1

Basic Notions of Algebraic Topology

> …geometry is the art of reasoning well from badly drawn figures; however, these figures, if they are not to deceive us, must satisfy certain conditions; the proportions may be grossly altered, but the relative positions of the different parts must not be upset …
>
> *Henri Poincaré*

Modern algebraic topology arose in order to provide quantitative tools for studying the “shape” of geometric objects without using distances. It assigns algebraic invariants (e.g., numbers) to geometric objects in a way that depends only on the relative, not absolute, positions of points. In this chapter, we motivate and introduce the basic ideas of algebraic topology. This material provides a conceptual framework for understanding the tools of topological data analysis and their application to real data. We do not provide a complete treatment, and in particular we omit proofs of the theorems. At the beginning of each section, we provide a reference to a comprehensive source for the material.

Although algebraic topology is not yet a standard tool in genomics, the study of shape is already ubiquitous – clustering techniques are widely used to analyze data in all domains of molecular biology. For example, we can represent the expression profile of genes in cancer patients as points in a high-dimensional Euclidean space. Patients that have similar expression profiles will have points that are close together. A clustering algorithm can then be employed to classify expression profiles of cancer patients and thereby illuminate some of the distinct molecular mechanisms underlying the disease.

Recall that a clustering algorithm assigns to a finite collection of points X equipped with a distance function ∂_X a partition of the points of X, i.e., a collection of subsets $C_i \subseteq X$ such that

1. the $\{C_i\}$ do not overlap, so $C_i \cap C_j = \emptyset$ for all $i \neq j$, and
2. together the $\{C_i\}$ cover all of X, so that $\bigcup_i C_i = X$.

These subsets C_i are the clusters. Typically, clustering algorithms seek to generate partitions so that points within a given cluster are closer together than points in distinct clusters.

A representative clustering algorithm, single-linkage clustering in Euclidean space, takes as input a set of points $X \subseteq \mathbb{R}^n$ and a fixed $\epsilon > 0$, and assigns points x and y to the same cluster if there is a path of points

$$x = x_0, x_1, x_2, \dots x_{k-1}, x_k = y$$

such that $\|x_i - x_{i-1}\| < \epsilon$ for $1 \leq i \leq k$. (Here for $x, y \in \mathbb{R}^n$, $\|x - y\|$ denotes the Euclidean distance between the points x and y, see Example 1.3.6.) In other words, we connect points if they are closer than ϵ; clusters are groups of connected points.

The methodology of clustering is motivated by the same focus on relative information as in algebraic topology. Specifically, clustering is a useful technique for analyzing data in circumstances in which the data is very noisy, so relative information is more reliable than absolute information. In fact, the connection between clustering and algebraic topology is very close: as we shall see in Section 1.3.2, single-linkage clustering has an interpretation in terms of a standard topological invariant.

In contrast to clustering techniques, which typically work on a collection of separated points (referred to as a “point cloud”), algebraic topology has traditionally concerned itself with continuous objects with infinitely many points which can be arbitrarily close together, e.g., a sphere. A first question we might ask is “what is the continuous analogue of the clustering algorithm described above?” Roughly speaking, the answer to this question will be as follows: a “cluster” should consist of all points which can be connected by a smooth path.

In order to make sense of this, we need a precise definition of a geometric object and of a smooth path through a geometric object. In the continuous setting, this is done using the notion of a *topological space*. The study of basic properties of topological spaces is typically referred to as *point-set topology*. We begin by giving a little background about sets and then reviewing the concept of a metric space, which provides a rich source of examples of topological spaces.

Guide for the Reader

Our expositional choices in this chapter (and in this part of the book more broadly) are motivated by our belief that in order to safely use mathematical tools, it is important to understand where they come from and how they fit into a broader ideological context. As a consequence, we have not adopted the maximally streamlined approach (which might start directly with simplicial complexes) to mathematical background. Instead, we have endeavored to "start from the beginning," and give a rapid but thorough introduction to the ideas of algebraic topology.

On the other hand, we are aware that the volume of material below might pose challenges to the energy of readers who have less math background. For someone interested in a minimal path through this section, we might recommend skipping to Section 1.8 and reading prior material as necessary to proceed. Strictly speaking, only Sections 1.8 through 1.12 are required for the rest of the book. Nonetheless, we hope that there are some readers from biology who find the broader introductory material useful.

1.1 Sets

All of the mathematical objects we will study herein are built on top of sets. Although the construction of rigorous axiomatizations of set theory is subtle and complicated, we can get by with a fairly naive view of the foundations. An excellent textbook that covers the material we use (and more) is Halmos' *Naive Set Theory* [224].

We will regard a *set* as simply an unordered collection of objects, referred to as *members* or *elements*. We require that the elements of a set be unique. A *finite* set has finitely many elements; otherwise, the set is *infinite*.

Example 1.1.1.

1. The empty set, denoted $\emptyset$, is the set with no elements.
2. The integers $\mathbb{Z}$ is the set $\{\ldots,-5,-4,-3,-2,-1,0,1,2,3,4,5,\ldots\}$; an element of $\mathbb{Z}$ is a number. Similarly, the natural numbers $\mathbb{N} = \{0,1,2,\ldots\}$, the rational numbers $\mathbb{Q}$ consisting of the fractions $\{\frac{p}{q}\}$ where p and q are relatively prime, and the real numbers $\mathbb{R}$ are sets. Note that rigorously constructing the real numbers as a set is complicated; although informally we are used to working with them as decimals, the construction requires some machinery we will discuss below.
3. The Euclidean vector spaces $\mathbb{R}^n$ are sets; the elements are the vectors $(x_1, x_2, \ldots, x_n)$, where each $x_i \in \mathbb{R}$.
4. The collection of possible bases in a DNA strand, $\{A, G, C, T\}$, is a set.
5. The expression vectors from a collection of samples from a cancerous tumor form a set, e.g., a set of vectors $\{(30,50,10,\ldots),(10,16,29,\ldots),\ldots\}$, where each element is a vector and each entry in an element of the set is an expression value at a particular position on a gene.
6. In general, a finite set can be specified as a list of elements, e.g., $\{a, b, 4\}$, which has elements a, b, and 4. These elements could be specified by a condition, e.g., the set of people named "Harold" in New York.

7. In contrast, “tall people in Boston” does not describe a set; the term “tall” is not an adequately specific description by itself. On the other hand “living people over six feet tall in Boston” is a well-defined characterization of a set.

There are several familiar constructions of new sets from old that will be of particular relevance for our work. First, given a set X, we can form new sets by taking only certain elements from X; we have seen examples of this above.

Definition 1.1.2. A *subset* Y of a set X is a set Y such that every element $y \in Y$ is an element of X. We write $Y \subseteq X$ to denote a subset of X.

Second, given a finite set of sets $\{X_i\} = \{X_1, X_2, \ldots, X_k\}$, we can form the set of tuples.

Definition 1.1.3. Let $\{X_i\}$ be a finite set of sets. The *Cartesian product* is the set specified as

$$\prod_i X_i = \{(x_1, x_2, \ldots, x_k) \mid x_i \in X_i\}.$$

Example 1.1.4.

1. Almost by definition, the standard xy-plane $\mathbb{R}^2$ can be identified with the product $\mathbb{R} \times \mathbb{R}$,
2. and more generally

$$\mathbb{R}^n \cong \prod_{i=1}^n \mathbb{R}.$$

Given two sets X and Y, we can form the *union*

$$X \cup Y = \{z \mid z \in X \text{ or } z \in Y\}$$

and *intersection*

$$X \cap Y = \{z \mid z \in X \text{ and } z \in Y\}.$$

More generally, for a collection $\{X_i\}$ of sets we can form the union $\cup_i X_i$ or intersection $\cap_i X_i$ of all of them.

If S_1 and S_2 are sets, a function $f\colon S_1 \to S_2$ is a rule that produces an element of S_2 for each element of S_1. We often refer to functions between sets as *maps* or *maps of sets*. Given two maps $f\colon X \to Y$ and $g\colon Y \to Z$, the composite $g \circ f$ takes $x \in X$ to $g(f(x)) \in Z$.

Definition 1.1.5. A map of sets $f\colon X \to Y$ is defined as follows.

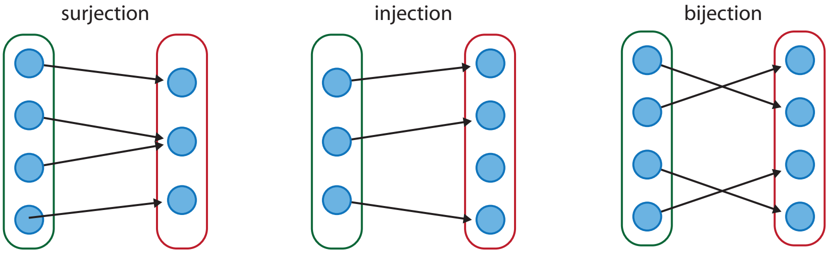

Figure 1.1 A surjective map hits everything. Injective maps take distinct elements to distinct places. Bijective maps are both injective and surjective.

- *Surjective* or *onto* if for every point $y \in Y$, there is at least one $x \in X$ such that $f(x) = y$; that is, f hits all the points of Y.
- *Injective* or *one-to-one* if for any two points $x, y \in X$ that are not the same, $f(x) \neq f(y)$; that is, no point of Y is hit more than once.
- *Bijective* if it is injective and surjective.

See Figure 1.1 for an illustration of these three properties of a map of sets.

Example 1.1.6.

1. The map $f\colon \mathbb{R} \to \mathbb{R}$ specified by $f(x) = x^2$ is not injective, since -2 and 2 both go to 4, and it is not surjective, since no negative numbers are hit.
2. The map $\{a, b, c\} \to \{d\}$ that takes every element to d is not injective since a and b both go to d, but it is surjective.

It is extremely useful to develop a criterion for considering sets “the same” that is weaker than requiring that they be identical. For this, we introduce the notion of the inverse of a function. Recall that the identity map $\mathrm{id}_X\colon X \to X$ is simply the function defined to be $f(x) = x$.

Definition 1.1.7. The function $f\colon X \to Y$ has *inverse* $g\colon Y \to X$ if the composite $g \circ f\colon X \to Y \to X$ is id_X, the identity on X, and the composite $f \circ g\colon Y \to X \to Y$ is id_Y, the identity on Y.

Notice that any bijective map $f\colon X \to Y$ has an inverse $f^{-1} : Y \to X$ where $f^{-1}(y)$ is defined to be the unique element of X that has image y.

Definition 1.1.8. Two sets X and Y are *isomorphic* if there exists a bijection $f\colon X \to Y$. We often write $X \cong Y$ and leave the functions f and g implicit.

We will refer to f as an *isomorphism*.

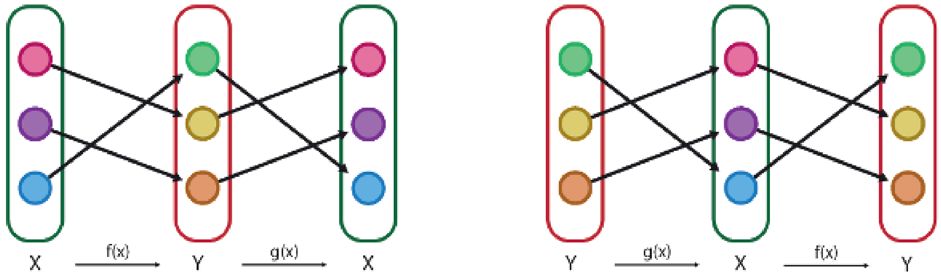

Figure 1.2 An isomorphism between two finite sets can be described in terms of a permutation.

Example 1.1.9.

1. Two finite sets X and Y are isomorphic if and only if they have the same number of elements. See Figure 1.2 for an example of an isomorphism between two finite sets.
2. The map $\mathbb{R}^2 \to \mathbb{R}^2$ that takes (x, y) to $(-x, -y)$ is an isomorphism; its inverse is itself.
3. The map $\mathbb{N} \to \mathbb{Z}$ that takes 0 to 0, 1 to 1, 2 to -1, and in general is specified by the formula

$$f(x) = \begin{cases} 0, & x = 0 \\ \frac{x+1}{2}, & x \text{ odd} \\ -\frac{x}{2}, & x \text{ even} \end{cases}$$

   is an isomorphism.

Elaborating on the observation that finite sets are isomorphic if and only if they have the same number of elements, we can use isomorphisms to talk about the size of infinite sets.

Definition 1.1.10. A set S is *countable* if there exists a bijection $f\colon \mathbb{N} \to S$, where $\mathbb{N}$ denotes the natural numbers $\{0, 1, 2, \ldots\}$.

Countable sets are the smallest kind of infinite sets.

Example 1.1.11.

1. Clearly the set of natural numbers $\mathbb{N}$ is countable. The set of integers $\mathbb{Z}$ is also countable, by the bijection given above in Example 1.1.9.
2. A little bit of work shows that the set of rational numbers $\mathbb{Q}$ is countable.

3. Famously, Cantor showed that $\mathbb{R}$ is uncountable, which means that it is *bigger* in a precise sense than any countable set. More generally, $\mathbb{R}^n$ is uncountable for any $n > 0$.

Two sets can be isomorphic in many different ways; for example, there are many isomorphisms between any two finite sets of the same size. In general, composing an isomorphism between two different sets X and Y with an isomorphism from Y to itself will produce a new isomorphism from X to Y.

We will often want to work with sets "up to isomorphism." Formally, we do this using the fact that isomorphism of sets is an *equivalence relation*.

Definition 1.1.12. Let S be a set and let $\sim$ be a relation on S, i.e., a collection of tuples (x, y) with $x, y \in S$. Given such a tuple (x, y), we write $x \sim y$. Then $\sim$ is an *equivalence relation* when the following holds.

1. For all $x, y \in S$, if $x \sim y$ then $y \sim x$.
2. For all $x \in S$ we have $x \sim x$.
3. For all $x, y, z \in S$, if $x \sim y$ and $y \sim z$, then $x \sim z$.

Isomorphism of sets clearly satisfies these properties. The collection of all sets isomorphic to X is called the *isomorphism class* of X; often we will be interested in a set only up to its isomorphism class. However, we have to be a little bit careful when formalizing the idea of an isomorphism class; the isomorphism class of a set is usually not itself a set! Instead, it is a larger object, referred to as a class. The issue is that Russell's paradox shows that the "set of all sets" cannot exist: the set of all sets would have to contain in particular the set that does not contain itself as an element, and this is a contradiction. The paradox rules out certain appealing but naive axioms about which sets can exist: in particular, certain constructions that intuitively seem like they should produce sets in fact do not, but rather produce larger objects.

1.2 Metric Spaces

It is very common to represent experimental data as a set of measurements, together with a distance between every pair of measurements. For example, genomic expression data is often presented as a collection of arrays of the form $\{x_1, x_2, \ldots, x_k\}$, where $x_i \in \mathbb{R}$ is a number representing the expression of the ith measured gene. The distance between two expression vectors could be the standard Euclidean distance or it could be a correlation function, depending on the specific situation. Mathematically, this kind of setup is captured by the notion of a *metric space*. There are many good treatments of metric spaces; Kaplansky's *Set Theory and Metric Spaces* is a particularly accessible elementary treatment [284].

A metric space is a set X equipped with a distance function, referred to as the metric, that satisfies a few simple axioms encapsulating the salient features of the usual Euclidean distance in $\mathbb{R}^n$. Specifically, we have the following definition.

Definition 1.2.1. A *metric space* is specified by a pair (X, ∂_X) where X is a set and ∂_X is a function

$$\partial_X \colon X \times X \to \mathbb{R}$$

that assigns a non-negative real number to each pair of points of X such that the following holds.

1. The metric ∂_X detects whether two points are the same, in the sense that

$$\partial_X(x_1, x_2) = 0 \iff x_1 = x_2.$$

2. The metric ∂_X is *symmetric* in that

$$\forall x, y \in X, \qquad \partial_X(x, y) = \partial_X(y, x).$$

3. The metric ∂_X satisfies the *triangle inequality*:

$$\forall x, y, z \in X, \qquad \partial_X(x, z) \leq \partial_X(x, y) + \partial_X(y, z).$$

The most interesting of these axioms is the triangle inequality. See Figure 1.3 for pictures of triangles on the surface of a cylinder and a sphere; the triangle inequality is evident. (Here the metric on these surfaces is computed by the length of shortest path.)

Remark 1.2.2. Particularly in biological applications, we sometimes encounter *dissimilarity measures* which are not quite metrics. For example, the Kullback-Leibler divergence (see Remark 3.2.32) is not symmetric, the Gromov-Hausdorff distance (see Definition 2.4.4) on the set of metric spaces can be zero for metric

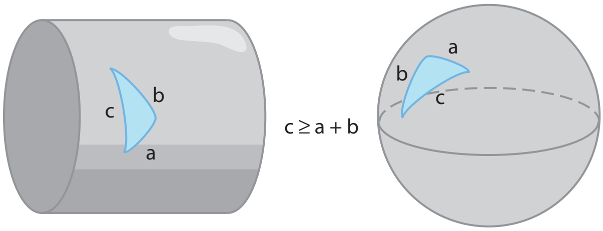

Figure 1.3 No matter how curved or distorted triangles in a metric space are, the length of any side must always be shorter than the sum of the other two sides.

spaces that are not identical, and many common dissimilarity measures (e.g., the Bray-Curtis dissimilarity measure [70]) do not satisfy the triangle inequality. In the first two kinds of examples, it is easy to construct a metric that captures the salient properties of the dissimilarity function – for instance, by symmetrizing (making a new metric $\partial'_X = \min(\partial_X(x, y), \partial_X(y, x))$) or identifying points such that $\partial_X(x, y) = 0$ when $x \neq y$. Fixing triangle inequality violations is more subtle (e.g., see [196] for interesting recent progress).

Example 1.2.3. The most familiar and important examples of metric spaces are the Euclidean spaces $\mathbb{R}^n$; these are defined as the n-tuples $\{(x_1, x_2, \dots, x_n) \mid x_i \in \mathbb{R}\}$ equipped with the standard distance metric

$$\partial_{\mathbb{R}^n}((x_1, x_2, \dots, x_n), (y_1, y_2, \dots, y_n)) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \dots + (x_n - y_n)^2}.$$

A natural family of examples of metric spaces come from metrics induced by weighted graphs. Particularly interesting examples of graph metrics come from trees with weighted edges; this kind of metric space will be important in work on modeling evolutionary phenomena using phylogenetic trees, as we will see in Section 5.2.

Example 1.2.4. A *graph* is specified by a set of vertices and a set of edges connecting the vertices. A weighted graph has weights (nonnegative numbers) attached to the edges. More precisely, a weighted graph is a tuple $G = (V, E, W)$ with vertex set V, edge set $E \subset V \times V$, and weights $W \colon E \to \mathbb{R}^{\geq 0}$.

Regarding this graph as undirected and stipulating that there are no edges with non-zero weight from any vertex v to itself, the graph metric on a weighted graph is a metric on the set of vertices of the graph. The metric is defined so that the distance between vertices v and w is the minimal length of a path connecting v and w:

$$\partial_G(v, w) = \min_{v, z_0, z_1, \dots, z_k, w \mid z_i \in V} \left(W(v, z_0) + \sum_{i=0}^{k-1} W(z_i, z_{i+1}) + W(z_k, w) \right).$$

(See Figure 1.4.)

The metrics we have described so far are continuous, in the sense that distances can in principle be any real number. But many interesting metrics are discrete. For example, the *Hamming distance*, which is a metric on strings that counts the number of differences, takes values in the natural numbers. The Hamming distance is a basic concept in information and coding theory.

