_i \in X_i$ for $i = 1, \dots, n$. (To be sure we are strictly following the axioms, we should note that this is a subset of the set of all functions from $\{1, \dots, n\}$ to $X_1 \cup \cdots \cup X_n$, which in turn is a subset of the power set of $\{1, \dots, n\} \times (X_1 \cup \cdots \cup X_n)$.) If $X_1 = \cdots = X_n = X$, the n-fold Cartesian product $X \times \cdots \times X$ is often written simply as X^n.

A Cartesian product comes equipped with *projection maps* $\pi_i \colon X_1 \times \cdots \times X_n \to X_i$, defined by $\pi_i(x_1, \dots, x_n) = x_i$. It is easy to see that each of these maps is surjective. If $f \colon S \to X_1 \times \cdots \times X_n$ is any function into a Cartesian product, the composite functions $f_i = \pi_i \circ f \colon S \to X_i$ are called its *component functions*. Any such function f is completely determined by its component functions, by the formula

$$f(y) = (f_1(y), \dots, f_n(y)).$$

More generally, the Cartesian product of an arbitrary indexed collection $\{X_\alpha\}_{\alpha \in A}$ of sets is defined to be the set of all functions $x \colon A \to \bigcup_{\alpha \in A} X_\alpha$ such that $x_\alpha \in X_\alpha$ for each α. It is denoted by $\prod_{\alpha \in A} X_\alpha$. Just as in the case of finite products, any Cartesian product comes equipped with projection maps $\pi_\beta \colon \prod_{\alpha \in A} X_\alpha \to X_\beta$, defined by $\pi_\beta(x) = x_\beta$.

Our last set-theoretic axiom asserts that it is possible to choose an element from each set in an arbitrary indexed collection.

- Axiom of choice: Given any nonempty indexed collection $\{X_\alpha\}_{\alpha \in A}$ of nonempty sets, there exists a function $c \colon A \to \bigcup_{\alpha \in A} X_\alpha$, called a *choice function*, such that $c(\alpha) \in X_\alpha$ for each α.

In other words, the Cartesian product of any nonempty indexed collection of nonempty sets is nonempty.

Here are two immediate applications of the axiom of choice.

Exercise A.8. Complete the proof of Lemma A.1 by showing that f is surjective if and only if it has a right inverse.

Exercise A.9. If there exists a surjective map from a countable set onto S, prove that S is countable.

The axiom of choice has a number of interesting equivalent reformulations; the relationships among them make fascinating reading, for example in [Hal74]. The only other formulations we will make use of are the following two (the well-ordering theorem in Problem 4-6, and Zorn's lemma in the proof of Lemma 13.36).

Theorem A.2 (The Well-Ordering Theorem). *Every set can be given a total ordering that is well-ordered.*

Theorem A.3 (Zorn's Lemma). *Let X be a partially ordered set in which every totally ordered subset has an upper bound. Then X has a maximal element.*

For proofs, see [Hal74] or [Dev93].

Metric Spaces

Metric spaces play an indispensable role in real analysis, and their properties provide the underlying motivation for most of the basic definitions in topology. In this section we summarize the important properties of metric spaces with which you should be familiar. For a thorough treatment of the subject, see any good undergraduate real analysis text such as [Rud76].

Euclidean Spaces

Most of topology, in particular manifold theory, is modeled on the behavior of Euclidean spaces and their subsets, so we begin with a quick review of their properties.

The Cartesian product $\mathbb{R}^n = \mathbb{R} \times \cdots \times \mathbb{R}$ of n copies of the real line is known as *n-dimensional Euclidean space*. It is the set of ordered n-tuples of real numbers. A point in $\mathbb{R}^n$ is denoted by $(x_1, \dots, x_n)$ or simply x. The numbers x_i are called its *components* or *coordinates*. Zero-dimensional Euclidean space $\mathbb{R}^0$ is, by convention, the singleton $\{0\}$.

We will use without further comment the fact that $\mathbb{R}^n$ is an n-dimensional vector space with the usual operations of scalar multiplication and vector addition. The geometric properties of $\mathbb{R}^n$ are derived from the *Euclidean dot product* $x \cdot y = x_1 y_1 + \cdots + x_n y_n$. In particular, the *norm* or *length* of a vector $x \in \mathbb{R}^n$ is given by

$$|x| = (x \cdot x)^{1/2} = \left((x_1)^2 + \cdots + (x_n)^2\right)^{1/2}.$$

The *angle* between two nonzero vectors x, y is defined to be $\cos^{-1}(x \cdot y)/(|x|\,|y|)$. Given two points $x, y \in \mathbb{R}^n$, the *line segment* between them is the set $\{tx + (1-t)y : 0 \le t \le 1\}$.

Continuity and convergence in Euclidean spaces are defined in the usual ways. A map $f\colon U \to V$ between subsets of Euclidean spaces is continuous if for any $x \in U$ and any $\varepsilon > 0$ there exists $\delta > 0$ such that $|x - y| < \delta$ implies $|f(x) - f(y)| < \varepsilon$. A sequence $\{x_i\}$ of points in $\mathbb{R}^n$ converges to $x \in \mathbb{R}^n$ if for any $\varepsilon > 0$ there exists N such that $i \ge N$ implies $|x_i - x| < \varepsilon$.

Metrics, Convergence, and Continuity

Metric spaces are generalizations of Euclidean spaces, in which none of the vector space properties are present and only the distance function remains.

If M is any set, a *metric* on M is a function $d\colon M \times M \to \mathbb{R}$, also called a *distance function*, satisfying the following three properties:

(i) Symmetry: For all $x, y \in M$, $d(x,y) = d(y,x)$.

(ii) Positivity: For all $x, y \in M$, $d(x,y) \geq 0$, and $d(x,y) = 0$ if and only if $x = y$.

(iii) Triangle inequality: For all $x, y, z \in M$, $d(x,z) \leq d(x,y) + d(y,z)$.

The pair (M, d) is called a *metric space*. (Actually, unless it is important to specify which metric is being considered, we often just say "M is a metric space," with the metric being understood from the context.)

Example A.4.

(a) If M is any subset of $\mathbb{R}^n$, the function $d(x,y) = |x - y|$ is a metric on M (Exercise A.10), called the *Euclidean metric*. Whenever we consider a subset of $\mathbb{R}^n$ as a metric space, unless we specify otherwise it will always be with the Euclidean metric.

(b) Similarly, if M is any metric space and X is a subset of M, then X inherits a metric simply by restricting the distance function of M to points in X.

(c) If X is any set, define a metric on X by setting $d(x,y) = 1$ unless $x = y$, in which case $d(x,y) = 0$. This is called the *discrete metric* on X.

Exercise A.10. Prove that $d(x,y) = |x - y|$ is a metric on any subset of $\mathbb{R}^n$.

Here are some of the standard definitions used in metric space theory. Let M be a metric space.

- For any $x \in M$ and $r > 0$, the (*open*) *ball* of radius r around x is the set
$$B_r(x) = \{y \in M : d(y,x) < r\},$$
and the *closed ball* of radius r around x is
$$\overline{B}_r(x) = \{y \in M : d(y,x) \leq r\}.$$

- Given a subset $A \subset M$, a point $x \in M$ is said to be a *limit point* (or *accumulation point* or *cluster point*) of A if every open ball around x contains a point of A other than x.

- A set $A \subset M$ is said to be *open* if it contains an open ball around each of its points.

- A set $A \subset M$ is said to be *closed* if it contains all its limit points.
- The *diameter* of a set $A \subset M$ is $\sup\{d(x,y) : x, y \in A\}$ (which may be infinite).
- A set $A \subset M$ is said to be *bounded* if it has finite diameter.

Exercise A.11. Let M be a metric space.

(a) Show that $A \subset M$ is open if and only if $M \smallsetminus A$ is closed.

(b) Show that an open ball in M is an open set, and a closed ball in M is a closed set.

(c) Show that the union of an arbitrary collection of open sets is open, and the intersection of finitely many open sets is open.

(d) Show that the intersection of an arbitrary collection of closed sets is closed, and the union of finitely many closed sets is closed.

(e) Show that a subset of M is bounded if and only if it is contained in an open ball if and only if it is contained in a closed ball.

Exercise A.12. In each part below, a subset S of a metric space M is given. In each case, decide whether S is open, closed, both, or neither.

(a) $M = \mathbb{R}$, and $S = [0, 1)$.

(b) $M = \mathbb{R}$, and $S = \mathbb{N}$.

(c) $M = \mathbb{Z}$, and $S = \mathbb{N}$.

(d) $M = \mathbb{R}^2$, and S is the set of points with rational coordinates.

(e) $M = \mathbb{R}^2$, and S is the unit disk $\{(x, y) \in \mathbb{R}^2 : x^2 + y^2 < 1\}$.

(f) $M = \mathbb{R}^3$, and S is the unit disk $\{(x, y, z) \in \mathbb{R}^3 : z = 0 \text{ and } x^2+y^2 < 1\}$.

(g) $M = \{(x, y) \in \mathbb{R}^2 : x > 0 \text{ and } y > 0\}$, and $S = \{(x, y) \in M : x^2+y^2 \leq 1\}$.

The definition of continuity in the context of metric spaces is a straightforward generalization of the Euclidean definition. If (M_1, d_1) and (M_2, d_2) are metric spaces, a map $f\colon M_1 \to M_2$ is said to be *continuous* if for every $x \in M_1$ and every $\varepsilon > 0$, there exists $\delta > 0$ such that $d_1(x, y) < \delta$ implies $d_2(f(x), f(y)) < \varepsilon$. Similarly, if $\{x_i\}$ is a sequence of points in a metric space (M, d), it is said to *converge* to $x \in M$, written $x_i \to x$ or $\lim_{i\to\infty} x_i = x$, if for any $\varepsilon > 0$ there exists N such that $i \geq N$ implies $d(x_i, x) < \varepsilon$.

Exercise A.13. If M and N are metric spaces and $f\colon M \to N$ is a map, show that f is continuous if and only if it takes convergent sequences to convergent sequences, i.e., if and only if $x_i \to y$ in M implies $f(x_i) \to f(y)$ in N.

Exercise A.14. Show that a subset A of a metric space M is closed if and only if, whenever $\{x_i\}$ is a sequence of points in A that converges in M, the limit lies in A.