Example 1.2.5. Fix an alphabet Σ, i.e., a set of symbols we will call letters. Let x and y be words of length n with letters in Σ. Then the *Hamming distance* between x and y is defined to be the number of positions at which the letters of x and y differ:

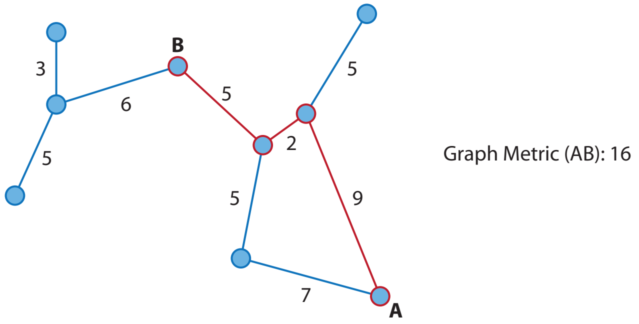

Figure 1.4 The length of the shortest path between A and B in the weighted graph gives the distance between them.

$$\partial_H(x, y) = \#\{i \mid x_i \neq y_i\}.$$

For example, if $\Sigma = \{A, C, G, T\}$, then

$$\partial_H(ACGT, ACAA) = 2.$$

An important point to emphasize is that there can be many distinct metrics on the same underlying set. For instance, in genomic data considered as words in $\{A, G, C, T\}$ there are, in addition to the Hamming distance, other well-motivated biologically relevant distances (see Section 5.2). As another example, a common distance metric used for gene expression data represented as points in $\mathbb{R}^n$ is the Pearson correlation distance.

Example 1.2.6. For $x, y \in \mathbb{R}^n$, define the *Pearson correlation distance* between x and y to be

$$\partial_{\mathrm{cor}}(x, y) = 1 - \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2(y_i - \bar{y})^2}},$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.

The existence of a distance function allows us to define many familiar notions from calculus; we review these now, as this is the prototype for the definitions of elementary topology. For instance, for each point x in a metric space and $\epsilon > 0$, we can specify the ϵ-neighborhoods of x to describe points that are close to x. Specifically, we have the *open balls* and *closed balls*

$$B_\epsilon(x) = \{z \in X \mid \partial_X(z, x) < \epsilon\} \qquad \text{and} \qquad \bar{B}_\epsilon(x) = \{z \in X \mid \partial_X(z, x) \leq \epsilon\}.$$

We can always *separate* two distinct points x and y by taking a ball B_1 around x and a ball B_2 around y such that $B_1 \cap B_2 = \emptyset$; if $\partial_X(x, y) = \epsilon$, we can set $B_1 = B_{\frac{\epsilon}{4}}(x)$ and $B_2 = B_{\frac{\epsilon}{4}}(y)$, for example. (See Figure 1.5.)

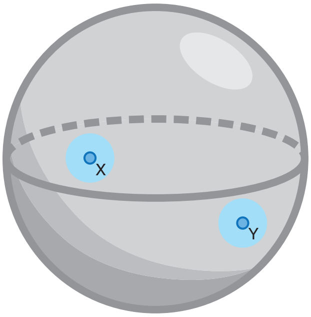

Figure 1.5 Any pair of distinct points in a metric space can be separated by open balls around them.

Elaborating on this, the existence of a metric allows us to talk about convergence of sequences. A sequence of points in X will be a function $\mathbb{N} \to X$, i.e., a sequence

$$\{x_i\} = x_0, x_1, x_2, x_3, \ldots$$

for $x_i \in X$.

Definition 1.2.7. For a metric space (X, ∂_X), an infinite sequence of points $\{x_i\}$ *converges* to a point $x \in X$ if for any $\epsilon > 0$, there exists a positive integer N such that $\partial_X(x_k, x) < \epsilon$ for all $k > N$.

Informally speaking, the definition of convergence simply means that if we go out far enough in the sequence, all the points are arbitrarily close to x. (See Figure 1.6 for a picture of what this means.)

Example 1.2.8. Consider the sequence

$$\left\{\frac{1}{n}\right\} = 1, \frac{1}{2}, \frac{1}{3}, \ldots, \frac{1}{100}, \ldots.$$

This sequence converges to 0; for any ϵ, it is clear that we can find an N such that for $n > N$,

$$\left|\frac{1}{n} - 0\right| = \frac{1}{n} < \epsilon.$$

Specifically, take N to be the smallest integer larger than $\frac{1}{\epsilon}$.

A more subtle notion is that of a *Cauchy sequence*; this is a sequence of points that *ought* to converge somewhere, in the following sense.

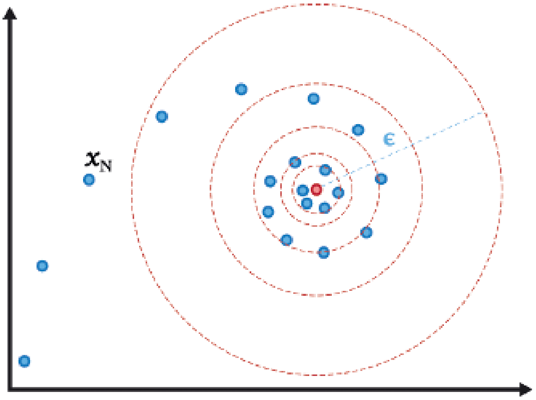

Figure 1.6 For any ball around the point of convergence, all but finitely many points of the convergent sequence are within that ball. (Note that in the picture there are only finitely many points, due to limits of resolution.)

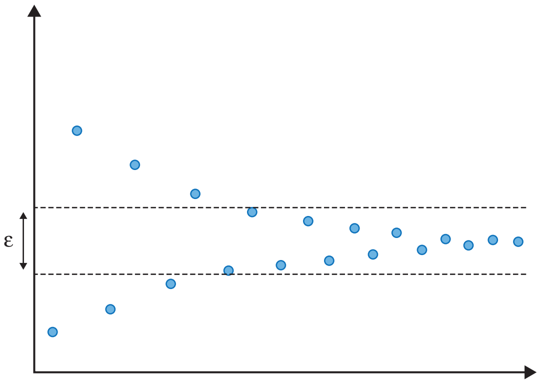

Figure 1.7 The points in a Cauchy sequence get closer and closer together but need not converge.

Definition 1.2.9. For a metric space (X, ∂_X), a *Cauchy sequence* is a sequence of points $\{x_i\}$ such that for all $\epsilon > 0$, there exists an N such that $\partial_X(x_j, x_k) < \epsilon$ for $j, k > N$.

Although the points in a Cauchy sequence get closer and closer together (see Figure 1.7), it is not necessarily the case that all Cauchy sequences converge to a point $x \in X$.

Example 1.2.10. Consider the set of rational numbers $\mathbb{Q}$ equipped with the standard metric, i.e., the distance between x and y is $\partial(x, y) = |x - y|$. Then the sequence

$$\{3, 3.1, 3.14, 3.141, 3.1415 \ldots\}$$

(where each new number in the sequence has an additional digit of π) is a Cauchy sequence and "wants" to converge to π, but π is not in $\mathbb{Q}$!

This possible failure of Cauchy sequences to coincide with convergent sequences motivates the following definition.

Definition 1.2.11. A metric space (X, ∂_X) is *complete* if every Cauchy sequence converges to a point $x \in X$.

Example 1.2.12. The Euclidean spaces $\mathbb{R}^n$ are all complete; $\mathbb{R}$ can in fact be constructed by formally adding to $\mathbb{Q}$ points for each Cauchy sequence to converge to.

As Example 1.2.12 indicates, there is a tension between the size of a metric space and whether it is complete; $\mathbb{Q}$ is countable but not complete. Adding points to $\mathbb{Q}$ to make it complete yields $\mathbb{R}$, which is uncountable. Although metric spaces of interest are often not countable, there is frequently a countable subset $X' \subset X$ that is dense, in the following sense.

Definition 1.2.13. A subset $X' \subset X$ is *dense* if for all $x \in X$ and $\epsilon > 0$ there exists a point $z \in X'$ such that $\partial_X(x, z) < \epsilon$. That is, for any point X, there exists an arbitrarily close approximation in X'.

For example, $\mathbb{Q}$ is dense in $\mathbb{R}$; any real number can be approximated to any precision by a finite-length decimal.

Definition 1.2.14. A metric space (X, ∂_X) is *separable* if there exists a countable subset $X' \subset X$ that is dense in X.

Example 1.2.15. All of the Euclidean spaces $\mathbb{R}^n$ are separable; any point can be approximated by a point with rational coordinates.

A closely related notion is the idea of an ϵ-net (Figure 1.8).

Definition 1.2.16. Let (X, ∂_X) be a metric space. A subset $X' \subset X$ is *ϵ-dense* if for every $x \in X$ there exists $z \in X'$ such that $\partial_X(x, z') < \epsilon$. (So a dense set is ϵ-dense for every ϵ.) An *ϵ-net* is a subset $X' \subset X$ that is ϵ-dense.

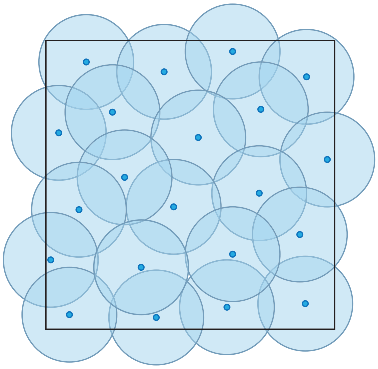

Figure 1.8 Any point in the square is within ϵ of the blue points at the centers of the circles.

In order to understand when ϵ-nets exist, we need to have ways to talk about the size of a metric space. In order to define the size, we first need to review the notion of inf and sup.

Definition 1.2.17. Given a subset $A \subset \mathbb{R}$, a *lower bound* for A is an element $x \in \mathbb{R}$ such that for all $a \in A$, $x \leq a$. Then the *infimum* $\inf(A)$ is the greatest lower bound, if one exists. Similarly, an *upper bound* for A is an element $y \in \mathbb{R}$ such that for all $a \in A$, $a \leq y$. Then the *supremum* $\sup(A)$ is the least upper bound, if one exists.

The sup and inf are distinct from the max and min, respectively, because they might not lie in A itself.

Definition 1.2.18. Let (X, ∂_X) be a metric space. The *diameter* of a subset $A \subset X$ is the supremum

$$\sup_{x,y \in X} \partial_X(x, y).$$

We must write sup rather than max because there might not be any pair of points which realizes the bound. (Note also that the diameter can be ∞, when there is no upper bound!)

Another way to talk about this is to observe that a subset $A \subset X$ has finite diameter when there exists $a \in A$ such that $A \subset B_\kappa(a)$ for some κ; more generally, such a set will be referred to as bounded. (See Figure 1.9 for an example of this.)

An even stronger notion is that of being *totally bounded*.

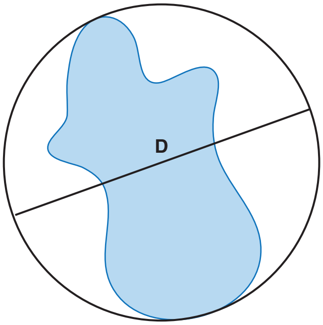

Figure 1.9 The diameter of a subset of a metric space can be approximated by taking a ball that completely encloses the subset.

Definition 1.2.19. Let (X, ∂_X) be a metric space. Then X is *totally bounded* if for every $\epsilon > 0$, there exists a finite cover of X by balls of radius ϵ, i.e., a collection of balls $\{B_\epsilon(x_i)\}$ whose union is X.

In $\mathbb{R}^n$ a subset is bounded if and only if it is totally bounded, but in general, a bounded space need not be totally bounded. For example, a metric space with infinitely many points such that all interpoint distances are 1 is bounded but not totally bounded.

Lemma 1.2.20. *Let (X, ∂_X) be a totally bounded metric space. Then for any ϵ we can find a finite ϵ-net in X.*

An important theme in modern mathematics is that the structure of mathematical objects (e.g., sets or metric spaces) can be completely understood in terms of functions between them. We describe a framework that allows us to be precise about this in Section 1.7 below (where we introduce basic concepts of category theory). From this perspective, an essential next step is to define a function between metric spaces.

At a minimum, a map between metric spaces (X, ∂_X) and (Y, ∂_Y) should involve a function of sets $f \colon X \to Y$. But we would like to require that the function also respect the metric structures on X and Y, in some sense. There are different ways to do this; we now discuss the familiar notion of a *continuous map*.

Definition 1.2.21. Let (X, ∂_X) and (Y, ∂_Y) be metric spaces. A map $f \colon X \to Y$ is *continuous* if for every sequence $\{x_i\}$ in X converging to x the sequence $\{f(x_i)\}$ converges in Y to $f(x)$.

An important property of continuous maps is that they compose.

Lemma 1.2.22. *Let (X, ∂_X), (Y, ∂_Y), and (Z, ∂_Z) be metric spaces. If $f\colon X \to Y$ and $g\colon Y \to Z$ are continuous, then so is the composition $g \circ f\colon X \to Z$.*

Continuity can also be defined in terms of a traditional ϵ-δ definition; this is easy to show directly. We explain this below in Example 1.3.20, where we generalize the notion of continuity to topological spaces.

For metric spaces, it is also sometimes useful to consider a stronger notion of continuous where the "expansion" of the map is bounded.

Definition 1.2.23. A map $f\colon X \to Y$ between metric spaces (X, ∂_X) and (Y, ∂_Y) is *Lipschitz* with constant κ if for all $x_1, x_2 \in X$ the inequality

$$\partial_Y(f(x_1), f(x_2)) \leq \kappa \partial_X(x_1, x_2)$$

holds.

Any Lipschitz map is continuous, but the converse does not hold in general.

1.3 Topological Spaces

The motivating idea of point-set topology is to relax the requirement of a distance and define a weaker and more flexible notion of *closeness* that still allows us to formalize the notions that lead to calculus (i.e., continuity and convergence). This is the basis for elementary analysis, which studies the foundations of calculus. A classic textbook for point-set topology is Munkres [369]; there are many excellent analysis books, of which Rudin [440] is a canonical example.

The basic observation that leads to the development of point-set topology is that most of the concepts we defined for metric spaces in Section 1.2 were or could be phrased in terms of the metric balls $B_\epsilon(x)$. A topological space can be thought of as simply a set with a well-behaved collection of subsets that act like metric balls. This abstraction is extremely useful, for a number of reasons: many metrics can lead to the same topology, some important topological spaces (notably those arising in algebraic geometry) do not come from a metric, and many basic constructions (e.g., gluing) are much more complicated to express in the context of a metric.

Definition 1.3.1. A *topological space* is a pair $(X, \mathcal{U})$, where X is a set and $\mathcal{U}$ is a collection of subsets of X, which we refer to as *open sets*. The open sets satisfy the following conditions.

1. Both the empty subset $\emptyset$ and X are elements of $\mathcal{U}$.
2. Any union of elements of $\mathcal{U}$ is an element of $\mathcal{U}$.
3. The intersection of a finite collection of elements of $\mathcal{U}$ is an element of $\mathcal{U}$.

A subset $Z \subseteq X$ is *closed* if the complement of Z in X is open.

Any metric space gives rise to a topological space.

Example 1.3.2. Let (X, ∂_X) be a metric space. Then we say that a subset $A \subset X$ is *open* if for every $z \in A$, there exists ϵ such that $B_\epsilon(z) \subseteq A$. The open sets make X into a topological space.

But the definition of a topological space is sufficiently flexible so as to allow a variety of strange examples. For instance, any set has two trivial topologies.

Example 1.3.3. Any set X can be given the following two topologies.

1. The *discrete topology*, in which any subset $Y \subset X$ is an open set. In particular, the points themselves are open sets. As the name suggests, in this topology the points should be thought of as maximally separated from one another.
2. The *indiscrete topology*, in which the only open sets are the entire set X and $\emptyset$. In this topology, the points should be thought of as being arbitrarily close to each other.

However, the most frequently occurring examples are very familiar. In order to specify a topological space, one typically gives a *base* for the topology.

Definition 1.3.4. A *base* for a topological space $(X,\mathcal{U})$ is a collection of open sets $\{U_\alpha\}$ such that any open set is a union of elements of the base. Given simply a set X, a collection of sets $\{B_\alpha\}$ is a base if every $x \in X$ is in some B_α and given $x \in B_\alpha \cap B_\beta$, there exists $B_\gamma \subseteq B_\alpha \cap B_\beta$ such that $x \in B_\gamma$.

The importance of the intrinsic definition is that we can define a topology on X given a base.

Lemma 1.3.5. *Given a set X and a base $\{B_\alpha\}$, we can define a topology on X where a set U is open if for each $x \in U$ there exists B_α such that $x \in B_\alpha \subseteq X$. (And we will often refer to this as the topology generated by a base.)*

As Lemma 1.3.5 makes clear, the base of a topological space is modeled on the open balls of a metric space.

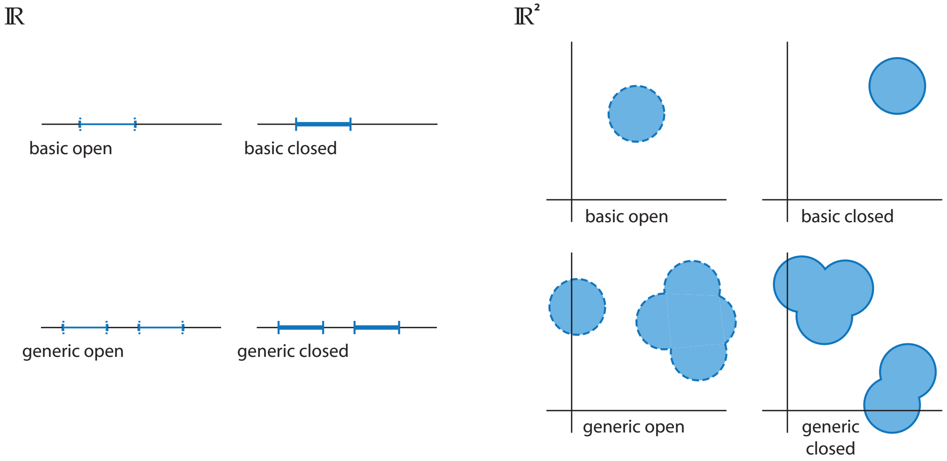

Figure 1.10 Basic open and closed sets in Euclidean space are *balls*; more general open and closed sets are generated by union and intersection.

Example 1.3.6. Euclidean space with the topology generated by the open balls $B_\epsilon(x) = \{y \in \mathbb{R}^n \mid ||x - y|| < \epsilon\}$, for $x \in \mathbb{R}^n$ and $\epsilon > 0$. See Figure 1.10 for some examples of open and closed sets in this topology.

Example 1.3.7. In fact, we can conveniently describe the topology of Example 1.3.2 on a metric space (X, ∂_X) as generated by the base of the open balls $B_\epsilon(x) = \{y \in X \mid \partial_X(y, x) < \epsilon\}$, for $x \in X$ and $\epsilon > 0$.

An important class of topological spaces are those with a countable base; these are called *second countable*. Example 1.3.6 is a second countable topological space; we can take the base using only the balls with rational radii.

The example of the topology induced by a metric has a particularly important property that we now highlight. Specifically, recall that in a metric space we can separate points in the sense that given two distinct points $x, y \in X$, we can choose balls $B_{\epsilon_1}(x)$ and $B_{\epsilon_2}(y)$ such that $B_{\epsilon_1}(x) \cap B_{\epsilon_2}(y) = \emptyset$; we simply take $\epsilon_1, \epsilon_2 < \frac{\partial_X(x,y)}{2}$. It turns out to be very useful to consider topological spaces that have this property, even if the topology is not generated by a metric.

Definition 1.3.8. A topological space $(X, \mathcal{U})$ is *Hausdorff* if for any pair of distinct points $x, y \in X$ there exist open sets U_x and U_y such that $x \in U_x$, $y \in U_y$, and $U_x \cap U_y = \emptyset$.

If $(X, \mathcal{U})$ is a topological space, any subset $Y \subset X$ can be given the structure of a topological space in a natural fashion induced from the topology on X. This

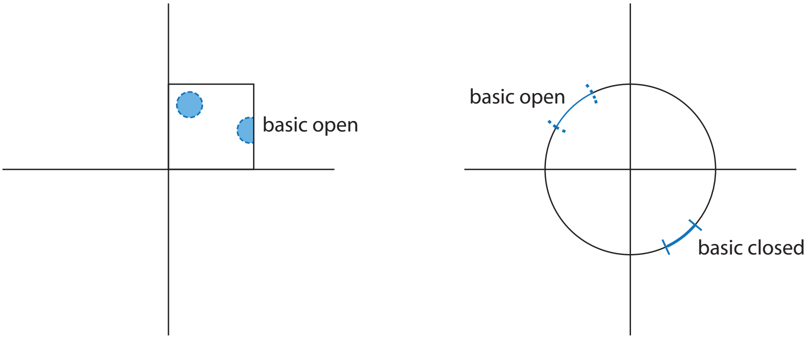

Figure 1.11 Left: Basic open sets in the subspace topology on the unit square. Right: Basic open sets in the subspace topology on the unit circle.

is referred to as the *subspace topology* on Y, and is a very important source of examples of topological spaces.

Definition 1.3.9. Let $(X, \mathcal{U})$ be a topological space and $Y \subset X$ a subset. Then the *subspace topology* on Y is defined by taking the open sets to be $\{Y \cap U \mid U \in \mathcal{U}\}$.

Example 1.3.10. The subspace topology on the unit square $[0,1]\times[0,1] \subset \mathbb{R}^2$ has basic open sets that are either balls (when the ball is completely contained within the square) or the intersection of balls with the square; see Figure 1.11 for examples.

Example 1.3.11. Let S^1 denote the standard unit circle in $\mathbb{R}^2$; that is, $S^1 = \{(x,y) \subset \mathbb{R}^2 \mid x^2 + y^2 = 1\}$. We topologize S^1 using the subspace topology as a subset of $\mathbb{R}^2$; see Figure 1.11 for examples.

Just as with sets, another standard way to produce new topological spaces from old is via the Cartesian product (recall Definition 1.1.3).

Definition 1.3.12. Let X and Y be topological spaces, with the topologies specified by open sets $\{U_\alpha\}$ and $\{V_\beta\}$ respectively. Then the product

$$X \times Y = \{(x,y) \mid x \in X, y \in Y\}$$

is a topological space with a base for the topology given by the open sets $\{U_\alpha \times V_\beta\}$; we refer to this as the *product topology*.

A topological space is designed to be a minimal structure in which we can talk about "closeness," in a precise sense.

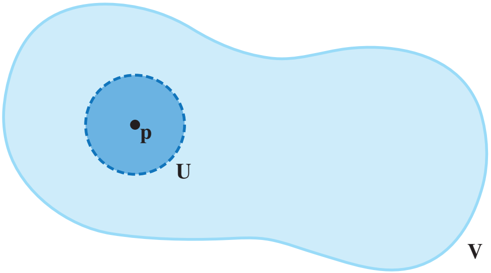

Figure 1.12 V is a neighborhood of p.

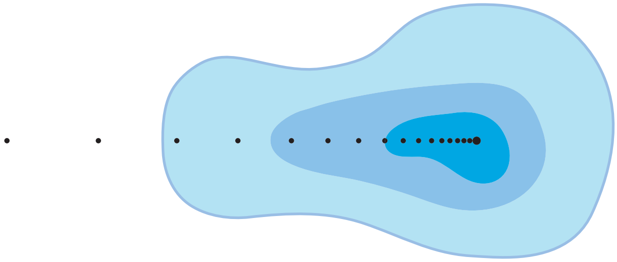

Figure 1.13 Smaller and smaller open sets around the point of convergence still contain all but finitely many points in the approaching sequence.

Definition 1.3.13. Given a point $x \in X$, we define a *neighborhood* of x to be a set $V \subseteq X$ such that there is an open set $U \subseteq V$ and $x \in U$. (See Figure 1.12.)

Immediately, we can use this definition to specify the notion of convergence of a sequence (Figure 1.13).

Definition 1.3.14. A sequence of points $\{x_i\}$ *converges* to p if for any neighborhood V of p there exists an N such that $x_n \in V$ for $n \geq N$.

Considering Example 1.3.6, we see that in Euclidean space this means that for any ϵ, there exists an n such that $x_n \in B_\epsilon(x)$, i.e., $\|x_n - x\| < \epsilon$. In particular, when restricted to $\mathbb{R}$, the definition recovers the usual notion from elementary calculus of convergence of a sequence. More generally, Definition 1.3.14 coincides with Definition 1.2.7 in a metric space given the metric topology.

Topological spaces also admit an extremely useful notion of size. This takes a little bit more work to specify without explicit reference to a distance function. In order to define this, we need the notion of a cover.

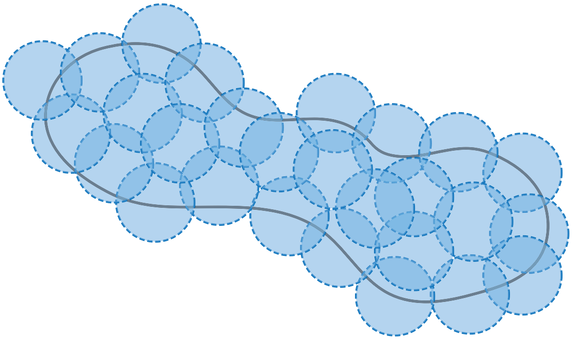

Figure 1.14 An open cover of U is a collection of open sets whose union contains U.

Definition 1.3.15. An *open cover* of a set U in a topological space X is a collection of open sets $\{U_\alpha\}$, with each $U_\alpha \subset X$, such that $U \subseteq \bigcup U_\alpha$.

For example, the collection of all balls $B_\epsilon(x)$ as x varies over the points of $\mathbb{R}^n$ is an open cover of $\mathbb{R}^n$. (See Figure 1.14.) A subcover of a cover is a subset whose union still contains U.

Definition 1.3.16. A topological space X is *compact* if any open cover of X has a finite subcover.

Example 1.3.17.

1. Every finite set is compact.
2. The sphere $\{x, y, z \,|\, x^2 + y^2 + z^2 = 1\}$ with the subspace topology is compact.
3. No Euclidean space $\mathbb{R}^n$ is compact for $n > 0$.

Compact sets are "small" in a basic sense. The notion of compactness is a way of formalizing the properties of the closed and bounded subsets of $\mathbb{R}^n$.

Theorem 1.3.18. *A subset $X \subseteq \mathbb{R}^n$ regarded as a metric space is compact if and only if it is closed and bounded.*

1.3.1 Maps between Topological Spaces

We now turn to consider the correct notion of a map between topological spaces. We want to restrict ourselves to maps $f: X \to Y$ which satisfy certain properties expressing compatibility with the topologies on X and Y. Roughly speaking, we want continuous maps to have the property that "nearby" points in X are taken to "nearby" points in Y.

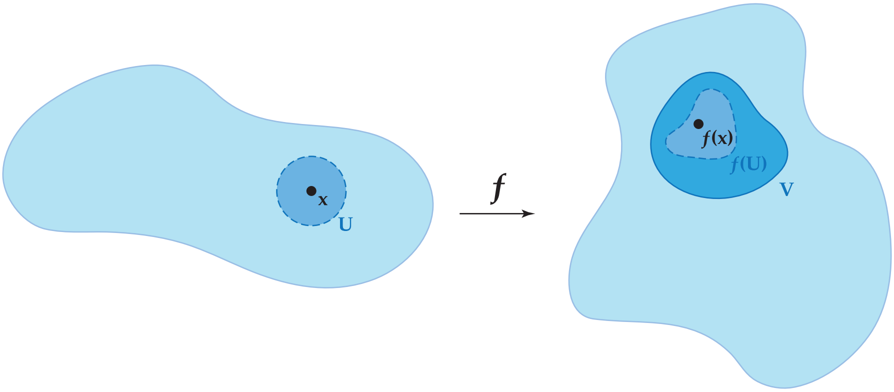

Figure 1.15 A function is continuous if for every neighborhood V around $f(x)$, we can find a neighborhood U of x whose image $f(U)$ sits inside it.

Definition 1.3.19. Let $(X, \mathcal{U}_X)$ and $(Y, \mathcal{U}_Y)$ be topological spaces. A map $f: X \to Y$ is *continuous at a point* x if for every neighborhood V of $f(x)$, there exists a neighborhood U of x such that $f(U) \subseteq V$ (Figure 1.15). The map f is *continuous* if it is continuous at every point $x \in X$.

It is instructive to work out exactly what this means in the case of the standard metric topology on $\mathbb{R}$.

Example 1.3.20. A map $f: \mathbb{R} \to \mathbb{R}$ is continuous at a point $x \in \mathbb{R}$ if for every open ball $B_\epsilon(f(x))$, there exists an open ball $B_\delta(x)$ such that $f(B_\delta(x)) \subseteq B_\epsilon(f(x))$. Put another way, for every $\epsilon > 0$, there exists $\delta > 0$ such that $|x - y| < \delta$ implies that $|f(x) - f(y)| < \epsilon$. That is, we have recovered precisely the usual ϵ-δ notion of continuity.

More generally, in any metric space, maps are continuous in the sense of Definition 1.2.21 if and only if they are continuous in the sense of Definition 1.3.19.

Generalizing Lemma 1.2.22, the composition of continuous maps is continuous.

Lemma 1.3.21. *Let $(X, \mathcal{U}_X)$, $(Y, \mathcal{U}_Y)$, and $(Z, \mathcal{U}_Z)$ be topological spaces and suppose we have continuous maps $f: X \to Y$ and $g: Y \to Z$. Then the composite $g \circ f: X \to Z$ is continuous.*

Continuous maps out of simple "test spaces" that are well understood play an important role in algebraic topology; for example, we can now define a path in terms of maps out of the unit interval.

Definition 1.3.22. A *path* from x to y in a topological space $(X, \mathcal{U}_X)$ is a continuous function $\gamma\colon [0,1] \to X$ such that $\gamma(0) = x$ and $\gamma(1) = y$. Here $[0,1]$ is given the subspace topology it inherits as a subset of $\mathbb{R}$.

The notion of a path captures many familiar examples, but the price of the generality is that strange examples are also permitted.

Example 1.3.23.

1. A path γ in $\mathbb{R}^n$ is just a curve that could be drawn without lifting up the pen (see Figure 1.16). Note that these can be surprisingly complicated: there are famous examples of "space-filling" curves, which are precisely paths that touch every point of $\mathbb{R}^2$.
2. A path γ in S^2 is a smooth curve on the surface of the sphere.
3. A path in a space given the discrete topology must be a constant map.

We now return to considering the continuous analogue of clustering; in light of Definition 1.3.22, this is straightforward – we replace the discrete paths by continuous ones.

Definition 1.3.24. Let $(X, \mathcal{U}_X)$ be a topological space. Two points $p, q \in X$ are *path-connected* if there exists a continuous path $\gamma\colon [0,1] \to X$ such that $\gamma(0) = p$ and $\gamma(1) = q$.

It is clear that the relation of being path-connected is an equivalence relation (reparametrizing paths to obtain transitivity), and so the following definition makes sense.

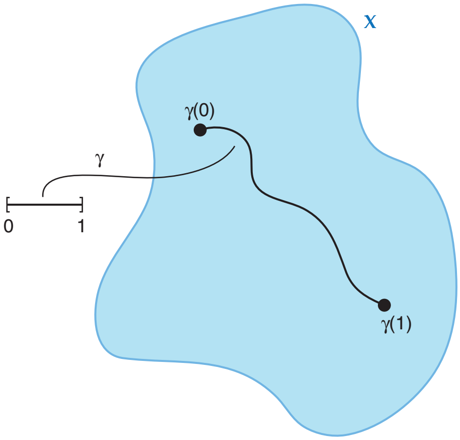

Figure 1.16 A path γ is a continuous map $\gamma\colon [0,1] \to X$.

Definition 1.3.25. We define the *path components* of a topological space $(X, \mathcal{U}_X)$ to be the collection of subsets of X such that x, y are in the same subset if and only if there is a path joining them.

We can think of the path components of X as giving a *continuous clustering* of the points of X; roughly speaking, two points are in distinct path components when they are separated by a "gap" in space. An important property of path components is that continuous maps of spaces give rise to maps of path components; this fact, referred to as *functoriality*, is essential for calculations (see Figure 1.17).

Lemma 1.3.26. *Let X and Y be topological spaces. Given a continuous map $f: X \to Y$, there is an induced map of sets between the path components of X and the path components of Y.*

1.3.2 Homeomorphisms

The construction of the set of path components is an example of a *topological invariant*; for two topological spaces that are "the same" in a suitable sense, the sets of path components should be isomorphic. To be precise about this, we need to describe when we will consider two topological spaces to be the same.

Definition 1.3.27. Topological spaces $(X, \mathcal{U}_X)$ and $(Y, \mathcal{U}_Y)$ are *homeomorphic* if there exists a bijection $f: X \to Y$ such that both f and f^{-1} are continuous maps.

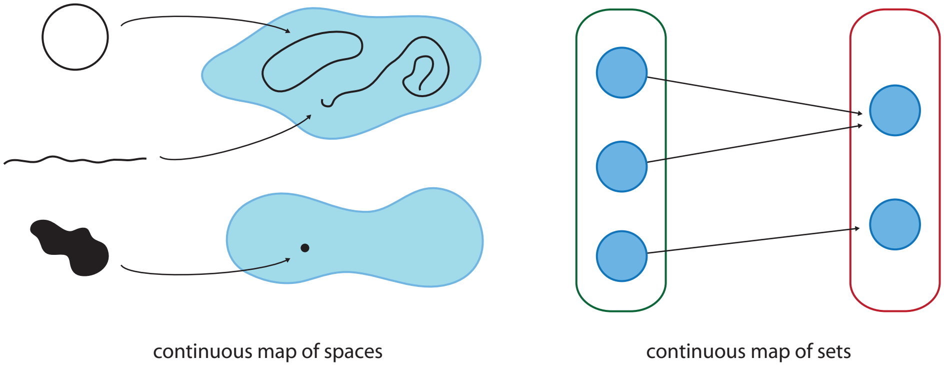

Figure 1.17 A continuous map of spaces induces a map of sets of path components. Here, the black space is the union of three components and the blue space the union of two.

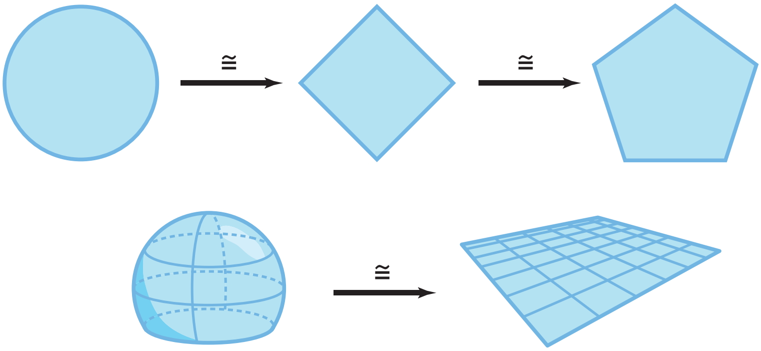

Figure 1.18 Two spaces are homeomorphic if there is a continuous bijection between them with continuous inverse. On the top, the circle is deformed into a pentagon. On the bottom, a sphere with the bottom cut off can be stretched onto a plane.

In this situation, we refer to f as a *homeomorphism*. Intuitively, two spaces are *homeomorphic* when they are related by a *continuous deformation*; roughly speaking, this means they are related by stretching and bending without introducing tears or gluing things together. See Figure 1.18 for a few examples of homeomorphic spaces.

Example 1.3.28.

1. The xy-plane $\mathbb{R}^2$ and a punctured two-dimensional sphere (i.e., a sphere where we have removed a point at one of the poles) are homeomorphic; there is a homeomorphism that “unwraps” the sphere. This homeomorphism is very familiar; this is a stereographic projection, used for example to make maps.
2. A square, a circle, and an octagon are all homeomorphic – we can define a homeomorphism by smoothing out the corners of the square and octagon, or alternatively adding kinks to the circle.
3. Famously, a coffee cup and a solid torus (a doughnut) are homeomorphic.

We can write $X \cong Y$ when two spaces X and Y are homeomorphic. The relation of homeomorphism is an equivalence relation on spaces:

1. it is reflexive (clearly $X \cong X$ via the identity map),
2. symmetric ($X \cong Y$ implies that $Y \cong X$), and
3. transitive (if $X \cong Y$ and $Y \cong Z$, composing the homeomorphisms shows that $X \cong Z$).

Recall from Lemma 1.3.26 that continuous maps of spaces induce maps of path components. When the continuous map in question is a homeomorphism, we can say something stronger.

Lemma 1.3.29. *Let $f\colon X \to Y$ be a homeomorphism. Then f induces a bijection between the set of path components of X and the set of path components of Y.*

We can interpret Lemma 1.3.29 to say that the number of path components is a topological invariant of a topological space. This numerical invariant is interesting, insofar as it allows us to distinguish spaces very easily.

Corollary 1.3.30. *Let X and Y be topological spaces. If X and Y have different numbers of path components, then X and Y are not homeomorphic. (Of course, two spaces with the same number of path components need not be homeomorphic!)*

We can directly relate the notion of path components to the problem of clustering discrete data, in a precise sense. First consider the case in which (M, ∂_M) is a finite metric subspace of $\mathbb{R}^n$. Fix a scale parameter $\epsilon \geq 0$. Then the topological space formed as the union

$$\mathcal{N} = \bigcup_{x \in M} \bar{B}_{\frac{\epsilon}{2}}(x)$$

has the property that the path components of $\mathcal{N}$ recover the clusters obtained via single-linkage clustering with parameter ϵ. However, a general finite metric space will not come with an embedding into $\mathbb{R}^n$; for this reason, it is useful to recast the clustering problem using a discretized topological model that encodes the same basic data.

To this end, we consider a construction which associates a graph to (M, ∂_M).

Definition 1.3.31. Let (M, ∂_M) be a finite metric space and fix $\epsilon \geq 0$. Define the associated *neighborhood graph* $G_\epsilon(M)$ to have *vertices* given by the points of M, and an *edge* (v_i, v_j) connecting v_i and v_j if and only if $\partial_M(v_i, v_j) \leq \epsilon$.

Regarding the graph as a topological space, we can give a graph-theoretic description of the path components.

Lemma 1.3.32. *Two vertices v_i and v_j in a graph G are in the same path component if there exists a collection of edges $(v_i, v_{k_1}), (v_{k_1}, v_{k_2}), \ldots, (v_{k_m}, v_j)$ where each pair of adjacent edges shares a vertex.*

It is now evident that the components of the graph associated to (M, ∂_M) correspond to the clusters given by single-linkage clustering with parameter ϵ.

1.4 Continuous Deformations and Homotopy Invariants

We have now arrived at the beginnings of homotopy theory; two excellent modern textbooks are by May [342] and Hatcher [235]. We observed in Lemma 1.3.29 in the previous section that the set of path components of a space X is a topological invariant, in the sense that if $f\colon X \to Y$ is a homeomorphism then the induced map on path components is an isomorphism. However, counting path components is much weaker than deciding whether two spaces are homeomorphic.

1. A circle and a point $\{x\}$ are not homeomorphic but have the same number of path components.
2. As an even simpler example, a disk $\{x \mid x \in \mathbb{R}^2, \|x\| \leq 1\}$ and a point $\{x\}$ have the same number of path components. However, they are clearly not homeomorphic (there is no map from $D^n \to \{x\}$ that is a bijection).

These examples motivate a search for a notion of equivalence that is weaker than homeomorphism and closer to comparing counts of path components. In particular, it seems reasonable to want a weaker kind of equivalence for which a point and a disk look the same but a point and a circle look different.

In order to introduce such a notion of equivalence, we will introduce the idea of a homotopy. A homotopy specifies a relationship between continuous maps from $X \to Y$; we will subsequently use this to define a kind of "approximate" homeomorphism.