A sequence $\{x_i\}$ in a metric space is said to be *Cauchy* if for every $\varepsilon > 0$, there exists N such that $i, j \geq N$ implies $d(x_i, x_j) < \varepsilon$. Every convergent sequence is Cauchy (Exercise A.15), but the converse is not true in general. A metric space in which every Cauchy sequence converges is said to be *complete*.

Exercise A.15. Prove that every convergent sequence in a metric space is Cauchy.

The following criterion for continuity is frequently useful (and in fact, as is explained in Chapter 2, is the main motivation for the definition of a topological space).

Lemma A.5 (Open Set Criterion for Continuity). *A map $f\colon M_1 \to M_2$ between metric spaces is continuous if and only if the inverse image of every open set is open: Whenever U is an open subset of M_2, $f^{-1}(U)$ is open in M_1.*

Exercise A.16. Prove Lemma A.5

Compactness

Let M be a metric space and K a subset of M. An *open cover* of K is a collection $\{U_\alpha\}_{\alpha\in A}$ of open subsets of M whose union contains K. A *subcover* is a subcollection that is still an open cover of K. A subset of M is said to be *compact* if every open cover has a finite subcover.

Two properties of compact sets—closedness and boundedness—follow immediately from the definition.

Proposition A.6. *Any compact subset of a metric space is closed and bounded.*

Proof. Let $K \subset M$ be compact, and let x be any point in K. The collection of open balls $\{B_r(x) : r > 0\}$ is an open cover of K, which therefore must have a finite subcover. Letting R be the radius of the largest of these finitely many balls, it follows that $K \subset B_R(x)$, which means that it is bounded.

To show that K is closed, we will show that its complement is open. Let q be any point of $M \smallsetminus K$. For each $p \in K$, let $\delta(p) = d(p, q)/2$; then the balls $B_{\delta(p)}(q)$ and $B_{\delta(p)}(p)$ are disjoint by the triangle inequality. The collection $\{B_{\delta(p)}(p) : p \in K\}$ is an open cover of K, and therefore has a finite subcover, say $\{B_{\delta(p_1)}(p_1), \dots, B_{\delta(p_k)}(p_k)\}$. Now let $r = \min\{\delta(p_1), \dots, \delta(p_k)\}$. Since $B_r(q)$ is disjoint from each of the balls $B_{\delta(p_i)}(p_i)$ and these balls cover K, it is disjoint from K. In other words, there is a ball around each $q \in M \smallsetminus K$ contained in $M \smallsetminus K$, so $M \smallsetminus K$ is open and thus K is closed. $\square$

In $\mathbb{R}^n$, the converse of this proposition is true. Before proving this result (the Heine–Borel theorem below), we need a preliminary lemma.

For any point $x \in \mathbb{R}^n$ and any $r > 0$, the *closed cube* of side r around x is the set

$$\overline{C}_r(x) = \{y \in \mathbb{R}^n : |x_i - y_i| \le r/2,\ i = 1, \dots, n\}.$$

It is easy to check that the diameter of $\overline{C}_r(x)$ is $r\sqrt{n}$.

Lemma A.7. *Suppose $\{C_i\}_{i=1}^{\infty}$ is a sequence of closed cubes in $\mathbb{R}^n$ that are nested, in the sense that $C_1 \supset C_2 \supset \cdots$. Then $\bigcap_k C_k$ is not empty.*

Proof. First consider the case $n = 1$, in which case the cubes are just closed intervals. Writing the intervals as $C_k = [a_k, b_k]$, the fact that they are nested means that $a_1 \le a_2 \le \cdots \le a_k < b_k \le b_{k-1} \le \cdots \le b_2 \le b_1$. Let $a = \sup\{a_k\}$ and $b = \inf\{b_k\}$. Then clearly, $a_k \le a \le b \le b_k$ for each k, so the interval $[a, b]$ (or the point a if $a = b$) is contained in $\bigcap_k C_k$.

For general n, just apply the preceding argument to each coordinate. □

Theorem A.8 (The Heine–Borel Theorem). *Every closed and bounded subset of $\mathbb{R}^n$ is compact.*

Proof. We begin by showing that any closed cube in $\mathbb{R}^n$ is compact. Let C be a closed cube of side r, and let $\mathcal{U}$ be an open cover of C. Suppose $\mathcal{U}$ has no finite subcover. Subdividing each of the sides of C in half yields a decomposition of C into 2^n closed cubes of side $r/2$ whose union is C. If each of these 2^n cubes were covered by finitely many sets of $\mathcal{U}$, then putting together these 2^n finite collections of open sets would give a finite subcover of C; thus at least one of them must not be covered by any finite subcollection of sets from $\mathcal{U}$. Call this smaller cube C_1. If we subdivide C_1 in the same way, one of the 2^n cubes in this subdivision must not admit a finite subcover by the same reasoning. Continuing by induction, we obtain a nested sequence of cubes $C = C_0 \supset C_1 \supset \cdots$ with the property that no C_k is covered by any finite collection of sets from $\mathcal{U}$. Each cube C_k has side length $r/2^k$.