Definition 1.4.1. Let X and Y be topological spaces. Then two continuous maps $f, g\colon X \to Y$ are *homotopic* if there exists a continuous map (called a *homotopy*) $h\colon X \times [0, 1] \to Y$ such that

$$\begin{cases} h(x, 0) = f(x) \\ h(x, 1) = g(x). \end{cases}$$

We write $f \simeq g$ when f and g are homotopic.

We think of $t \in [0, 1]$ as parametrizing a family of maps interpolating between f and g; for each t, h induces a continuous map $h(-, t)\colon X \to Y$. The continuity condition on h means that these maps vary "smoothly" as the parameter changes. In fact, for maps to Euclidean space, this description can be made precise as follows. (See also Figure 1.19.)

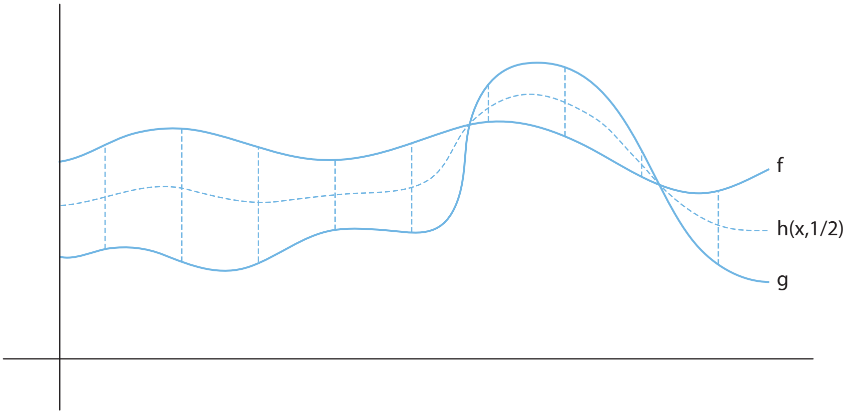

Figure 1.19 Two maps $\mathbb{R} \to \mathbb{R}$ are homotopic via linear interpolation. We can think of this as if we represented the graphs of f and g as rubber bands and dragged one to the other.

Example 1.4.2. Any two continuous maps $f, g\colon \mathbb{R}^m \to \mathbb{R}^n$ are homotopic; the homotopy is specified by interpolation as

$$h(x, t) = (1 - t)f(x) + t(g(x)).$$

The relation of being homotopic is an equivalence relation on the set Map(X, Y) of continuous maps between topological spaces X and Y. As in the previous examples, only transitivity is non-trivial to check. Assume that for $f, g, h\colon X \to Y$, we have $f \simeq g$ via the homotopy H_1 and $g \simeq h$ via the homotopy H_2. Then a homotopy H_3 defined as

$$\begin{cases} H_3(t, x) = H_1(2t, x) & 0 \le t \le \frac{1}{2} \\ H_3(t, x) = H_2(2t - 1, x) & \frac{1}{2} < t \le 1 \end{cases}$$

shows that $f \circ h$.

The notion of a homotopy now allows us to weaken the definition of homeomorphism; we will consider continuous maps $f\colon X \to Y$ that admit continuous inverses *up to homotopy*. Specifically, we have the following definition.

Definition 1.4.3. Let X and Y be topological spaces. Then X and Y are *homotopy equivalent* if there exist continuous maps

$$f\colon X \to Y \qquad \text{and} \qquad g\colon Y \to X$$

such that

$$f \circ g \simeq \mathrm{id}_Y \qquad \text{and} \qquad g \circ f \simeq \mathrm{id}_X.$$

(Here id_X and id_Y denote the identity maps on X and Y.) In this case, we write $X \simeq Y$ and we refer to f and g as *homotopy equivalences*.

Example 1.4.4.

1. Any spaces X and Y which are homeomorphic (via maps f and g) are also homotopy equivalent; the required homotopies are

$$\begin{aligned} h_1 \colon X \to X \qquad & h_1(x,t) = x \\ h_2 \colon Y \to Y \qquad & h_2(y,t) = y \end{aligned}$$

   since $f \circ g = \mathrm{id}_X$ and $g \circ f = \mathrm{id}_Y$.
2. For a disk $B_\epsilon(x) \subset \mathbb{R}^2$, the inclusion $i\colon \{x\} \to B_\epsilon(x)$ and the constant map $p\colon B_\epsilon(x) \to \{x\}$ induces a homotopy equivalence. The composite $p \circ i$ is equal to the identity, and for the composite $i \circ p$, we use the "radial contraction"

$$h((r,\theta),t) = (tr,\theta),$$

   where here we are representing the disk using polar coordinates. See the left panel of Figure 1.20 below for a picture of this process.
3. Recall (from Example 1.3.11) that S^1 denotes the standard unit circle. A cylinder $[0,1] \times S^1$ is homotopy equivalent to the circle; the maps are the inclusion $S^1 \to [0,1] \times S^1$ that takes $(x,y) \mapsto (0,(x,y))$ and the collapse that takes $(t,(x,y)) \mapsto (x,y)$. Once again, the composite of the inclusion and the collapse is the identity and the other composite is homotopic to the identity via the homotopy

$$h(t,(s,x,y)) = (ts,x,y).$$

   See the right panel of Figure 1.20 for a picture of this process.

Homotopy equivalence is an equivalence relation on spaces:

1. it is reflexive (clearly $X \simeq X$ via the identity homotopy),
2. symmetric ($X \simeq Y$ implies that $Y \simeq X$, using the same homotopy in the opposite direction), and
3. transitive; this is the only property that is not immediate. The key idea is that given homotopy equivalences $f_1\colon X \to Y$ and $f_2\colon Y \to Z$ (with inverses g_1 and g_2), we can build a homotopy from $(f_2 \circ f_1) \circ (g_1 \circ g_2)$ to the identity of Z by using the homotopy from $f_1 \circ g_1$ to the identity of Y on the interval $[0, \frac{1}{2}]$ and the homotopy from $f_2 \circ g_2$ to the identity of Z on the interval $[\frac{1}{2}, 1]$.

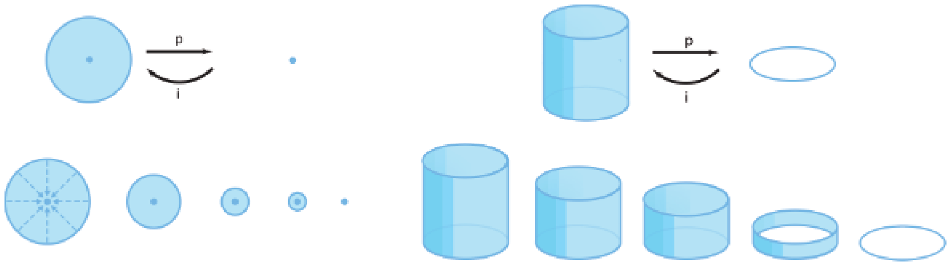

Figure 1.20 Radially shrinking a disk realizes the homotopy equivalence between a point and a disk. A cylinder shrinks along its length to a circle.

Definition 1.4.5. We will refer to the equivalence class of a space under the relation of homotopy equivalence as its *homotopy type*.

To understand homotopy equivalence, it is useful to consider the notion of a *deformation retraction*.

Definition 1.4.6. Let $A \subset X$ be a subspace. Then A is a *deformation retraction* of X if there exists a homotopy $H\colon X \times I \to X$ such that $H(x, 0) = x$, $H(x, 1) \in A$, and $H(a, 1) = a$.

A deformation retraction specifies a homotopy equivalence between A and X. Not all homotopy equivalences are deformation retractions, but one can show that two spaces X and Y are homotopy equivalent if and only if there is a space Z such that X and Y are each deformation retractions of Z.

Lemma 1.3.29 showed that counting path components of a space was a homeomorphism invariant. In fact, it is an invariant of homotopy equivalence.

Lemma 1.4.7. *Let X and Y be topological spaces such that there is a homotopy equivalence $f\colon X \to Y$. Then f induces a bijection between the set of path components of X and the set of path components of Y.*

In order to study homotopy equivalences, it turns out to be useful to consider the set obtained by taking *homotopy classes* of maps; two continuous maps are in the same homotopy class if they are homotopic.

Definition 1.4.8. Let X and Y be topological spaces. The set of *homotopy classes* of maps from X to Y, denoted $\{X, Y\}$, is the set of equivalence classes in $\mathrm{Map}(X, Y)$ under the equivalence relation given by homotopy.

1.4.1 Homotopy Groups

An essential insight from early in the development of algebraic topology is the idea that homotopy classes of maps from certain "test spaces" capture the homotopy type of a topological space. The test spaces we need are the standard spheres.

Definition 1.4.9. Let D^n denote the n-dimensional unit disk in $\mathbb{R}^n$ defined as

$$D^n = \left\{(x_1, \ldots, x_n) \in \mathbb{R}^n \mid \sum_{i=1}^{n} x_i^2 \leq 1\right\}$$

and let S^{n-1} denote the $(n-1)$-dimensional unit sphere in $\mathbb{R}^n$ defined as

$$S^{n-1} = \left\{(x_1, \ldots, x_n) \in \mathbb{R}^n \mid \sum_{i=1}^{n} x_i^2 = 1\right\}.$$

Observe that there is a natural inclusion $S^{n-1} \to D^n$ as the boundary.

Notice that $D^1 = [-1, 1] \subseteq \mathbb{R}^1$, $S^0 = \{-1, 1\} \subseteq \mathbb{R}^1$, and so forth. We regard D^n and S^{n-1} as topologized using the subspace topology, with regard to the standard topology on $\mathbb{R}^n$.

Now we define the *homotopy groups*. These will be sets with some additional algebraic structure, which we will describe informally below and then more precisely in Section 1.6.4. For this definition, we use the notion of a *based homotopy*, which is simply a homotopy $H \colon X \times I \to Y$ that has the property that for specified basepoints $x \in X$ and $y \in Y$, $H(x, t) = y$ for all t.

Definition 1.4.10. Let X be a topological space and $x \in X$ a point. Choose a point $p \in S^n$. Then for $n \geq 0$, as a set, the nth *homotopy group* $\pi_n(X, x)$ is the set of based homotopy classes $\{S^n, X\}$ where the point p is sent to x.

Up to isomorphism, the homotopy groups are independent of the choice of basepoint in the spheres S^n, but might change depending on the chosen basepoint in the target space X. For example, if X has many path components, then $\pi_n(X, x)$ will depend on which path component x lies in.

Example 1.4.11.

1. When $n = 0$, $\pi_0(X, x)$ is the set of path components of X.
2. When $n = 1$, $\pi_1(X, x)$ is called the *fundamental group*, the set of homotopy classes of loops in X that start and end at x. (See Figure 1.21.)
3. The fundamental group $\pi_1(S^1, x)$, where x is any point of the circle, has elements in bijection with $\mathbb{Z}$; each homotopy class of maps from $S^1 \to S^1$ can

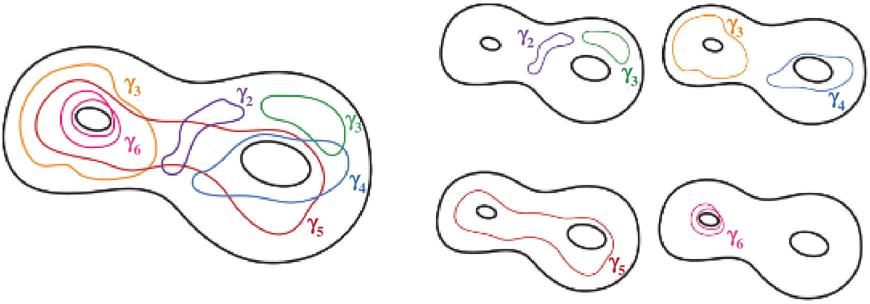

Figure 1.21 The fundamental group of a space X is the set of homotopy classes of loops.

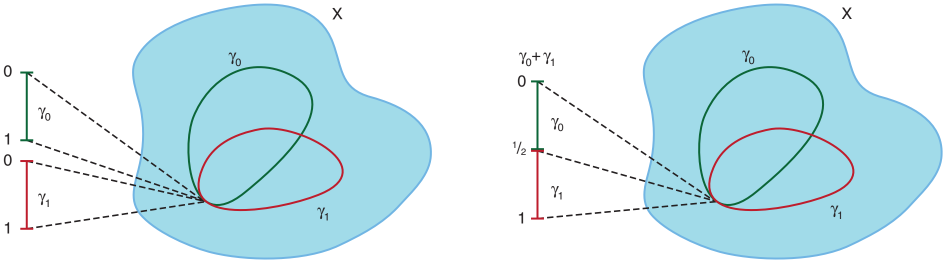

Figure 1.22 A loop $S^1 \to X$ is represented by a map $[0, 1] \to X$ with the same value on 0 and 1. Two loops γ_0 and γ_1 are added by reparameterizing, doing γ_0 on $[0, \frac{1}{2}]$ and γ_1 on $(\frac{1}{2}, 1]$.

be characterized by how many times it wraps around, and in which direction it goes.

The fundamental group of X records information about “holes” in X; a loop is homotopic to the constant map at a point unless it goes around a hole in X. (Of course, the loop might go around many times or it might wind around multiple holes; the intricacies of the geometry are reflected in the additional algebraic structure.)

When $n \geq 1$, π_n has additional algebraic structure; given two basepoint preserving maps from $S^1 \to X$, we can “add” them to get a new loop by doing first one, then the other. (See Figure 1.22.)

More generally, given two pointed maps $f_1, f_2 \colon S^n \to X$, we can make a new one by “pinching” a radial belt of the sphere to a point, forming two copies of the sphere, and then considering the new map that does f_1 on one bulb and f_2 on the other. (See Figure 1.23.)

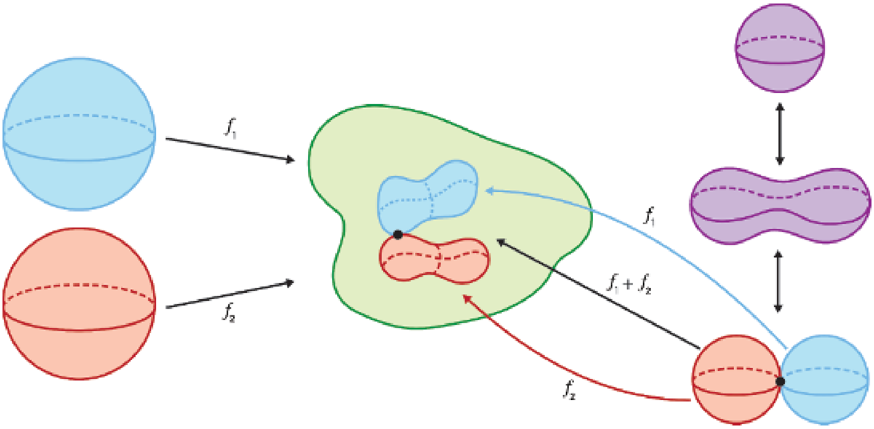

Figure 1.23 Two maps $f_1, f_2 \colon S^n \to X$ are added by taking a sphere, pinching it around the radius to produce two spheres joined at a point, and then doing f_1 on one “bulb” and f_2 on the other.

In fact, not only can we add, we can subtract as well. In Section 1.6, we quickly review the abstract framework for this kind of algebraic structure; in Section 1.6.4, we return to discuss the homotopy groups in more detail.

Another important property of the homotopy groups is that they behave nicely in the presence of continuous maps. Specifically, restating Lemma 1.4.7 in this language, we have the following result.

Lemma 1.4.12. *Let X and Y be topological spaces and $f \colon X \to Y$ a homotopy equivalence. Then for any $x \in X$ there is an isomorphism of sets $\pi_0(X, x) \cong \pi_0(Y, f(x))$.*

More generally, we have the following result.

Proposition 1.4.13. *Let X and Y be topological spaces and $f \colon X \to Y$ a homotopy equivalence. Then for any $x \in X$, there is an isomorphism of sets $\pi_n(X, x) \cong \pi_n(Y, f(x))$.*

The most pressing question about the homotopy groups is now to what degree there is a converse to Proposition 1.4.13. An answer to this question and a justification of the use of spheres as test objects is provided by the theory of CW complexes.

1.5 Gluing and CW Complexes

When contemplating practical work with topological spaces, a very natural question arises: how do we concretely specify the data of a topological space? Definition 1.3.1 is very well suited for abstract reasoning, but is not usually convenient as a way to present a generic space. In particular, since our eventual goals involve devising algorithms for computing topological invariants that are tractable on computers, we want to develop means of encoding topological spaces that are discrete.

If we restrict attention to the question of working with spaces up to homotopy equivalence, then we obtain additional flexibility. The idea is now to model a given homotopy type by particularly nice spaces; in a precise sense, it turns out that we can always replace an arbitrary topological space by one which has a very regular topological structure. This approach is based on an inductive description of a topological space in terms of building blocks that are easily understood, namely disks and spheres.

In order to describe the topology on spaces built up in this way, we begin by describing the *quotient topology*. To motivate this construction, consider the interval $[0, 1]$, topologized with the subspace topology from $\mathbb{R}$. Gluing together the two endpoints $\{0\} \subset [0, 1]$ and $\{1\} \subset [0, 1]$ should produce a circle. The quotient topology is a way to make this precise.

Proposition 1.5.1. *Let X be a topological space and Y a set. Let $p\colon X \to Y$ be a surjective map. Then we can make Y a topological space by specifying that a subset $U \subset Y$ is open when $p^{-1}(U)$ is an open set in X. We call the topology on Y the quotient topology.*

Equivalently, given a continuous surjection of topological spaces $p\colon X \to Y$, we can identify a criterion for when the topology on Y is the quotient topology.

Proposition 1.5.2. *Given a surjective map of topological spaces $p\colon X \to Y$, we say that p is a quotient map provided that $U \subseteq Y$ is an open set in Y if and only if $p^{-1}(U) \subseteq X$ is an open set in X. In this case, the topology on Y is the quotient topology.*

We can now identify the usual topology on the unit circle S^1 as the quotient topology.

Example 1.5.3. Let $p\colon [0, 1] \to S^1$ be the map specified by $x \mapsto (\cos(2\pi x), \sin(2\pi x))$. Then p is a quotient map. (See figure 1.24.)

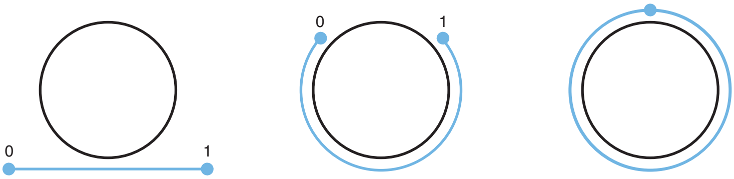

Figure 1.24 The unit interval wraps around the circle, joined at the endpoints.

Given a topological space, it is often useful to have a more intrinsic way of producing a surjective map $f\colon X \to Y$, where Y is a set; by intrinsic, we mean defined in terms of some sort of "gluing" data on X. For this, we need the notion of a partition.

Definition 1.5.4. Given a topological space X, we let a *partition* of X be a decomposition

$$X = \bigcup X_i, \quad \text{where} \quad X_i \cap X_j = \emptyset, i \neq j.$$

A partition specifies an equivalence relation on the points of X, where x and y are equivalent when $x, y \in X_i$.

The basic idea is that all of the points in each X_i are going to be glued together.

Definition 1.5.5. Given a partition $\{X_i\}$ of X, the *quotient space* of the partition is a topological space with points the set of partitions. The topology is induced by the surjective map $X \to \{X_i\}$ which takes $x \in X$ such that $x \in X_i$ to X_i. Put another way, we are topologizing the set of equivalence classes determined by the partition.

For instance, if we take the partition of $[0,1]$ specified by $\{0,1\}$ and the points $\{x\}$ in the open interval $(0,1)$, we generate the usual topology on S^1 as in Example 1.5.3. A rich source of partitions comes from circumstances in which we want to glue a space X to a space Y along a map from $Z \subset X$ to Y.