By Lemma A.7, there is a point $x \in \bigcap_k C_k$. Because $\mathcal{U}$ is a cover of C, x must be in one of the sets of $\mathcal{U}$, say $x \in U_0$. Because U_0 is open, there is a ball $B_\varepsilon(x) \subset U_0$. Because C_k has diameter $r\sqrt{n}/2^k$ and $x \in C_k$, as soon as k is sufficiently large $C_k \subset B_\varepsilon(x) \subset U_0$, which contradicts the fact that C_k cannot be covered by finitely many sets of $\mathcal{U}$. This proves that any cube is compact.

Now suppose $K \subset \mathbb{R}^n$ is any closed and bounded subset. Because it is bounded, it is contained in some cube $\overline{C}_r(0)$. If $\mathcal{U}$ is an open cover of K, the collection $\mathcal{U} \cup \{\mathbb{R}^n \smallsetminus K\}$ is an open cover of $\overline{C}_r(0)$. (Here we use the fact that K is closed.) Finitely many of these sets, say $\{U_1, \dots, U_m, \mathbb{R}^n \smallsetminus K\}$, cover $\overline{C}_r(0)$, and then it is clear that $\{U_1, \dots, U_m\}$ cover K. □

The Heine–Borel theorem is not true if $\mathbb{R}^n$ is replaced by a general metric space, as the following exercise shows.

Exercise A.17.

(a) In $\mathbb{Z}$ with the discrete metric, show that any infinite subset is closed and bounded, but not compact.

(b) Similarly, in the metric space $(0, \infty)$, show that the set $(0, \frac{1}{2}]$ is closed and bounded, but the open cover of it by intervals of the form $(1/n, 1)$ for $n \in \mathbb{N}$ has no finite subcover.

Most of the applications of compactness depend on the following theorem.

Theorem A.9. *If M and N are metric spaces, $f\colon M \to N$ is continuous, and $K \subset M$ is compact, then $f(K)$ is compact.*

Exercise A.18. Prove Theorem A.9.

For example, the following theorem is of fundamental importance in real analysis.

Theorem A.10 (Euclidean Extreme Value Theorem). *Any continuous real-valued function on a closed and bounded subset of $\mathbb{R}^n$ attains its maximum and minimum values.*

Proof. Let $f\colon K \to \mathbb{R}$ be such a function. Since K is closed and bounded, it is compact by the Heine–Borel theorem. By the preceding theorem, $f(K)$ is compact and therefore closed and bounded in $\mathbb{R}$. In particular, it contains its supremum and infimum. $\square$

Group Theory

We will assume only basic group theory such as one is likely to encounter in most undergraduate algebra courses. You can find much more detail about all of this material in, for example, [Hun90] or [Her75].

Basic Definitions

A *group* is a set G together with a map $G \times G \to G$, usually called *multiplication* and written $(g, h) \mapsto gh$, satisfying

(i) Associativity: For all $g, h, k \in G$, $(gh)k = g(hk)$.

(ii) Existence of identity: There is an element $1 \in G$ such that $1g = g1 = g$ for all $g \in G$.

(iii) Existence of inverses: For each $g \in G$, there is an element $h \in G$ such that $gh = hg = 1$.

The *order* of a group G is its cardinality as a set. The *trivial group* is the unique group of order 1; it is the group consisting of the identity alone. One checks easily that the inverse of any element is unique (so the usual notation g^{-1} for inverses makes sense), and that $(gh)^{-1} = h^{-1}g^{-1}$. Similarly, the identity is unique.

A group G is said to be *abelian* if $gh = hg$ for all $g, h \in G$. The group operation in an abelian group is frequently written additively, $(g, h) \mapsto g + h$, in which case the identity element is denoted by 0 and the inverse of g is denoted by $-g$.

A subset of G that is itself a group with the same multiplication is called a *subgroup* of G. Clearly, a subset is a subgroup if and only if it is closed under multiplication and contains the inverse of each of its elements.

If S is any set, the set of permutations of S is a group under composition, called the *permutation group* of S. In particular, if S is a finite set, any permutation of S can be factored as a product of *transpositions*, which are permutations that interchange two elements and leave all others fixed. The factorization into transpositions is not uniquely determined, but the parity (evenness or oddness) of the number of transpositions is the same for every such factorization. A permutation is called *even* or *odd* depending on whether it decomposes into an even or odd number of transpositions.

Exercise A.19. Let S_n denote the group of permutations of the set $\{1, \ldots, n\}$, called the *symmetric group* on n elements.

(a) Show that the map $\operatorname{sgn}\colon S_n \to \{\pm 1\}$ given by

$$\operatorname{sgn}(s) = \begin{cases} +1 & \text{if } s \text{ is even,} \\ -1 & \text{if } s \text{ is odd} \end{cases}$$

is a surjective homomorphism. (Here we consider $\{\pm 1\}$ as a group under multiplication.)

(b) Show that every element of S_n can be written as a product of transpositions of the following type:

$$\begin{aligned} s_k(k) &= k + 1; \\ s_k(k+1) &= k; \\ s_k(i) &= i \quad \text{if } i \neq k \text{ or } k+1. \end{aligned}$$

If $G_1, \ldots, G_n$ are groups, their *direct product* is the set $G_1 \times \cdots \times G_n$ with the group structure defined by the multiplication law

$$(g_1, \ldots, g_n)(g_1', \ldots, g_n') = (g_1 g_1', \ldots, g_n g_n')$$

and with identity element $(1, \ldots, 1)$. More generally, the direct product of an arbitrary indexed collection of groups $\{G_\alpha\}_{\alpha \in A}$ is the Cartesian product set $\prod_{\alpha \in A} G_\alpha$ with multiplication defined componentwise: $(gg')_\alpha = g_\alpha g_\alpha'$.

A map $f\colon G \to H$ between groups is called a *homomorphism* if it preserves multiplication: $f(gh) = f(g)f(h)$. The *image* of f is $f(G) \subset H$, often written $\operatorname{Im} f$, and its *kernel* is the set $f^{-1}(1)$, denoted by $\operatorname{Ker} f$. A bijective homomorphism is called an *isomorphism*. An isomorphism from G to itself is called an *automorphism*.

Exercise A.20. Let $f\colon G \to H$ be a homomorphism.

- Show that f is injective if and only if $\operatorname{Ker} f = \{1\}$.
- If f is bijective, show that f^{-1} is also a homomorphism.
- Show that the image and the kernel of f are subgroups.
- If $K \subset G$ is a subgroup, show that $f(K)$ is a subgroup of H.

An element $g \in G$ defines a map $C_g\colon G \to G$ by $C_g(h) = ghg^{-1}$. This map, called *conjugation by g*, is easily seen to be an automorphism of G, so the image under C_g of any subgroup $H \subset G$ (written symbolically as gHg^{-1}) is another subgroup of G. Two subgroups H, H' are *conjugate* if $H' = gHg^{-1}$ for some $g \in G$.

Exercise A.21. Show that conjugacy is an equivalence relation on the set of all subgroups of G.

The set of subgroups conjugate to a given subgroup $H \subset G$ is called the *conjugacy class* of H.

Cosets and Quotient Groups

Given a subgroup $H \subset G$ and an element $g \in G$, the *left coset* of H determined by g is the set

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

The right coset Hg is defined similarly. Define a relation called *congruence modulo H* by declaring that $g \equiv g' \pmod{H}$ if and only if $g^{-1}g' \in H$.

Exercise A.22. Show that congruence modulo H is an equivalence relation on G, and its equivalence classes are precisely the left cosets of H.

The set of left cosets of H in G is denoted by G/H. (This is just the partition of G defined by congruence modulo H.) The cardinality of G/H is called the *index* of H in G.

A subgroup $K \subset G$ is said to be *normal* if it is invariant under all conjugations, that is if $gKg^{-1} = K$ for all $g \in G$. Clearly, every subgroup of an abelian group is normal.

Exercise A.23. Show that a subgroup $K \subset G$ is normal if and only if $gK = Kg$ for every $g \in G$, so that every left coset of K is also a right coset.

Exercise A.24. Show that the kernel of any homomorphism is a normal subgroup.

Normal subgroups give rise to one of the most important constructions in group theory. Given a normal subgroup $K \subset G$, define a multiplication operator on the set G/K of left cosets by

$$(gK)(g'K) = (gg')K.$$

Lemma A.11. *This multiplication is well-defined on cosets and turns G/K into a group.*

Proof. First we need to show that the product does not depend on the representatives chosen for the cosets: If $gK = g'K$ and $hK = h'K$, then $(gh)K = (g'h')K$. From Exercise A.22, the fact that g and g' determine the same coset means that $g^{-1}g' \in K$, which is the same as saying $g' = gk$ for some $k \in K$. Similarly, $h' = hk'$ for $k' \in K$. Because K is normal, $h^{-1}kh$ is an element of K. Writing this element as k'', we have $kh = hk''$. It follows that

$$g'h' = gkhk' = ghk''k',$$

which shows that $g'h'$ and gh determine the same coset.

Now we just note that the group properties are satisfied: Associativity of the multiplication in G/K follows from that of G; the element $1K = K$ of G/K acts as an identity; and $g^{-1}K$ is the inverse of gK. □

When K is a normal subgroup of G, the group G/K is called the *quotient group* of G by K. The natural projection map $\pi\colon G \to G/K$ that sends each element to its coset is a surjective homomorphism whose kernel is K.

The following lemma tells how to define homomorphisms from a quotient group.

Lemma A.12. *Let $K \subset G$ be a normal subgroup. Given a homomorphism $f\colon G \to H$ such that $K \subset \operatorname{Ker} f$, there is a unique homomorphism $\widetilde{f}\colon G/K \to H$ such that the following diagram commutes:*

$$\begin{tikzcd} G \arrow[r, "f"] \arrow[d, "\pi"'] & H \\ G/K. \arrow[ur, "\widetilde{f}"'] & \end{tikzcd} \tag{A.1}$$

(A diagram such as (A.1) is said to *commute*, or to be *commutative*, if the maps between two spaces obtained by following arrows around either side of the diagram are equal. So in this case commutativity means that $\widetilde{f} \circ \pi = f$.)

Proof. Since $\pi(g) = gK$, if such a map exists, it has to be given by the formula $\widetilde{f}(gK) = f(g)$; this proves uniqueness. To prove existence, we just define $\widetilde{f}$ by this formula, so it obviously makes the diagram commute. It is well-defined, because if $g \equiv g' \pmod{K}$, then $g' = gk$ for some $k \in K$, and therefore $f(g') = f(gk) = f(g)f(k) = f(g)$. It is clear from the definition of multiplication in G/K that it is a homomorphism. $\square$

In the situation of the preceding lemma, we say that the homomorphism f *passes to the quotient* or *descends to the quotient*.