Definition 1.5.6. Let X and Y be topological spaces, $Z \subseteq X$ a subspace of X, and $f\colon Z \to Y$ a continuous map. Define a partition on the disjoint union $X \coprod Y$ with sets

$$\begin{cases} \{x\} & \forall x \in X - Z, \\ \{y\} & \forall y \in Y - f(Z), \\ \{z, f(z)\} & \forall z \in Z. \end{cases}$$

The *gluing* $X \cup_f Y$ is the quotient space associated to this partition.

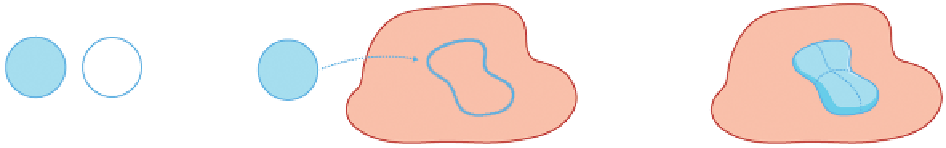

Figure 1.25 Cell attachment involves gluing on a disk along its boundary; here, the boundary circle of the blue disk is glued to the red loop on the surface.

For example, Definition 1.5.6 allows us to regard S^1 as obtained by gluing two copies of $[0, 1]$ along the map that identifies the endpoints. More generally, Definition 1.5.6 allows us to regard S^n as built by gluing two copies of D^n along the boundary $S^{n-1} = \partial D^n \subset D^n$.

Now, we will describe an inductive process for constructing a topological space by repeatedly gluing on disks along their boundaries, as follows (see Figure 1.25).

1. Let X_0 be a set of points, given the discrete topology. These are the *zero cells*.
2. Form X_1 by *attaching* copies of D^1 to X_0 by gluing them along their boundaries – that is, we are given the data of continuous maps

$$f_\alpha \colon \partial D_1 = S^0 \to X_0$$

   (referred to as *attaching maps*), and for each one we look at the quotient $D_1 \cup_{f_\alpha} X_0$ of the disjoint union $X_0 \coprod D^1$ where we identify the points $z \in S^0 \subseteq D^1$ and $f_\alpha(z) \in X_0$. The intervals glued in during this stage are referred to as 1-cells.
3. Then we repeat, attaching copies of D^2 to X_1 by gluing them along their boundaries – in this case, the data of the attaching maps is given by continuous maps $f_\beta \colon S^1 \to X_1$, and we form the corresponding union $D^2 \coprod_f X_0$. The disks glued in during this stage are referred to as 2-cells.
4. And so on . . .

Formalizing this, we have the following definition.

Definition 1.5.7. A *finite CW complex* is a topological space obtained as a finite union $\bigcup_i X_i$ in which each stage X_i is obtained from X_{i-1} by gluing on copies of D^i as above. (The topology is the natural quotient topology induced by the gluing, and is independent of the order in which cells are attached.)

The subspace $X_n \subset X$ is referred to as the n-skeleton, and consists of k-cells for $k \leq n$; if there are no cells of dimension larger than m, then the CW complex A is referred to as m-dimensional. Notice that the essential data of the CW complex is contained in the number of cells and the attaching maps, and the n-skeleton encodes all of the attaching data for objects of dimension less than n.

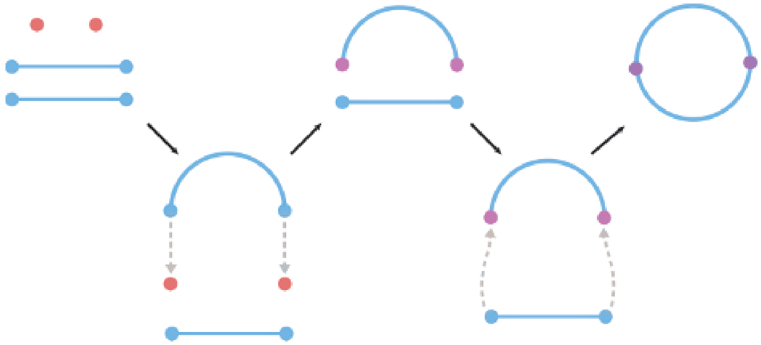

Figure 1.26 The circle S^1 can be formed by gluing two intervals along their boundaries.

Remark 1.5.8. It is also possible to consider an infinite attachment process of this kind, but the construction of the topology on the infinite union requires some care.

Example 1.5.9.

1. Any graph can be realized as a CW complex with one 0-cell for each vertex and a 1-cell for each edge (glued to the relevant vertices).
2. The circle S^1 can be given the structure of a CW complex in which $X_0 = \{0\}$ and X_1 is obtained by the map that attaches $[-1, 1]$ to 0 via the map from $\{-1, 1\}$ that takes both points to 0.
3. The circle can also be given many CW structures, as follows: take n 0-cells (points), where $n \geq 2$. Label these points as $\{x_1, \ldots, x_n\}$. Then take n 1-cells (intervals) and attach them sequentially to connect x_1, x_2, then x_2, x_3, then x_i, x_{i+1}, and finally x_n, x_1. (See Figure 1.26.)
4. In general, a sphere can be given a CW structure by taking a single 0-cell and a single n-cell and gluing the n-cell to the 0-cell along the map that sends the entire boundary to the point.
5. A torus (the surface of a doughnut) can be given the structure of a CW complex by taking a single 0-cell, two 1-cells, and a 2-cell. The two 1-cells are glued to the 0-cell to form a figure-eight, and then the 2-cell is glued to the figure-eight to make the torus. (See Figure 1.27.)

We now describe two ways to construct new CW complexes out of old that cover many interesting examples.

Definition 1.5.10. Let X and Y be CW complexes. Then $X \times Y$ has the structure of a CW complex where the cells are the products of the cells of X and Y.

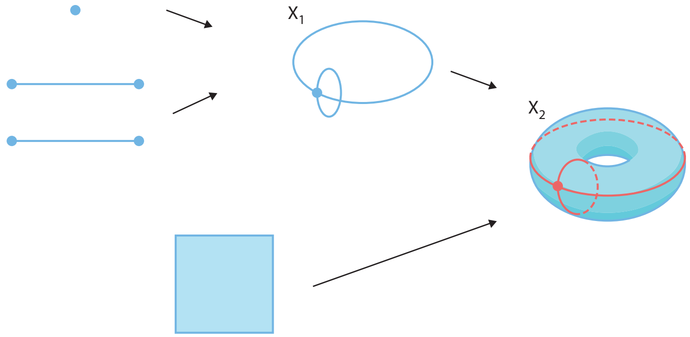

Figure 1.27 The torus can be built up by gluing together two intervals and then a two-cell to the resulting figure-eight.

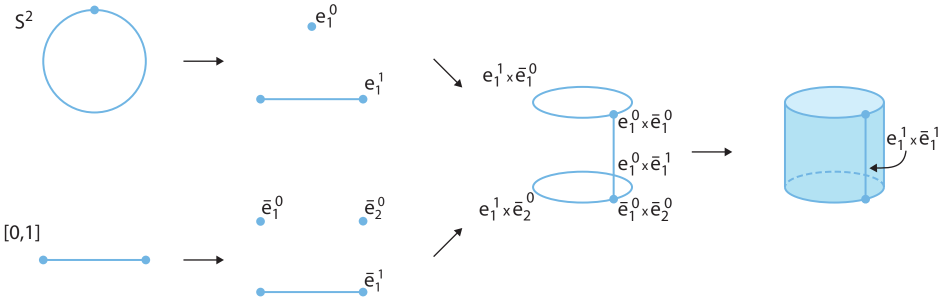
S^2
e_1^0
e_1^1
$e_1^1 \times \bar{e}_1^0$
$e_1^0 \times \bar{e}_1^0$
$e_1^0 \times \bar{e}_1^1$
$e_1^1 \times \bar{e}_1^1$
$e_1^1 \times \bar{e}_2^0$
$\bar{e}_1^0 \times \bar{e}_2^0$
[0,1]
$\bar{e}_1^0$
$\bar{e}_2^0$
$\bar{e}_1^1$

Figure 1.28 The cylinder is the product of a circle and an interval.

To be more explicit, given a cell D^n attached to X along $f\colon S^{n-1} \to X$ and D^m attached to Y along $g\colon S^{m-1} \to Y$, we can attach a cell $D^{n+m} \cong D^n \times D^m$ to $X \times Y$ along the map $S^{n+m-1} \to X \times Y$ determined by the homeomorphism

$$S^{n+m+1} \cong (D^n \times S^{m-1}) \cup (S^{n-1} \times D^m),$$

the maps f and g, and the inclusions $D^n \to X$ and $D^m \to Y$.

Example 1.5.11.

1. The standard cylinder $S^1 \times [0,1]$ can be given a CW complex structure as the product of the CW complex S^1 and the CW complex $[0,1]$. (See Figure 1.28.)
2. The torus can be given a CW complex structure as the product of the CW complexes $S^1 \times S^1$.

A *subcomplex* of a CW complex is just a closed subspace determined by taking only some of the cells.

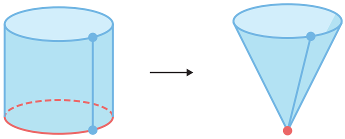

Figure 1.29 Collapsing one of the copies of S^1 inside the cylinder $S^1 \times [0,1]$ to a point results in a cone.

Definition 1.5.12. Let X be a CW complex and A a subcomplex; then the *quotient* X/A has a CW complex structure consisting of the cells of X that are not contained in A, along with a new 0-cell representing A. (An attaching map $\gamma\colon S^n \to X$ gives rise to an attaching map $\gamma'\colon S^n \to X \to X/A$.)

Taking the cylinder from Example 1.5.11 and taking the quotient $S^1 \times [0,1]/S^1 \times \{0\}$ gives rise to a model for the CW complex structure on a cone; see Figure 1.29.

There are three essential results about CW complexes that justify focus on these combinatorial models of spaces.

1. Replacing an attaching map in a CW complex by a homotopic map does not change the homotopy type.
2. A homotopy equivalence $X \to Y$ of CW complexes can be detected algebraically in terms of the homotopy groups π_n.
3. Any reasonable topological space can be approximated up to homotopy equivalence by a CW complex, and for an arbitrary topological space there is an approximation up to a weak kind of equivalence. (See Definition 1.6.32 below.)

The first observation tells us that the data of a CW complex is entirely contained in the homotopy classes of the attaching maps. The second and third observations imply that if we are working up to homotopy equivalence, CW complexes are a good model for general spaces and that homotopy equivalence classes can be studied algebraically. That is, CW complexes provide a class of spaces which are constructed according to a recipe from basic building blocks and are well suited to work up to homotopy equivalence. To make the last two observations precise (notably in Theorem 1.6.31), we need to develop some algebraic background.

The next section, which briefly reviews abstract algebra (notably group theory and ring theory), may be particularly difficult for readers new to the subject. On a first reading, a quick perusal of Section 1.6.6 for a refresher on linear algebra might suffice; such readers could then skip to Section 1.7, which introduces ideas from category theory.

1.6 Algebra

A central goal of algebraic topology is to produce suitable *algebraic* invariants of topological spaces to allow us to determine whether two spaces are homeomorphic or homotopy equivalent. For example, the function which takes a topological space to the number of path components is an example of such an invariant; by Lemma 1.4.12, this invariant can serve to distinguish certain spaces with different homotopy types.

Early on in the development of the subject, it was recognized that more discriminatory power could be obtained by considering more structured algebraic objects than numbers as repositories for topological invariants. For example, the set of path components is a richer invariant than simply its size. The point is that there are no maps between numbers, but there are maps of sets – and we have seen in Lemma 1.3.26 that a continuous map of spaces induces a map of sets of path components.

It turns out that keeping even more algebraic structure leads to invariants that are computable and very informative. For example, consider the problem of distinguishing the circle from the figure-eight. Looking at homotopy classes of maps from S^1, both of these have an infinite number. But in the circle, the homotopy classes are all "multiples" of the basic one which wraps around once, and in the figure-eight all of the homotopy classes are built from combinations of the classes which wrap around one circle or the other. Algebraic invariants provide a way to make precise the intuitive notion of being "built from" or "generated by" these basic loops, and therefore let us tell these spaces apart.

In order to describe these algebraic invariants, we now turn to a quick review of the background from abstract algebra that we need. Again, our treatment is very terse and selective; we refer the reader to one of the many excellent treatments of abstract algebra, for example Artin's *Algebra* [22] or Lang's *Undergraduate Algebra* [314]. We begin by reviewing the theory of groups.

1.6.1 Groups

A group is a set with the additional structure of an "addition" operation.

Definition 1.6.1. A set G is equipped with the structure of a group if there is a distinguished element $e \in G$ and functions

$$G \times G \to G \qquad (g_1, g_2) \mapsto g_1 +_G g_2$$

and

$$G \to G \qquad g \mapsto -g$$

such that

1.
$$\forall x \in G, \qquad e +_G x = x = x +_G e,$$

2.
$$\forall x \in G, \qquad x +_G (-x) = e = (-x) +_G x,$$

3. and
$$\forall x, y, z \in G, \qquad x +_G (y +_G z) = (x +_G y) +_G z.$$

We will often write $g_1 + g_2$ rather than $g_1 +_G g_2$ and usually write 0 for e, in analogy with the notation. We sometimes use "multiplicative" notation and write $g_1 g_2$ rather than $g_1 +_G g_2$, 1 for e, and g^{-1} for the inverse of g.

Put another way, a group is a set equipped with an "addition" operation that is associative, has a unit element, and such that every element $x \in G$ has an inverse. The definition of a group is an abstraction of familiar objects from arithmetic.

Example 1.6.2.

1. The integers $\mathbb{Z}$ under the standard addition operation form a group; $x +_{\mathbb{Z}} y = x + y$ for $x, y \in \mathbb{Z}$. The unit is $0 \in \mathbb{Z}$, and the inverse of x is $-x$.
2. The real numbers $\mathbb{R}$ under the standard addition operation form a group; $x +_{\mathbb{R}} y = x + y \in \mathbb{R}$. The unit is $0 \in \mathbb{R}$ and the inverse of x is $-x$.
3. The non-zero real numbers $\mathbb{R} - \{0\}$ under multiplication form a group; the operation is $(x, y) \mapsto xy$ for $x, y \in \mathbb{R} - \{0\}$. The unit is $1 \in \mathbb{R}$ and the inverse of x is $\frac{1}{x}$. (It is the existence of inverses that requires us to restrict to non-zero reals!)
4. The set of all polynomials in $\mathbb{R}$ of degree k in a single variable t,
$$\mathcal{P}_k = \{a_0 + a_1 t + \ldots + a_k t^k \,|\, a_0, a_1, \ldots, a_k \in \mathbb{R}\},$$
is a group under addition of polynomials, i.e.,
$$(a_0 + a_1 t + \ldots + a_k t^k) + (b_0 + b_1 t + \ldots + b_k t^k) = (a_0 + b_0) + (a_1 + b_1)t + \ldots + (a_k + b_k)t^k.$$
The identity element is 0 and the inverse of $p(x)$ is $-p(x)$.
5. The set $C(\mathbb{R})$ of all continuous functions $f \colon \mathbb{R} \to \mathbb{R}$ is a group under pointwise addition, i.e.,
$$f +_{C(\mathbb{R})} g = (f + g)(x) = f(x) + g(x).$$
The identity element is the zero function $f(x) = 0$ and the inverse of a function f is $-f$.
6. The set of $n \times n$ matrices with real elements $M_n(\mathbb{R})$ is a group under matrix addition. The unit element is the zero matrix and the inverse of A is the matrix $-A$.

7. The set of invertible $n \times n$ matrices $\mathrm{GL}_n(\mathbb{R})$ is a group under matrix multiplication where the unit element is the identity matrix and the inverse of A is the inverse matrix A^{-1}.

The example of $\mathrm{GL}_n(\mathbb{R})$ is particularly interesting, since this group has the property that the operation is not commutative, i.e., $AB \neq BA$ in general.

Definition 1.6.3. A group G is *abelian* if for all $x, y \in G$, we have $x +_G y = y +_G x$.

Example 1.6.4. All of the examples above in Example 1.6.2 are abelian except for $\mathrm{GL}_n(\mathbb{R})$.

The examples of groups we have discussed above are "numerical." But historically, groups arose from symmetries and rigid transformations of physical objects; for example, the set of rotations of an object in space forms a group. More abstractly, the symmetries of a finite set form a group.

Example 1.6.5.

1. The set of symmetries of a square is the group generated by two elements r and f; r is the counterclockwise rotation and f is the flip across a diagonal. These are subject to certain relations, as indicated in Figure 1.30; the group has 8 elements. In general, the *dihedral groups D_n* describe the symmetries of a regular n-gon in the plane, and have $2n$ elements.
2. The set of rotations of the unit cube $[-1, 1] \times [-1, 1] \times [-1, 1] \subset \mathbb{R}^3$ about the z-axis is the circle group S^1; we can parametrize the elements as $e^{i\theta}$, with group operation $e^{i\theta_1} e^{i\theta_2} = e^{i(\theta_1+\theta_2)}$. (See Figure 1.31.)

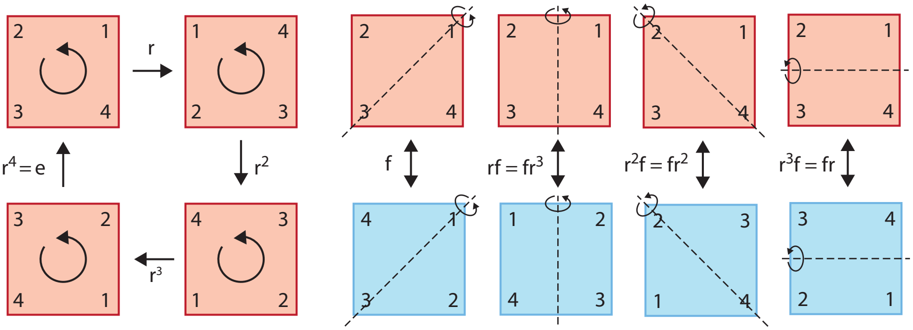

Figure 1.30 The rotation and the flip across a diagonal specify two basic symmetries of a square. Together these generate a group of order 8, the dihedral group D_4.

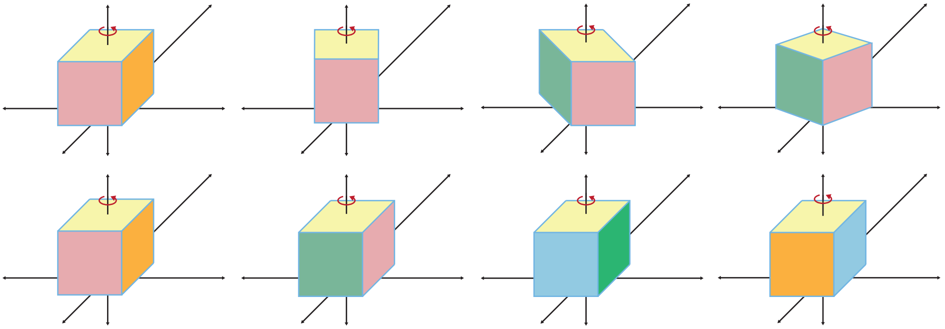

Figure 1.31 There is a natural action of the circle on a cube that rotates the cube around the z-axis. On the top row, we see some snapshots of this rotation. There is a natural subgroup isomorphic to $\mathbb{Z}/4$ inside of S^1 determined by rotations by $90°$; the action of this group on the cube is shown on the bottom row.