The most important fact about quotient groups is the following result, which says in essence that the projection onto a quotient group is the model for all surjective homomorphisms.

Theorem A.13 (First Isomorphism Theorem). *Let G, H be groups, and let $f\colon G \to H$ be a surjective homomorphism. Then $K = \operatorname{Ker} f$ is a normal subgroup of G, and f induces an isomorphism $\widetilde{f}\colon G/K \to H$ by $\widetilde{f}(gK) = f(g)$.*

Proof. From the preceding lemma, $\widetilde{f}(gK) = f(g)$ defines a homomorphism $\widetilde{f}\colon G/K \to H$. Because f is surjective, $\widetilde{f}$ is surjective: For any $h \in H$ there is an element $g \in G$ with $f(g) = h$, and then $\widetilde{f}(gK) = h$. To show that $\widetilde{f}$ is injective, suppose $1 = \widetilde{f}(gK) = f(g)$. This means that $g \in \operatorname{Ker} f = K$, so $gK = K$ is the identity element of G/K. $\square$

Exercise A.25. Suppose $f\colon G \to H$ is a surjective homomorphism, and $K \subset G$ is a normal subgroup. Show that $f(K)$ is normal in H.

Cyclic Groups

Let G be a group and $g \in G$. The set $\langle g \rangle = \{g^n : n \in \mathbb{Z}\}$ is obviously a subgroup of G, called the *cyclic subgroup* generated by g. The group G is said to be *cyclic* if $G = \langle g \rangle$ for some element $g \in G$. In this case, the element g is called a *generator* of G.

Example A.14 (Cyclic Groups).

(a) The group $\mathbb{Z}$ of integers (under addition) is an infinite cyclic group generated by 1.

(b) For any $n \in \mathbb{Z}$, the cyclic subgroup $\langle n \rangle$ is normal because $\mathbb{Z}$ is abelian. The quotient group $\mathbb{Z}/\langle n \rangle$ is called the group of *integers modulo n*. It is easily seen to be a cyclic group of order n, with the coset of 1 as a generator.

Exercise A.26. Show that any infinite cyclic group is isomorphic to $\mathbb{Z}$ and any finite cyclic group is isomorphic to $\mathbb{Z}/\langle n \rangle$, where n is the order of the group.

Exercise A.27. Show that every subgroup of a cyclic group is cyclic.

Exercise A.28. If G is a cyclic group and $f\colon G \to G$ is any homomorphism, show there is an integer n such that $f(\gamma) = \gamma^n$ for all $\gamma \in G$. If G is infinite, show that n is uniquely determined by f.

References

[Dev93] Keith Devlin. *The Joy of Sets: Fundamentals of Contemporary Set Theory*. Springer-Verlag, New York, second edition, 1993.

[DH07] Max Dehn and Poul Heegaard. Analysis situs. *Enzyklopädie der Math. Wiss.*, III.1.1:153–220, 1907.

[Die89] Jean Dieudonné. *A History of Algebraic and Differential Topology 1900–1960*. Birkhäuser Boston, Cambridge, 1989.

[Fre82] Michael Hartley Freedman. The topology of four-dimensional manifolds. *J. Differential Geom.*, 17:357–453, 1982.

[Hal74] Paul R. Halmos. *Naive Set Theory*. Springer-Verlag, New York, 1974.

[Her75] Israel N. Herstein. *Topics in Algebra*. Wiley, New York, second edition, 1975.

[Hun90] Thomas W. Hungerford. *Abstract Algebra: An Introduction*. Saunders College Publishing, Philadelphia, 1990.

[Lee97] John M. Lee. *Riemannian Manifolds: An Introduction to Curvature*. Springer-Verlag, New York, 1997.

[LMM94] Silvio Levy, Delle Maxwell, and Tamara Munzer. Outside in. The Geometry Center (video), 1994.

[Mas89] William S. Massey. *Algebraic Topology: An Introduction.* Springer-Verlag, New York, 1989.

[Max77] Nelson Max. Turning a sphere inside out. International Film Bureau (video), 1977.

[Moi77] Edwin E. Moise. *Geometric Topology in Dimensions 2 and 3.* Springer-Verlag, New York, 1977.

[Mun75] James R. Munkres. *Topology: A First Course.* Prentice–Hall, Englewood Cliffs, New Jersey, 1975.

[Mun84] James R. Munkres. *Elements of Algebraic Topology.* Addison–Wesley, Reading, Massachusetts, 1984.

[Rad25] Tibor Radó. Über den Begriff der Riemannschen Fläche. *Acta Litt. Sci. Szeged.*, 2:101–121, 1925.

[Ran96] A. A. Ranicki. On the Hauptvermutung. In A. A. Ranicki, editor, *The Hauptvermutung Book*, volume 1 of *K-Monographs in Mathematics*, pages 3–31. Kluwer Acad. Publ., Dordrecht, 1996.

[Rud76] Walter Rudin. *Principles of Mathematical Analysis.* McGraw-Hill, New York, third edition, 1976.

[Sco83] Peter Scott. The geometries of 3-manifolds. *Bull. London Math. Soc.*, 15:401–487, 1983.

[SFL98] John M. Sullivan, George Francis, and Stuart Levy. The optiverse. National Center for Supercomputing Applications (video), 1998.

[Sie92] Allan J. Sieradski. *An Introduction to Topology and Homotopy.* PWS-Kent, Boston, 1992.

[Sma58] Stephen Smale. A classification of immersions of the two-sphere. *Trans. Amer. Math. Soc.*, 90:281–290, 1958.

[Sma61] Stephen Smale. Generalized Poincaré conjecture in dimensions greater than 4. *Ann. Math.*, 74:391–406, 1961.

[Spa89] Edwin H. Spanier. *Algebraic Topology.* Springer-Verlag, New York, 1989.

[Sti82] John Stillwell. The word problem and the isomorphism problem for groups. *Bull. Amer. Math. Soc. (N.S.)*, 6:33–56, 1982.

[Sti93] John Stillwell. *Classical Topology and Combinatorial Group Theory.* Springer-Verlag, New York, second edition, 1993.

[Thu97] William P. Thurston. *Three-dimensional geometry and topology*, volume 1. Princeton University Press, Princeton, 1997.

[Whi78] George W. Whitehead. *Elements of Homotopy Theory*. Springer-Verlag, New York, 1978.

Index

Graduate Texts in Mathematics

(continued from page ii)

72 STILLWELL. Classical Topology and Combinatorial Group Theory. 2nd ed.
73 HUNGERFORD. Algebra.
74 DAVENPORT. Multiplicative Number Theory. 2nd ed.
75 HOCHSCHILD. Basic Theory of Algebraic Groups and Lie Algebras.
76 IITAKA. Algebraic Geometry.
77 HECKE. Lectures on the Theory of Algebraic Numbers.
78 BURRIS/SANKAPPANAVAR. A Course in Universal Algebra.
79 WALTERS. An Introduction to Ergodic Theory.
80 ROBINSON. A Course in the Theory of Groups. 2nd ed.
81 FORSTER. Lectures on Riemann Surfaces.
82 BOTT/TU. Differential Forms in Algebraic Topology.
83 WASHINGTON. Introduction to Cyclotomic Fields. 2nd ed.
84 IRELAND/ROSEN. A Classical Introduction to Modern Number Theory. 2nd ed.
85 EDWARDS. Fourier Series. Vol. II. 2nd ed.
86 VAN LINT. Introduction to Coding Theory. 2nd ed.
87 BROWN. Cohomology of Groups.
88 PIERCE. Associative Algebras.
89 LANG. Introduction to Algebraic and Abelian Functions. 2nd ed.
90 BRØNDSTED. An Introduction to Convex Polytopes.
91 BEARDON. On the Geometry of Discrete Groups.
92 DIESTEL. Sequences and Series in Banach Spaces.
93 DUBROVIN/FOMENKO/NOVIKOV. Modern Geometry—Methods and Applications. Part I. 2nd ed.
94 WARNER. Foundations of Differentiable Manifolds and Lie Groups.
95 SHIRYAEV. Probability. 2nd ed.
96 CONWAY. A Course in Functional Analysis. 2nd ed.
97 KOBLITZ. Introduction to Elliptic Curves and Modular Forms. 2nd ed.
98 BRÖCKER/TOM DIECK. Representations of Compact Lie Groups.
99 GROVE/BENSON. Finite Reflection Groups. 2nd ed.
100 BERG/CHRISTENSEN/RESSEL. Harmonic Analysis on Semigroups: Theory of Positive Definite and Related Functions.
101 EDWARDS. Galois Theory.
102 VARADARAJAN. Lie Groups, Lie Algebras and Their Representations.
103 LANG. Complex Analysis. 3rd ed.
104 DUBROVIN/FOMENKO/NOVIKOV. Modern Geometry—Methods and Applications. Part II.
105 LANG. $SL_2(\mathbf{R})$.
106 SILVERMAN. The Arithmetic of Elliptic Curves.
107 OLVER. Applications of Lie Groups to Differential Equations. 2nd ed.
108 RANGE. Holomorphic Functions and Integral Representations in Several Complex Variables.
109 LEHTO. Univalent Functions and Teichmüller Spaces.
110 LANG. Algebraic Number Theory.
111 HUSEMÖLLER. Elliptic Curves.
112 LANG. Elliptic Functions.
113 KARATZAS/SHREVE. Brownian Motion and Stochastic Calculus. 2nd ed.
114 KOBLITZ. A Course in Number Theory and Cryptography. 2nd ed.
115 BERGER/GOSTIAUX. Differential Geometry: Manifolds, Curves, and Surfaces.
116 KELLEY/SRINIVASAN. Measure and Integral. Vol. I.
117 SERRE. Algebraic Groups and Class Fields.
118 PEDERSEN. Analysis Now.
119 ROTMAN. An Introduction to Algebraic Topology.
120 ZIEMER. Weakly Differentiable Functions: Sobolev Spaces and Functions of Bounded Variation.
121 LANG. Cyclotomic Fields I and II. Combined 2nd ed.
122 REMMERT. Theory of Complex Functions. *Readings in Mathematics*
123 EBBINGHAUS/HERMES et al. Numbers. *Readings in Mathematics*
124 DUBROVIN/FOMENKO/NOVIKOV. Modern Geometry—Methods and Applications. Part III.
125 BERENSTEIN/GAY. Complex Variables: An Introduction.
126 BOREL. Linear Algebraic Groups. 2nd ed.
127 MASSEY. A Basic Course in Algebraic Topology.
128 RAUCH. Partial Differential Equations.
129 FULTON/HARRIS. Representation Theory: A First Course. *Readings in Mathematics*
130 DODSON/POSTON. Tensor Geometry.
131 LAM. A First Course in Noncommutative Rings.
132 BEARDON. Iteration of Rational Functions.
133 HARRIS. Algebraic Geometry: A First Course.
134 ROMAN. Coding and Information Theory.
135 ROMAN. Advanced Linear Algebra.