3. The set of rotations of $\mathbb{R}^3$ about the origin forms a group, the *special orthogonal group* $SO(3)$. This can be described as the set of orthogonal 3×3 matrices (i.e., matrices A such that $A^{-1} = A^T$) with determinant 1. The group operation is matrix multiplication. The identity is the identity map (i.e., the rotation that leaves everything fixed) and the inverse of a rotation is the "opposite" rotation.
4. Let S be an ordered set with n elements. The set of permutations of S (i.e., bijective maps $S \to S$) forms a group. The identity is the permutation that leaves every element of S in place, the group operation is given by composition of permutations, and the inverse of a permutation is the permutation that "undoes" it.

Another important arithmetic example comes from *modular arithmetic*.

Definition 1.6.6. For x and y in $\mathbb{Z}$, define $x = y \mod n$ if $x - y = kn$, for some $k \in \mathbb{Z}$. The *congruence class* of x modulo n is a subset of the form

$$\{x + kn \,|\, k \in \mathbb{Z}\}.$$

The classical long division algorithm implies that a congruence class has a unique smallest nonnegative representative, the remainder r when we write $x = qn + r$ via long division.

Example 1.6.7. The set of congruence classes modulo n, which we can represent as $\{0, 1, 2, \ldots, n-1\}$, forms a group that we denote by $\mathbb{Z}/n$. The identity element is 0, addition is given by letting the sum of x and y be $x + y \mod n$, and the inverse of x is $n - x \mod n$.

The preceding example has a special structure; it is a *cyclic group*, in the sense that every element other than the identity is generated by sums of a distinguished generator, for example 1. The integers $\mathbb{Z}$ are an *infinite cyclic group*, with generator 1. But not all groups are cyclic; for example, $SO(3)$ is very far from being cyclic.

1.6.2 Homomorphisms

A fundamental tenet of modern mathematics is that to understand a collection of mathematical objects it is essential to understand the maps between them. An important aspect of this principle is that invariants should "take maps to maps." We have already seen this at work in the context of topological spaces and continuous maps: a continuous map of spaces induces a map between sets of path components. In Section 1.7, we will describe an abstract framework for formalizing this insight.

In the meantime, we want to describe the correct notion of a map between groups. Recall that we singled out the class of continuous maps when describing functions between topological spaces; these were the functions that were suitably compatible with the topologies of the domain and range. Correspondingly, we are primarily interested in functions between groups which respect the group structure, in the sense of the following definition.

Definition 1.6.8. A map $f\colon G_1 \to G_2$ is a *group homomorphism* if

$$f(0) = 0 \qquad \text{and} \qquad f(x +_{G_1} y) = f(x) +_{G_2} f(y) \quad \forall x, y \in G_1.$$

Example 1.6.9.

1. The natural inclusion $\mathbb{Z} \to \mathbb{R}$ is a group homomorphism.
2. The projection $\mathbb{Z} \to \mathbb{Z}/m$ specified by the formula

$$x \mapsto x \mod m$$

   is a group homomorphism.
3. The derivative

$$\begin{aligned} \frac{d}{dt}\colon \mathcal{P}_k &\to \mathcal{P}_{k-1} \\ a_0 + a_1 t + a_2 t^2 + \ldots + a_k t^k &\mapsto a_1 + 2a_2 t + \ldots + k a_k t^{k-1} \end{aligned}$$

   is a group homomorphism.
4. The trace of a square matrix with real entries (the sum of the diagonal elements) specifies a group homomorphism

$$\mathrm{Tr}\colon M_n(\mathbb{R}) \to \mathbb{R}.$$

Associated to a homomorphism $f\colon G_1 \to G_2$ are certain distinguished subsets of G_1 and G_2.

Definition 1.6.10. Let $f\colon G_1 \to G_2$ be a group homomorphism (Figure 1.32).

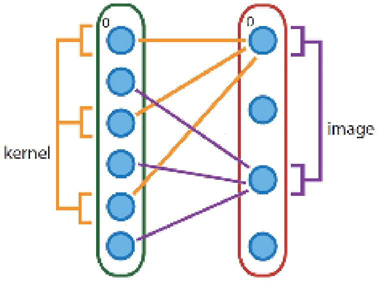

Figure 1.32 The kernel of a homomorphism f is the set of points that go to 0; the image is the set of points that f hits.

- The *kernel* of f, $\ker f \subseteq G_1$, is the set of elements x such that $f(x) = 0$.
- The *image* of f, $\operatorname{im} f \subseteq G_2$, is the set of elements y such that $y = f(x)$ for some x.

Generalizing the notion of an isomorphism of sets from Definition 1.1.8, we have the following version in the context of groups and group homomorphisms.

Definition 1.6.11. A group homomorphism $f: G_1 \to G_2$ is an *isomorphism* if there exists an inverse group homomorphism $g: G_2 \to G_1$ such that f and g demonstrate an isomorphism of sets between G_1 and G_2.

Equivalently, we have the following characterization.

Lemma 1.6.12. *Let G_1 and G_2 be groups. A group homomorphism $f: G_1 \to G_2$ is an isomorphism if and only if it is a bijection. As a consequence, f is an isomorphism if and only if* $\ker f = \{0\}$ *and* $\operatorname{im} f = G_2$.

Both $\ker f$ and $\operatorname{im} f$ are themselves groups, with operations inherited from G_1 and G_2 respectively. These are *subgroups* of G_1 and G_2, as we now explain.

1.6.3 New Groups from Old

Many groups of interest arise via constructions that start from an existing group. The simplest is to consider subsets of a group that inherit the structure of a group themselves.

Definition 1.6.13. A *subgroup* of a group G is a subset $H \subseteq G$ such that H is a group in its own right with the operation and unit inherited from G. That is,

1. the identity element $0 \in G$ is an element of H,
2. for any $h \in H$, $-h$ is in H, and

3. for all $h_1, h_2 \in H$, the sum $h_1 + h_2$ is in H.

We have already seen some examples of subgroups.

Example 1.6.14.

1. The special orthogonal group $SO(3)$ is a subgroup of $\mathrm{GL}_3(\mathbb{R})$.
2. The set $\{x \in \mathbb{Z} \mid x \text{ even}\}$ is a subgroup of $\mathbb{Z}$ under addition.
3. The set $\mathcal{P}_k$ of degree at most k polynomials is a subgroup of $\mathcal{P}_{k+1}$.
4. The set $\mathcal{P}_k$ of degree at most k polynomials is a subgroup of $C(\mathbb{R})$.

The following lemma provides many other examples of subgroups.

Lemma 1.6.15. *Let $f \colon G_1 \to G_2$ be a group homomorphism. Then* $\ker f \subseteq G_1$ *is a subgroup of* G_1 *and* $\operatorname{im} f \subseteq G_2$ *is a subgroup of* G_2.

The preceding lemma is a simple exercise in the properties of group homomorphisms; for the first part, if $f(g_1) = 0$ and $f(g_2) = 0$, then

$$f(g_1 + g_2) = f(g_1) + f(g_2) = 0 + 0 = 0.$$

Given a suitable subgroup $H \subset G$, we can "collapse it out" by forming the quotient group G/H of G by a subgroup H, which is akin to the quotient topology discussed above in Proposition 1.5.1. The idea is to specify that in G/H all elements of H are identified. We will define the quotient in the setting of an abelian group G; when G is not abelian, only certain subgroups permit the construction of the quotient group.

Definition 1.6.16. Let G be an abelian group and $H \subset G$ a subgroup. Then the *quotient group* G/H is given by the set of *cosets* $gH = \{gh \,|\, h \in H\}$ as g varies, with group operation $(g_1H)(g_2H) = (g_1g_2)H$.

(Note that a small check is required to verify that the definition of the quotient group is independent of choice of coset representative.)

Example 1.6.17. For $\mathbb{Z}$ and the subgroup $3\mathbb{Z} = \{3k \,|\, k \in \mathbb{Z}\}$, the quotient $\mathbb{Z}/3\mathbb{Z}$ is isomorphic to the construction of $\mathbb{Z}/3$ described in Example 1.6.7.

A basic structural property of group homomorphisms can be usefully described in terms of the quotient group.

Theorem 1.6.18. *Let G_1 and G_2 be groups and $f\colon G_1 \to G_2$ be a group homomorphism. Then there is an isomorphism*

$$\operatorname{im} f \cong G_1/\ker f.$$

(This is true even if G_1 is not abelian; the kernel of a homomorphism allows the construction of the quotient.)

As an elaboration of this result, we can describe a large class of groups in terms of *generators and relations*.

Definition 1.6.19. A group G is *finitely generated* if there exists a finite set $S \subseteq G$ such that any $g \in G$ can be written as a (finite) sum of elements in S.

For example, any finite group is of course finitely generated. The integers $\mathbb{Z}$ are finitely generated with generator 1. On the other hand, the rationals $\mathbb{Q}$ are not finitely generated. Clearly, a finitely generated group must be countable; therefore, $\mathbb{R}$ is not finitely generated.

Definition 1.6.20. A group is *free* if there exists a collection of elements $\{g_\alpha\}$ (called the *generators*) such that every element $g \in G$ can be uniquely written as a finite sum

$$\sum_i n_i g_{\alpha_i}$$

for $n_i \in \mathbb{Z}$.

Free groups are easy to work with because group homomorphisms $F \to G$, where F is free, can be described simply as set maps from the generators of F to G. That is, to specify such a group homomorphism f, it suffices to give the data of where each generator lands in G,

$$f\left(\sum_i n_i g_{\alpha_i}\right) = \sum_i n_i f(g_{\alpha_i}).$$

Theorem 1.6.21. *Any finitely generated group G is isomorphic to the quotient of a free group by a subgroup described by specifying products of generators that are equal to* 1.

We refer to the generators of the free group as the generators of G and the products describing the subgroup as the relations of G. From an algorithmic perspective, a presentation of a group in terms of generators and relations is essential.

Example 1.6.22.

1. The integers $\mathbb{Z}$ can be represented as having the identity element 0, a single generator 1, and no relations. Here the element -1 must exist and is distinct from 1, and in general we have a description as

$$\mathbb{Z} \cong \{\ldots, -1 + (-1) + (-1), -1 + (-1), -1, 0, 1, 1 + 1, 1 + 1 + 1, \ldots\}.$$

2. The cyclic group $\mathbb{Z}/3$ is the quotient of the free group $\mathbb{Z}$ by the subgroup of relations $\{3k \,|\, k \in \mathbb{Z}\}$. Another way to express this is that $\mathbb{Z}/3$ can be described as having an identity element, a single generator g, and the single relation $g^3 = 1$. Then explicitly this representation describes $\mathbb{Z}/3$ as the set $\{1, g, g^2\}$ with the usual multiplication of polynomials as the group operation; $g^{-1} = g^2$, since $(g)(g^2) = g^3 = 1$.

Remark 1.6.23. Note that an interesting problem arises in this context, namely, the problem of deciding when two "words" representing group elements are equal. For instance, in the group with generator $\{x\}$ and relation $x^4 = 1$, one might ask whether x^8 and x^{16} are the same. This is known as the *word problem* for a group, and it is an important classical result that this is *undecidable*. That is, there does not exist any algorithm (computer program) to solve this problem in general! This hardness result is the core of many demonstrations that certain mathematical questions are undecidable.

However, for our purposes it will suffice to consider free abelian groups. A free abelian group with one generator is an infinite cyclic group and is isomorphic to $\mathbb{Z}$. In order to describe free groups with more generators, we need the notion of a product.

Definition 1.6.24. Let G_1 and G_2 be groups (not necessarily abelian). Then the Cartesian product $G_1 \times G_2$ denotes the group structure on the Cartesian product of sets with identity element $(0_{G_1}, 0_{G_2})$, operation

$$(g_1, g_2) + (g_1', g_2') = (g_1 + g_1', g_2 + g_2'),$$

and the inverse of (g_1, g_2) is $(-g_1, -g_2)$.

Lemma 1.6.25. *A free abelian group with k generators is isomorphic to a product of k copies of $\mathbb{Z}$: one copy of $\mathbb{Z}$ for each generator.*

In this case, any finitely generated abelian group G is isomorphic to a quotient $\mathbb{Z}^n/H$, for a subgroup $H \in \mathbb{Z}^n$; here n is the size of a set S of generators. More precisely, we have the following fundamental characterization, the *structure theorem for finitely generated abelian groups*.

Theorem 1.6.26. *Let G be a finitely generated abelian group. Then there is an isomorphism*

$$G \cong \underbrace{\mathbb{Z} \times \mathbb{Z} \times \ldots \mathbb{Z}}_{k} \times \mathbb{Z}/p_1^{n_1} \times \mathbb{Z}/p_2^{n_2} \times \ldots \times \mathbb{Z}/p_m^{n_m}.$$

Here the p_i are prime and not necessarily distinct.

The number k of factors of $\mathbb{Z}$ is known as the *rank* of G. The part of G that does not consist of copies of $\mathbb{Z}$ is often referred to as the *torsion*. The rank is unique and the torsion is unique up to rearrangement.

1.6.4 The Group Structure on $\pi_n(X, x)$

We now return to justify referring to the homotopy groups $\pi_n(X, x)$ (from Definition 1.4.10) as groups. Specifically, we explain the following theorem.

Theorem 1.6.27. *When $n > 0$, the set of homotopy classes of maps*

$$\pi_n(X, x) = \{(S^n, *), (X, x)\}$$

can be given the structure of a group, where the identity element is the constant map and the composition is given by "composing" maps.

We begin by considering the case of $\pi_1(X, x)$. Given two loops $\gamma_1, \gamma_2 \colon S^1 \to X$, we can produce a new loop as follows. Regard the maps γ_1 and γ_2 as paths (maps from $[0, 1]$ to X) such that

$$\gamma_1(0) = \gamma_2(0) = \gamma_1(1) = \gamma_2(1) = x.$$

Then define $\gamma_1\gamma_2 \colon [0, 1] \to X$ to be the loop specified by the formula

$$(\gamma_1\gamma_2)(t) = \begin{cases} \gamma_1(2t) & 0 \le t < \frac{1}{2}, \\ \gamma_2(2t-1) & \frac{1}{2} \le t \le 1. \end{cases}$$

That is, we reparameterize and do γ_1 on the first half of the interval and γ_2 on the second half of the interval (see Figure 1.33). Since $\gamma_1(0) = \gamma_2(1) = x$, this defines a map $S^1 \to X$.

Note that the composition we have just defined is not associative prior to passing to homotopy classes of maps; that is, $(\gamma_1\gamma_2)\gamma_3$ is not the same map as $\gamma_1(\gamma_2\gamma_3)$. Specifically, given $\gamma_1, \gamma_2, \gamma_3 \colon S^1 \to X$, $(\gamma_1\gamma_2)\gamma_3$ does γ_1 on $[0, \frac{1}{4})$, γ_2 on $[\frac{1}{4}, \frac{1}{2})$ and γ_3 on $[\frac{1}{2}, 1]$ whereas $\gamma_1(\gamma_2\gamma_3)$ does γ_1 on $[0, \frac{1}{2})$, γ_2 on $[\frac{1}{2}, \frac{3}{4})$, and γ_3 on $[\frac{3}{4}, 1]$. However, there is a natural straight-line homotopy connecting $(\gamma_1\gamma_2)\gamma_3$ to $\gamma_1(\gamma_2\gamma_3)$; see Figure 1.34.

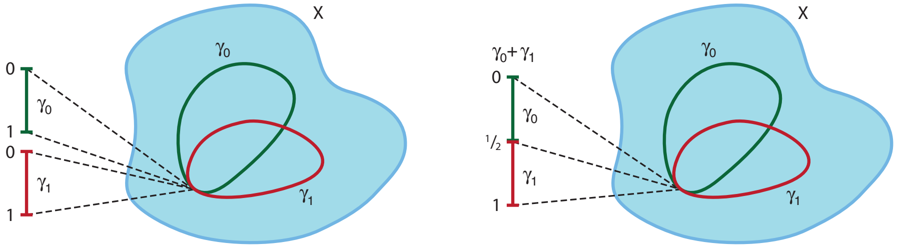

Figure 1.33 Two loops γ_0 and γ_1 are added by reparameterizing, doing γ_0 on $[0, \frac{1}{2})$ and γ_1 on $[\frac{1}{2}, 1]$.

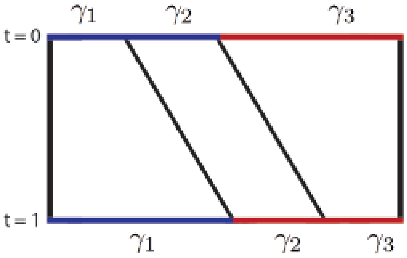

Figure 1.34 A linear homotopy connects the two associativity parameterizations.

Analogously, we define the inverse of $\gamma\colon S^1 \to X$ to be the loop traversed in the opposite direction:

$$\gamma^{-1}(t) = \gamma(1-t).$$

Once again, note that $\gamma\gamma^{-1}$ is not equal to the constant map until we pass to homotopy classes of maps; there is a homotopy connecting $\gamma\gamma^{-1}$ to the constant map that takes all of S^1 to x.

Generalizing this, we can put a group structure on $\pi_n(X, x)$ for $n > 1$ as follows. We regard maps from $S^n \to X$ as maps from $[0, 1]^n \to X$ which take the boundary of $[0, 1]^n$ to x and again compose by reparametrizing. We have choices about how to reparameterize; fixing an index $1 \le i \le n$, we define

$$\gamma_1\gamma_2(x_1, x_2, \ldots, x_n) = \begin{cases} \gamma_1(x_1, x_2, \ldots, 2x_i, \ldots, x_n) & x_i \in [0, \frac{1}{2}) \\ \gamma_1(x_1, x_2, \ldots, 2x_i - 1, \ldots, x_n) & x_i \in [\frac{1}{2}, 1]. \end{cases}$$

(See Figure 1.35.)

Once again, there is a homotopy that makes this associative. In fact, for $n > 1$, we have the following improvement of Theorem 1.6.27.

Theorem 1.6.28. *For $n > 1$, the homotopy group $\pi_n(X, x)$ is abelian.*

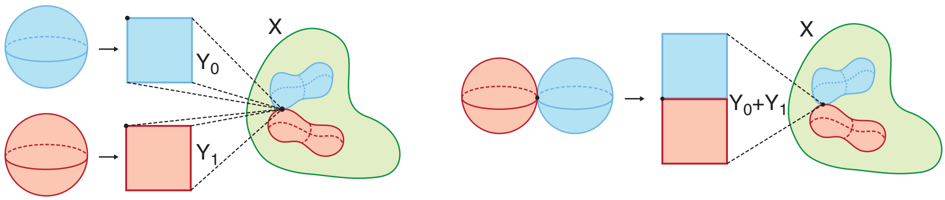

Figure 1.35 Two maps from spheres γ_0 and γ_1 are added by reparameterizing, doing γ_0 on the upper square and γ_1 on the lower square.

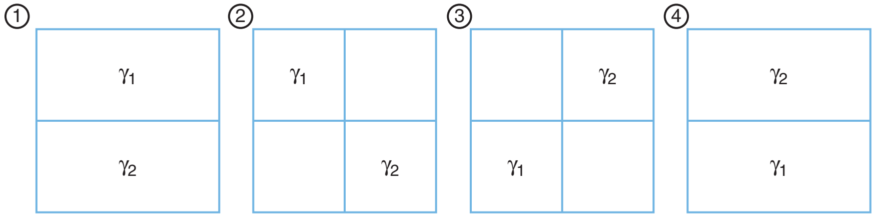

Figure 1.36 The commutativity homotopy involves moving two squares past each other. Here the unlabeled squares are sent to the basepoint.

A picture of the commutativity homotopy that proves Theorem 1.6.28 is shown in Figure 1.36.

Given a continuous map $f\colon X \to Y$, composition defines a map

$$\pi_n(X, x) \to \pi_n(Y, f(x))$$

via

$$\gamma\colon S^1 \to X \mapsto (f \circ \gamma)\colon S^1 \to X \to Y.$$

In fact, this map specifies a group homomorphism when $n > 0$.