136 ADKINS/WEINTRAUB. Algebra: An Approach via Module Theory.
137 AXLER/BOURDON/RAMEY. Harmonic Function Theory.
138 COHEN. A Course in Computational Algebraic Number Theory.
139 BREDON. Topology and Geometry.
140 AUBIN. Optima and Equilibria. An Introduction to Nonlinear Analysis.
141 BECKER/WEISPFENNING/KREDEL. Gröbner Bases. A Computational Approach to Commutative Algebra.
142 LANG. Real and Functional Analysis. 3rd ed.
143 DOOB. Measure Theory.
144 DENNIS/FARB. Noncommutative Algebra.
145 VICK. Homology Theory. An Introduction to Algebraic Topology. 2nd ed.
146 BRIDGES. Computability: A Mathematical Sketchbook.
147 ROSENBERG. Algebraic K-Theory and Its Applications.
148 ROTMAN. An Introduction to the Theory of Groups. 4th ed.
149 RATCLIFFE. Foundations of Hyperbolic Manifolds.
150 EISENBUD. Commutative Algebra with a View Toward Algebraic Geometry.
151 SILVERMAN. Advanced Topics in the Arithmetic of Elliptic Curves.
152 ZIEGLER. Lectures on Polytopes.
153 FULTON. Algebraic Topology: A First Course.
154 BROWN/PEARCY. An Introduction to Analysis.
155 KASSEL. Quantum Groups.
156 KECHRIS. Classical Descriptive Set Theory.
157 MALLIAVIN. Integration and Probability.
158 ROMAN. Field Theory.
159 CONWAY. Functions of One Complex Variable II.
160 LANG. Differential and Riemannian Manifolds.
161 BORWEIN/ERDÉLYI. Polynomials and Polynomial Inequalities.
162 ALPERIN/BELL. Groups and Representations.
163 DIXON/MORTIMER. Permutation Groups.
164 NATHANSON. Additive Number Theory: The Classical Bases.
165 NATHANSON. Additive Number Theory: Inverse Problems and the Geometry of Sumsets.
166 SHARPE. Differential Geometry: Cartan's Generalization of Klein's Erlangen Program.
167 MORANDI. Field and Galois Theory.
168 EWALD. Combinatorial Convexity and Algebraic Geometry.
169 BHATIA. Matrix Analysis.
170 BREDON. Sheaf Theory. 2nd ed.
171 PETERSEN. Riemannian Geometry.
172 REMMERT. Classical Topics in Complex Function Theory.
173 DIESTEL. Graph Theory. 2nd ed.
174 BRIDGES. Foundations of Real and Abstract Analysis.
175 LICKORISH. An Introduction to Knot Theory.
176 LEE. Riemannian Manifolds.
177 NEWMAN. Analytic Number Theory.
178 CLARKE/LEDYAEV/STERN/WOLENSKI. Nonsmooth Analysis and Control Theory.
179 DOUGLAS. Banach Algebra Techniques in Operator Theory. 2nd ed.
180 SRIVASTAVA. A Course on Borel Sets.
181 KRESS. Numerical Analysis.
182 WALTER. Ordinary Differential Equations.
183 MEGGINSON. An Introduction to Banach Space Theory.
184 BOLLOBAS. Modern Graph Theory.
185 COX/LITTLE/O'SHEA. Using Algebraic Geometry.
186 RAMAKRISHNAN/VALENZA. Fourier Analysis on Number Fields.
187 HARRIS/MORRISON. Moduli of Curves.
188 GOLDBLATT. Lectures on the Hyperreals: An Introduction to Nonstandard Analysis.
189 LAM. Lectures on Modules and Rings.
190 ESMONDE/MURTY. Problems in Algebraic Number Theory.
191 LANG. Fundamentals of Differential Geometry.
192 HIRSCH/LACOMBE. Elements of Functional Analysis.
193 COHEN. Advanced Topics in Computational Number Theory.
194 ENGEL/NAGEL. One-Parameter Semigroups for Linear Evolution Equations.
195 NATHANSON. Elementary Methods in Number Theory.
196 OSBORNE. Basic Homological Algebra.
197 EISENBUD/ HARRIS. The Geometry of Schemes.
198 ROBERT. A Course in p-adic Analysis.
199 HEDENMALM/KORENBLUM/ZHU. Theory of Bergman Spaces.
200 BAO/CHERN/SHEN. An Introduction to Riemann–Finsler Geometry.
201 HINDRY/SILVERMAN. Diophantine Geometry: An Introduction.
202 LEE. Introduction to Topological Manifolds.