Lemma 1.6.29. *Let $f\colon X \to Y$ be a continuous map of spaces. There are induced group homomorphisms for $n > 0$*

$$\pi_n(X, x) \to \pi_n(Y, f(x)).$$

The importance of the homotopy groups as algebraic invariants is provided by the following two theorems. First, homotopy groups are invariants of the homotopy type.

Proposition 1.6.30. *Let $f\colon X \to Y$ be a homotopy equivalence. Then the induced group homomorphism*

$$\pi_n(X, x) \to \pi_n(Y, f(x))$$

is an isomorphism.

Although the converse to this is not in general true, we have the following basic result.

Theorem 1.6.31 (Whitehead). *Let $f\colon X \to Y$ be a continuous map of CW complexes such that the induced maps $\pi_n(X, x) \to \pi_n(Y, f(x))$ are isomorphisms for every $n \geq 0$ and $x \in X$. Then f is a homotopy equivalence between X and Y.*

We say that a map f that induces isomorphisms of homotopy groups as in Theorem 1.6.31 is a *weak homotopy equivalence*.

Definition 1.6.32. Let $f\colon X \to Y$ be a continuous map of topological spaces. Then f is a *weak homotopy equivalence* (or *weak equivalence*) if the induced group homomorphisms

$$\pi_n(X, x) \to \pi_n(Y, f(x))$$

are isomorphisms for every $n \geq 0$ and $x \in X$.

This is a central definition in modern algebraic topology. Moreover, it turns out that any topological space X is weakly homotopy equivalent to a CW complex. (Warning: note that not every space is homotopy equivalent to a CW complex. For example, the sequence $\{\frac{1}{n}\}$ along with its limit point 0 is not homotopy equivalent to a CW complex. See also discussion of the "long line", e.g., in [369, §10].)

Weak homotopy equivalence is not an equivalence relation on spaces, but we work with the transitive closure, which is the smallest equivalence relation it generates.

Definition 1.6.33. We will refer to the equivalence class of a space under the relation of weak homotopy equivalence as its *weak homotopy type*.

We now have a number of different equivalence relations on topological spaces. These relations are progressively weaker – the relationship between them can be summarized as follows.

1. If two spaces X and Y are homeomorphic, then they are homotopy equivalent.
2. If two spaces X and Y are homotopy equivalent, then they are weakly homotopy equivalent.

Theorem 1.6.31 shows that for CW complexes, the latter two equivalence relations coincide. In contrast, determining when a homotopy equivalence is even homotopic to a homeomorphism is quite difficult; a restricted version of this problem is the subject of *surgery theory*.

Although homotopy groups are easy to define and the Whitehead theorem implies that they are complete invariants of the homotopy type of a CW complex (in the presence of a continuous map), the best known algorithms for computing them in general are intractable. As a consequence, we are led to search for algebraic invariants which are rich enough to distinguish a wide class of spaces but can be easily computed.

1.6.5 Rings and Fields

We return to the basic examples of the abelian groups $(\mathbb{Z}, +, 0)$ and $(\mathbb{R}, +, 0)$, the integers and the real numbers with group operation given by addition. These groups have additional structure, namely a second operation – multiplication. Moreover, multiplication interacts nicely with addition, for example, the distributive property tells us that $x(y + z) = xy + xz$.

Definition 1.6.34. A *ring* is a set R that has an abelian group structure (with operation denoted by $+$ and identity by 0) along with a distinguished element $1 \in R$ and an additional operation

$$R \times R \to R \qquad (x, y) \mapsto xy$$

such that

$$\forall x \in G, \quad 1x = x = x1,$$

and

$$x(yz) = x(yz).$$

In addition, we require that the new operation satisfy the distributive law with respect to the abelian group structure:

$$x(y + z) = xy + xz.$$
$$(x + y)z = xz + yz.$$

A ring has both an additive identity element (typically written 0) and a multiplicative identity element (typically written 1). A *multiplicative inverse* for an element $x \in R$ is an element y such that $xy = 1$; typically we write x^{-1} for the multiplicative inverse. An element $x \in R$ that has a multiplicative inverse is called a *unit*. Not all elements of a ring have multiplicative inverses.

Definition 1.6.35. A *field* F is a ring such that for all $x \in R$ such that $x \neq 0$ (where 0 denotes the additive identity), x has a multiplicative inverse x^{-1} such that $xx^{-1} = x^{-1}x = 1$.

Example 1.6.36.

1. The integers $\mathbb{Z}$ with addition and multiplication form a ring, but not a field as there is no multiplicative inverse for any $x \neq \pm 1$.
2. The rational numbers $\mathbb{Q}$ with addition and multiplication form a ring and in fact a field; the inverse of $\frac{p}{q}$ is $\frac{q}{p}$, which is well defined as long as $p \neq 0$.
3. The set of congruence classes $\mathbb{Z}/m$ forms a ring, where multiplication is also computed by taking the remainder of xy when divided by m. When m is prime, this is in fact a field; the inverse can be computed using the long division algorithm. The fields $\mathbb{Z}/p$ are referred to as finite fields of order p.

In addition to $\mathbb{R}$, the most important fields for our purposes are the rational numbers $\mathbb{Q}$ and the finite fields $\mathbb{Z}/p$ (which are often denoted $\mathbb{F}_p$). For any field F, we can consider a vector space with F as the scalars. Although we assume that the reader has some familiarity with linear algebra in the context of the fields $\mathbb{R}$ and $\mathbb{C}$, we quickly review linear algebra from a more abstract perspective.

1.6.6 Vector Spaces and Linear Algebra

Linear algebra studies the geometric structure of solutions to systems of linear equations; these turn out to form lines and (hyper)planes. It is a central example of the power of using algebraic structures to encode geometry. There are an enormous number of textbooks on linear algebra. For an abstract treatment, Axler's book [24] is very clearly written. For applications, Meyer's book is an excellent introduction [349].

The basic object in linear algebra is the vector space, which is an abstraction of some parts of the structure of Euclidean space.

Definition 1.6.37. Let F be a field. An *F-vector space* is an abelian group V with an additional operation called *scalar multiplication*

$$F \times V \to V \qquad (x, v) \mapsto xv$$

that is

1. associative, $x_1(x_2 v) = (x_1 x_2)v$,
2. distributive with respect to addition in F, $(x_1 + x_2)v = x_1 v + x_2 v$,
3. distributive with respect to the group operation in V, $x(v_1 + v_2) = xv_1 + xv_2$, and
4. compatible with the multiplicative unit in F, $1v = v$.

We call the elements of V vectors.

Example 1.6.38.

1. The field F itself gives a first example of a vector space.
2. The set $\{0\}$ is a vector space.
3. When $F = \mathbb{R}$, familiar examples of vector spaces are given by $\mathbb{R}^n$, where $\mathbb{R}$ acts by multiplication in each component.
4. More generally, for any field F, the product $F^n = \prod_{i=1}^{n} F$ of n copies of F is a vector space where F acts by componentwise multiplication.

Other basic examples of vector spaces are given by *subspaces*.

Definition 1.6.39. A *subspace* W of a vector space V is a subgroup such that $kw \in W$ for all $k \in F, w \in W$. (That is, W is closed under addition in V and scalar multiplication.)

Vector spaces can sometimes be decomposed into pieces by subspaces.

Definition 1.6.40. Let U and W be subspaces of the vector space V. If $U \cap W = \{0\}$, the *direct sum* $U \oplus W$ is defined to be the collection

$$U \oplus W = \{u + w \mid u \in U, w \in W\}.$$

More generally, given two vector spaces V_1 and V_2 we can define the external direct sum $V_1 \oplus V_2$ to consist of pairs (v_1, v_2) for $v_1 \in V_1$ and $v_2 \in V_2$, with the operations defined coordinatewise. Then regarding V_1 and V_2 as subspaces of $V_1 \oplus V_2$ (via $\{(v_1, 0)\}$ and $\{(0, v_2)\}$, respectively), $V_1 \oplus V_2$ arises as their direct sum as in Definition 1.6.40.

Although a priori it appears that subspaces could take on many forms, in fact, it turns out that all examples of finite-dimensional vector spaces look like the examples in 1.6.38. For example, the subspaces of $\mathbb{R}^2$ are $\{0\}$, $\mathbb{R}^2$ itself, and lines that pass through the origin. Each such line looks like a copy of $\mathbb{R}$. Similarly, the subspaces of $\mathbb{R}^3$ are $\{0\}$, lines through the origin (which look like $\mathbb{R}$), planes through the origin (which look like $\mathbb{R}^2$), and $\mathbb{R}^3$ itself. To be precise about this fact, we need the notion of a basis, which generalizes the idea of the coordinate axes in Euclidean space.

Definition 1.6.41. Let V be a vector space. For a subset $B = \{b_1, b_2, \ldots, b_n\} \subseteq V$,

1. *B spans V* if any vector $z \in V$ can be written as a sum

$$z = \sum_{i=1}^{n} a_i b_i, \quad a_i \in F,$$

i.e., any vector admits a representation as a weighted sum of basis elements,

2. the set B is *linearly independent* if the only solution to the equation

$$\sum_{i=1}^{n} a_i b_i = 0$$

is $a_i = 0$ for all i, and
3. B is a basis for a vector space V if it *spans* and is *linearly independent*.

Linear independence is a way of saying that a set of vectors has no redundancy, in the following sense.

Lemma 1.6.42. *The set B is linearly independent if and only if when $z \in V$ can be written as a sum*

$$z = \sum_{i=1}^{n} a_i b_i,$$

then this representation is unique, i.e., the values $\{a_i\}$ are unique.

Example 1.6.43.

1. In $\mathbb{R}^2$, the standard unit vectors along the axes $(1,0)$ and $(0,1)$ form a basis.
2. In $\mathbb{R}^2$, the vectors $(3,4)$ and $(-1,1)$ form a basis. In fact, any two non-collinear vectors form a basis. (See Figure 1.37 and Figure 1.38 for an example.)
3. In $\mathbb{R}^3$, any three vectors that do not all lie in the same plane form a basis.
4. More generally, in $\mathbb{R}^n$, any n vectors that do not all lie in the same *hyperplane* (i.e., subspace of strictly smaller dimension) form a basis.

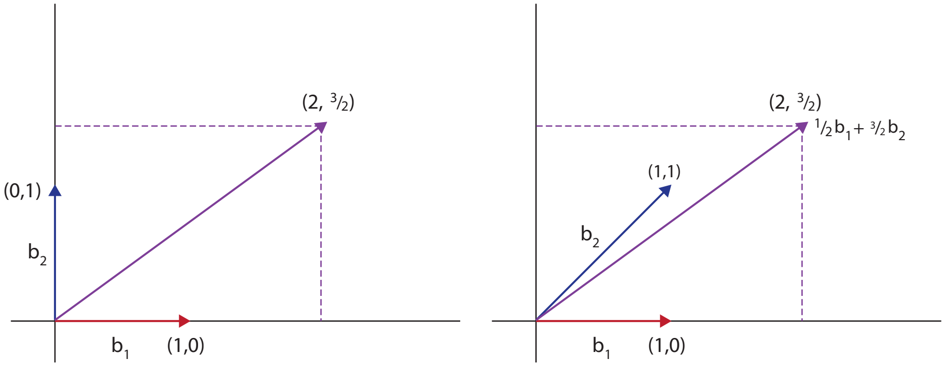

Figure 1.37 Any vector in $\mathbb{R}^2$ can be written uniquely as a linear combination $a_1 v_1 + a_2 v_2$ as long as v_1 and v_2 do not lie on the same line. We illustrate this for the vector (2,3/2).

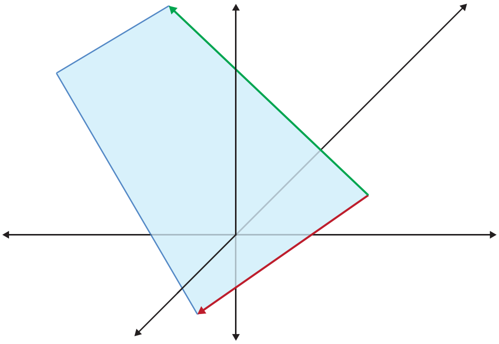

Figure 1.38 Any two-dimensional subspace of $\mathbb{R}^3$ is a plane; two vectors that specify the plane provide a basis.

By providing coordinates for describing points in vector spaces, bases are essential for calculation. They also give rise to the notion of dimension of a vector space.

Proposition 1.6.44. *Any basis for a vector space V has the same size.*

In light of the preceding proposition, the following definition makes sense.

Definition 1.6.45. The *dimension* of a vector space V is the size of a basis.

In fact, the dimension is a complete invariant of finite-dimensional vector spaces. To be precise, we need to define the notion of a map between vector spaces.

Definition 1.6.46. Let V and W be vector spaces. A *linear transformation* $f\colon V \to W$ is a map of sets such that

$$f(ax + by) = af(x) + bf(y).$$

That is, a linear transformation is a group homomorphism that preserves scalar multiplication.

The kernel and image of a linear transformation $f\colon V_1 \to V_2$ are subgroups of V_1 and V_2 respectively. In fact, they are vector spaces themselves.

Lemma 1.6.47. *Let $f\colon V_1 \to V_2$ be a linear transformation. Then* $\ker f$ *is a subspace of V_1 and* $\operatorname{im} f$ *is a subspace of V_2.*

One of the appealing things about linear transformations is that they can be expressed in a concise and algorithmically tractable way. Since a vector space is the set of linear combinations of basis elements, a linear transformation can be specified simply in terms of its action on the basis. Put another way, linear transformations can be specified by matrices; the ith column of the matrix describes the effect of the linear transformation applied to the basis vector b_i.

Definition 1.6.48. A linear transformation $f\colon V \to W$ is an *isomorphism* if it is injective and surjective, or equivalently if there is an inverse transformation $g\colon W \to V$ such that $g \circ f = \mathrm{id}_V$ and $f \circ g = \mathrm{id}_W$.

Theorem 1.6.49. *Any vector space of dimension n is isomorphic to F^n.*

The homotopy groups $\pi_n(X, x)$ are groups that are very hard to compute. The basic topological invariants that will be our algorithmic focus take values in vector spaces; the fact that a linear transformation can be specified by a matrix will ensure that computation is tractable. Before we introduce these invariants, we will have a brief interlude about category theory, which provides a formal context to describe the invariants.

1.7 Category Theory

The basic topological invariants we study are *functions* that take as input topological spaces (represented by CW complexes or simplicial complexes) and output finitely generated abelian groups or vector spaces:

$$\left\{\begin{matrix}\text{finite}\\ \text{simplicial complexes}\end{matrix}\right\} \to \left\{\begin{matrix}\text{abelian}\\ \text{groups}\end{matrix}\right\}.$$

However, these invariants are better than functions, as they turn out to have an additional essential property: they take continuous maps between spaces to group homomorphisms. We have already seen an example of this in Lemma 1.6.29, which states that a continuous map $f\colon X \to Y$ induces a group homomorphism $\pi_k(X, x) \to \pi_k(Y, f(x))$. Formalizing this property of algebraic invariants was one of the original motivations for the invention of *category theory*.

Category theory provides a language for capturing common phenomena in different domains. For example, the notion of an isomorphism has appeared in a variety of different contexts in this chapter. A motivating idea at the core of the development of category theory is the notion that properties of mathematical objects (e.g., topological spaces) can often be characterized entirely in terms of maps from other objects. We have seen this philosophy at work already in our discussion of homotopy groups. Properties that can be expressed purely in terms of such data are often referred to as *formal*; a common slogan is that category theory is a way to make formal things formal. We give a very brief overview of category theory; the classic text is Mac Lane [337]. Riehl has written two excellent recent books, [428] which is a more elementary introduction and [427] which is an in-depth discussion from the perspective of algebraic topology. Spivak's book [480] strives to provide context for categorical notions in applications.

Definition 1.7.1. A *category* $\mathcal{C}$ is a collection of objects $\mathrm{ob}(\mathcal{C})$ and for each pair of objects $x, y \in \mathrm{ob}(\mathcal{C})$ a set of *morphisms* or maps $\mathrm{Hom}_{\mathcal{C}}(x, y)$ satisfying the following conditions.

1. For all objects $w, x, y \in \mathcal{C}$, there is a composition map

$$\mathrm{Hom}_{\mathcal{C}}(x, y) \times \mathrm{Hom}_{\mathcal{C}}(w, x) \to \mathrm{Hom}_{\mathcal{C}}(w, y)$$

   that takes the morphisms $f\colon w \to x$ and $g\colon x \to y$ to the composite morphism $g \circ f\colon w \to y$.
2. There is a distinguished element $\mathrm{id}_x \in \mathrm{Hom}_{\mathcal{C}}(x, x)$, the identity map.
3. Composition is associative and unital. Associativity means that given $f \in \mathrm{Hom}(w, x)$, $g \in \mathrm{Hom}(x, y)$, and $h \in \mathrm{Hom}(y, z)$, we have the equality of composites

$$(h \circ g) \circ f = h \circ (g \circ f).$$

   Unitality means that

$$\mathrm{id}_x \circ f = f = f \circ \mathrm{id}_w.$$

The composition map is written in the "backwards" order above in order to align with the standard notation for composition, i.e. $(g \circ f)(-) = g(f(-))$.

Remark 1.7.2. The sophisticated reader will notice that we are being incautious about set theory and using the somewhat vague term "collection"; as we discussed in Section 1.1, Russell's paradox tells us that there is no "set of all sets," and so there cannot be a set of objects for the category of sets. We refer the reader to the category theory references for more discussion of this point.

We have many familiar examples of categories underlying the notions we have already seen.

Example 1.7.3.

1. The category Set with objects sets and morphisms maps of sets.
2. The category Grp with objects groups and morphisms homomorphisms.
3. The category Vect with objects vector spaces and morphisms linear transformations.
4. The category Top with objects topological spaces and morphisms continuous maps.

5. The category Met of metric spaces and metric maps (i.e., maps $f\colon X \to Y$ such that $\partial_Y(f(x_1), f(x_2)) \leq \partial_X(x_1, x_2)$).
6. The category Ho(Top) with objects topological spaces and morphisms homotopy classes of continuous maps.
7. A partially ordered set forms a category. For example, $\mathbb{N}$ is a category with objects the elements of $\mathbb{N}$ and a morphism between x and y if $x \leq y$.

Moreover, for any category we can obtain new categories by taking subsets of the collection of objects and morphisms.

Definition 1.7.4. A category $\mathcal{D}$ is a *subcategory* of a category $\mathcal{C}$ if each object of $\mathcal{D}$ is an object of $\mathcal{C}$ and for every $x, y \in \mathrm{ob}(\mathcal{D})$, we have

$$\mathrm{Hom}_{\mathcal{D}}(x, y) \subseteq \mathrm{Hom}_{\mathcal{C}}(x, y).$$

When we have equality in the previous inclusion, $\mathcal{D}$ is called a *full* subcategory of $\mathcal{C}$.

Example 1.7.5.

1. The category Ab with objects abelian groups and morphisms homomorphisms is a full subcategory of Grp.
2. The category of finite dimensional vector spaces and linear transformations is a full subcategory of Vect.
3. The category of topological spaces and morphisms the homeomorphisms is a subcategory of Top, although it is not full.

In any category, there is an intrinsic notion of two things being "the same" that comes directly from the data of the category.

Definition 1.7.6. Let $\mathcal{C}$ be a category. A map $f \in \mathrm{Hom}_{\mathcal{C}}(x, y)$ is an isomorphism if there exists $g \in \mathrm{Hom}_{\mathcal{C}}(y, x)$ such that

$$f \circ g = \mathrm{id}_y \in \mathrm{Hom}_{\mathcal{C}}(y, y) \qquad \text{and} \qquad g \circ f = \mathrm{id}_x \in \mathrm{Hom}_{\mathcal{C}}(x, x).$$

The notion of a categorical isomorphism encompasses all of the definitions we have seen so far.

Example 1.7.7.

1. In Set, an isomorphism is an isomorphism of sets (as defined in Definition 1.1.8).
2. In Grp, an isomorphism is an isomorphism of groups (as defined in Definition 1.6.11).
3. In Vect, an isomorphism is an isomorphism of vector spaces (as defined in Definition 1.6.48).

4. In Top, an isomorphism is a homeomorphism (as defined in Definition 1.3.27).
5. In Ho(Top), an isomorphism is the equivalence class of a homotopy equivalence (as defined in Definition 1.4.3).

Since the only properties we can express in a category are described in terms of morphisms and the result of composing morphisms, the notion of a *commutative diagram* is of basic importance. A commutative diagram refers to a collection of objects and morphisms such that any morphisms between two objects coincide. For example, in the commutative square

$$\begin{tikzcd} A \arrow[r, "f"] \arrow[d, "h"'] & B \arrow[d, "g"] \\ C \arrow[r, "i"] & D \end{tikzcd}$$

we are expressing the compatibility requirement that $g \circ f = i \circ h$ as a morphism in $\mathrm{Hom}_{\mathcal{C}}(A, D)$.

The structure of the category itself can encode many interesting properties of objects; we now give some examples.

Definition 1.7.8. An *initial object* in a category is an object c such that $\mathrm{Hom}_{\mathcal{C}}(c, z)$ consists of a single point for any z. That is, there is a unique morphism from c to any other object.

Dually, a *terminal object* is an object d such that $\mathrm{Hom}_{\mathcal{C}}(z, d)$ consists of a single point for any z.

These notions are not necessarily unique, although they are unique up to isomorphism, i.e., any two initial or terminal objects are isomorphic.

Example 1.7.9.

1. In Set, the initial object is the empty set $\emptyset$ and any one-point set is a terminal object. We will denote a choice of terminal object by $*$.
2. In Grp the initial object is the trivial group and the terminal object is also the trivial group.
3. In Top the initial object is $\emptyset$ and the one-point space is a terminal object. We will again denote a choice of terminal object by $*$.

The point here (no pun intended) is that the special properties of the one-point set or the one-point space can be expressed in a way which generalizes to any category; the properties can be expressed solely in terms of data about maps to and from other objects.

Moreover, commutative diagrams allow us to succinctly express algebraic properties. For instance, a group is an object G in the category Set along with a morphism $m\colon G \times G \to G$, a morphism $u\colon * \to G$, and a morphism $i\colon G \to G$ such that the following holds.

1. The diagram

$$\begin{tikzcd}
G\times G\times G \arrow[r, "m\times \mathrm{id}"] \arrow[d, "\mathrm{id}\times m"'] & G\times G \arrow[d, "m"] \\
G\times G \arrow[r, "m"'] & G.
\end{tikzcd}$$

   commutes; this expresses associativity.
2. The diagrams

$$\begin{tikzcd}
G \arrow[r, "\mathrm{id}\times u"] \arrow[dr, "\mathrm{id}"'] & G\times G \arrow[d, "m"] \\
 & G
\end{tikzcd}$$

   and

$$\begin{tikzcd}
G \arrow[r, "u\times \mathrm{id}"] \arrow[dr, "\mathrm{id}"'] & G\times G \arrow[d, "m"] \\
 & G
\end{tikzcd}$$

   commute; this expresses the property of the identity element.
3. The diagrams

$$\begin{tikzcd}
G \arrow[r, "\Delta"] \arrow[d, "u"'] & G\times G \arrow[d, "\mathrm{id}\times i"] \\
G & G\times G \arrow[l, "m"]
\end{tikzcd}$$

and

$$\begin{tikzcd}
G \arrow[r, "\Delta"] \arrow[d, "u"'] & G\times G \arrow[d, "i\times \mathrm{id}"] \\
G & G\times G \arrow[l, "m"]
\end{tikzcd}$$

commute, where $\Delta\colon G \to G{\times}G$ is the diagonal map specified by the assignment $x \mapsto (x, x)$ and $u\colon G \to G$ is the composite $G \to * \to G$ specified by the unique map $G \to *$ and the unit map $u\colon * \to G$. These diagrams express the property of the inverse.

We can also describe gluing constructions (e.g., the attaching of cells in Definition 1.5.6) purely in terms of categorical data. Suppose that we have a diagram

$$\begin{tikzcd}
A \arrow[r, "f"] \arrow[d, "g"'] & B \\
C &
\end{tikzcd}$$

in some category $\mathcal{C}$. Explicitly, this means that

1. A, B, and C are objects in the category $\mathcal{C}$,
2. f is an element of $\mathrm{Hom}_{\mathcal{C}}(A, B)$ and g is an element of $\mathrm{Hom}_{\mathcal{C}}(A, C)$.

We will refer to the data of this diagram as D. We now want to explain how to give a general construction of an object that is produced by “gluing” B to C along A.

To motivate the abstract definition, it is instructive to consider how to describe such a construction. Within category theory, the only way we can express the properties of such a gluing is to talk about morphisms either into or out of it, i.e., to talk about the gluing in terms of its relationship to other objects. Let us consider how to specify a map out of the gluing of B and C along A, to some other object X. Such a map should be determined by maps

$$B \to X \qquad \text{and} \qquad C \to X$$

that agree on the image of $A \to B$ and $A \to C$. Moreover, we would like the gluing to be the “smallest” such object. We can make all of this precise as follows.

Definition 1.7.10. The *pushout* of D is an object P equipped with morphisms $p_1 \colon B \to P$ and $p_2 \colon C \to P$ such that the square

$$\begin{tikzcd}
A \arrow[r, "f"] \arrow[d, "g"'] & B \arrow[d, "p_1"] \\
C \arrow[r, "p_2"'] & P
\end{tikzcd}$$

commutes, and for any pair of morphisms $a \colon B \to X$ and $b \colon C \to X$ such that $a \circ f = b \circ f$ there is a unique morphism $h \colon P \to X$ such that the diagram

$$\begin{tikzcd}
A \arrow[r, "f"] \arrow[d, "g"'] & B \arrow[d, "p_1"] \arrow[ddr, bend left, "a"] & \\
C \arrow[r, "p_2"'] \arrow[drr, bend right, "b"'] & P \arrow[dr, dashed, "h"] & \\
 & & X
\end{tikzcd}$$

commutes.

The requirement that for *any* maps a and b there is a map $h \colon P \to X$ enforces the condition that P be the smallest candidate, up to isomorphism; if there were another object P' that satisfied the same property as P, then P would map to P' and P' would map to P and by the uniqueness of the induced mappings P and P' would be isomorphic.

Example 1.7.11.

1. In Set, the pushout of the maps $\emptyset \to \{0,1,2\}$ and $\emptyset \to \{7,8,9\}$ is the set $\{0,1,2,7,8,9\}$.
2. More generally, the pushout in Set of the maps $\emptyset \to B$ and $\emptyset \to C$ is the disjoint union of B and C, i.e., the set consisting of all the elements of B and C.
3. In Set, the pushout of the maps

$$f \colon \{0,1\} \to \{3,4,5\} \qquad f(0) = 3,\ f(1) = 4$$

and

$$g \colon \{0,1\} \to \{a,b,c\} \qquad g(0) = g(1) = a$$

is the union of $\{3,4,5\}$ and $\{a,b,c\}$ with a identified with 3 and 4.

4. More generally, the pushout in Set of maps $A \to B$ and $A \to C$ is the set specified by taking the disjoint union of B and C and identifying $f(a)$ and $g(a)$.

As a set, the pushout in the category of topological spaces is described by the pushout in sets. However, we need to specify the topology on this identification. We have already seen how to perform this kind of construction in our discussion of the quotient topology.

Example 1.7.12. Let $f\colon A \to B$ be a continuous map of topological spaces. The pushout of the diagram

$$\begin{tikzcd}
A \arrow[r, "f"] \arrow[d] & B \\
* &
\end{tikzcd}$$

where $A \to *$ is the unique map taking all of A to $*$, is the quotient space generated by the partition of B given by $\{b\}$ for $b \in B - f(A)$ and $f(A)$. That is, the pushout is isomorphic to the quotient $B/f(A)$.

Example 1.7.13. Let B be a cylinder $S^1 \times [0,1]$, C a point $*$, and A be the circle S^1. Take $f\colon A \to B$ to be the inclusion $S^1 \to S^1 \times [0,1]$ specified by $x \mapsto (x,0)$ and $g\colon A \to *$ to be the unique map taking all $x \in S^1$ to the point $*$. Then the pushout

$$\begin{tikzcd}
S^1 \arrow[r, "f"] \arrow[d, "g"'] & S^1\times[0,1] \arrow[d] \\
* \arrow[r] & (S^1\times[0,1])/S^1
\end{tikzcd}$$

is a cone (see Figure 1.39).

The description of the quotient topology in terms of the pushout gives rise to the following interesting characterization.

Corollary 1.7.14. *Let $f\colon A \to B$ be a continuous map of topological spaces. A map from the quotient space $B/f(A) \to X$ is determined by a map $B \to X$ which takes all of A to a point.*

More generally, we use the quotient topology to describe the pushout in topological spaces.

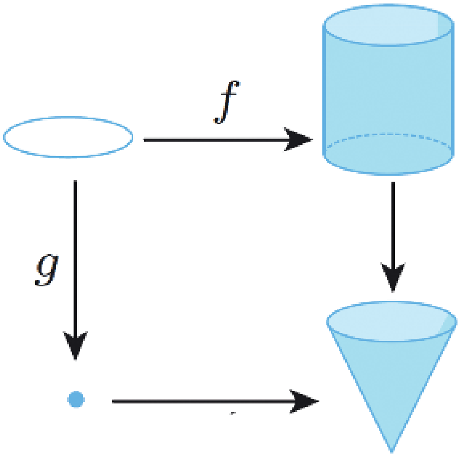

Figure 1.39 The cone can be formed by collapsing one end of a cylinder to a point.

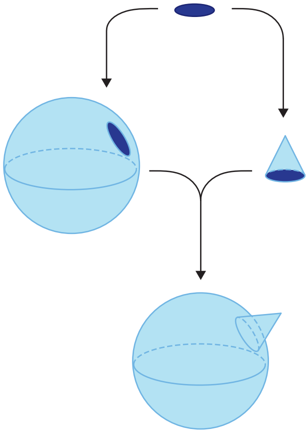

Figure 1.40 Gluing along a common subspace.

Example 1.7.15. The pushout of $f\colon A \to B$ and $g\colon A \to C$ is the quotient of the disjoint union $B \coprod C$ given by identifying the points $f(a)$ and $g(a)$ for each $a \in A$. For example, if f and g are injective, we look at the partition of $B \coprod C$ given by the points in $B \setminus f(A)$, the points in $C \setminus g(A)$, and all subsets of the form $\{f(a), g(a)\}$ for $a \in A$.

As this last example suggests, the gluing in CW complexes can also be described in terms of pushouts (Figure 1.40). Specifically, the constructions $D^n \coprod_f X_i$ arising in the description of CW complexes (in Definition 1.5.7) are precisely pushouts.

Example 1.7.16.

1. Let B and C be the subspaces of $\mathbb{R}^3$ defined as

$$B = \{(x, y, z) \mid x^2 + y^2 + z^2 = 1, z \geq 0\}$$

and

$$C = \{(x, y, z) \mid x^2 + y^2 + z^2 = 1, z \leq 0\},$$

and let A be the circle

$$\{(x, y, 0) \mid x^2 + y^2 = 1\}.$$

Take $f: A \to B$ and $g: A \to C$ to be the evident inclusions. Then the pushout is precisely the unit sphere

$$S^2 = \{(x, y, z) \mid x^2 + y^2 + z^2 = 1\}.$$

2. More generally, we have the following pushout diagram

$$\begin{tikzcd}
S^{n-1} \arrow[r] \arrow[d] & D^n \arrow[d] \\
D^n \arrow[r] & S^n
\end{tikzcd}$$

3. We can do the same kind of construction with solid disks and hemispheres; see Figure 1.41.

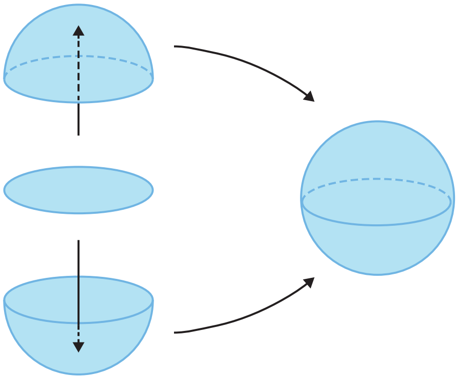

Figure 1.41 The solid sphere can be represented as the pushout of two hemispheres along a shared bounding disk.

1.7.1 Functors

For our purposes, perhaps the most important definition from category theory is the notion of a function between categories, called a *functor*. The topological invariants we study will all be functors from geometric categories to algebraic ones, for example, the function that assigns the set of path components to a topological space X.

Definition 1.7.17. Let $\mathcal{C}$ and $\mathcal{D}$ be categories. A *functor* $F\colon \mathcal{C} \to \mathcal{D}$ is specified by

1. a function
$$F\colon \mathrm{ob}(\mathcal{C}) \to \mathrm{ob}(\mathcal{D}),$$
2. for all $x, y \in \mathrm{ob}(\mathcal{C})$ a function
$$F\colon \mathrm{Hom}_{\mathcal{C}}(x, y) \to \mathrm{Hom}_{\mathcal{D}}(Fx, Fy)$$
such that $F(\mathrm{id}_x) = \mathrm{id}_{Fx}$ (the maps preserve the identity) and $Fg \circ Ff = F(g \circ f)$ (the maps are compatible with the composition).

We can reinterpret and strengthen Lemma 1.3.26 in this language.

Lemma 1.7.18. *The assignment of path components is a functor from the category* Top *to the category* Set.

Functorial constructions are ubiquitous in mathematics.

1. The functor Grp → Set that forgets the group structure is an example of a *forgetful functor*.
2. The functor Set → Grp that takes a set to the free group on generators the elements of the set is a functor.
3. The functor Top → Ho(Top) that takes each space to itself and each continuous map to its homotopy class is a functor.
4. The assignment of a vector space to its double dual and each linear transformation to its double dual transformation is a functor from Vect to itself. (The assignment of a vector space to its dual reverses the direction of the arrows, and specifies what is known as a *contravariant* functor.)

In the language of this section, we can now describe algebraic topology as the study of functors from Top to an algebraic category (e.g., Grp or Vect). For example, let Top_* be the category of *based spaces*, i.e., the objects are pairs (X, x) of a topological space and a “basepoint” $x \in X$ and a morphism $(X, x) \to (Y, y)$ is a continuous map $f: X \to Y$ such that $f(x) = y$. Then Lemma 1.6.29 can be interpreted and strengthened as the following assertion.

Lemma 1.7.19. *For $n > 0$, the construction $\pi_n(X, x)$ specifies a functor from* Top$_*$ *to* Grp.

All of the invariants we study will be functorial, and in fact we will see that the functoriality of our invariants is one of the essential facts that ensures their good properties in algorithmic contexts.

Remark 1.7.20. Correspondingly, one might hope to cast a certain amount of molecular biology as the study of suitable functors from genotype to phenotype. Here the initial problem of setting up categories of genotype and phenotype, where for instance morphisms might represent mutation and certain physical changes, is of basic interest.

The final notion we need from category theory is the idea of a *natural transformation*; this is a map between functors.

Definition 1.7.21. Let F and G be functors from $\mathcal{C}$ to $\mathcal{D}$. A natural transformation $\tau: F \to G$ is specified by:

1. a map $\tau_x: F(x) \to G(x)$ for every object $x \in \mathrm{ob}(\mathcal{C})$, and
2. commuting squares

$$\begin{tikzcd}
F(x) \arrow[r] \arrow[d, "\tau_x"] & F(y) \arrow[d, "\tau_y"] \\
G(x) \arrow[r] & G(y)
\end{tikzcd}$$

for every morphism $x \to y$ in $\mathrm{Hom}_{\mathcal{C}}(x, y)$.

Example 1.7.22.

1. The most important example for us comes in the context of functors $\mathbb{N} \to \mathcal{C}$, for a category $\mathcal{C}$. A functor $F: \mathbb{N} \to \mathcal{C}$ is specified by a sequence

$$F(0) \to F(1) \to F(2) \to \ldots,$$

and so a natural transformation $\tau\colon F \to G$ is determined by the commuting diagrams

$$\begin{tikzcd}
F(0) \arrow[r] \arrow[d, "\tau_0"] & F(1) \arrow[r] \arrow[d, "\tau_1"] & F(2) \arrow[r] \arrow[d, "\tau_2"] & \dots \\
G(0) \arrow[r] & G(1) \arrow[r] & G(2) \arrow[r] & \dots
\end{tikzcd}$$

2. For any category $\mathcal{C}$ and object $x \in \mathrm{ob}(\mathcal{C})$, there is a functor

$$\hom(x, -)\colon \mathcal{C} \to \mathrm{Set}$$

that takes an object y to the set of maps $\mathrm{Hom}(x, y)$ and a map $f\colon y_1 \to y_2$ to the map $\mathrm{Hom}(x, y_1) \to \mathrm{Hom}(x, y_2)$ induced by composition with f. Now, for any pair of functors $\hom(x_1, -)$ and $\hom(x_2, -)$, any map $x_2 \to x_1$ induces a natural transformation $\hom(x_1, -) \to \hom(x_2, -)$. (This is a version of the Yoneda lemma.)

1.8 Simplicial Complexes

Our most basic model of a geometric object is a topological space, which we introduced in Section 1.3. Topological spaces are too general to be feasible for algorithmic purposes, however. In Section 1.5, we introduced CW complexes, which are a more restrictive notion of a topological space; this data is a recipe for building a space from spheres and disks. Although CW complexes are an incredibly useful notion in modern algebraic topology, they are still not concise enough for algorithmic purposes. The issue is that describing the data of an attaching map $f\colon S_n \to X_n$ in general requires an infinite amount of information. That is, despite the fact that there are a limited number of building blocks, the instructions about how to glue them together are not simple enough.

We now describe an older model of topological spaces, the category of simplicial complexes, that is entirely discrete: here a space will be specified by gluing simple pieces together in a very small number of ways. As long as we are willing to work up to homotopy equivalence or weak homotopy equivalence, it will turn out that this is a general model of topological spaces. Our treatment follows the fantastic introduction given in [368].

Simplicial complexes are generalizations of graphs. And in this guise, there are many examples of simplicial complexes that are studied by systems biologists. For example, any of the networks that are described as graphs (e.g., protein interaction networks, regulatory networks, ecological interaction networks) are simplicial complexes. Thus, in a precise sense the theory we are developing here is a way to talk about higher dimensional networks.

Suppose that we are given points $\{x_0, \dots, x_k\}$ in $\mathbb{R}^n$. We will assume that these points satisfy the condition that the set of vectors in $\mathbb{R}^n$ represented by the differences

$$\{x_1 - x_0, x_2 - x_0, \ldots, x_k - x_0\}$$

are linearly independent. For example, a set $\{x_0, x_1, x_2\}$ will satisfy this condition if the points do not all lie on the same line.

Definition 1.8.1. The *k-simplex* spanned by the points $\{x_0, \ldots, x_k\}$ is the set of all points

$$z = \sum_{i=0}^{k} a_i x_i, \qquad \sum_{i=0}^{k} a_i = 1.$$

For a given z, we refer to a_i as the *ith barycentric coordinate*.

Example 1.8.2.

1. A 0-simplex is a point.
2. A 1-simplex is a line segment (with endpoints the points x_0 and x_1).
3. A 2-simplex is a triangle with vertices the points $\{x_0, x_1, x_2\}$.

(See Figure 1.42 for examples of geometric simplices.)

The simplices are the basic building blocks for a simplicial complex; roughly speaking, a simplicial complex is a collection of simplices glued along their edges (or "edges" of their edges).

Definition 1.8.3. The *interior* of a simplex S spanned by the points $\{x_0, \ldots, x_k\}$, denoted $\text{int}(S)$, is the subset of points where $a_i > 0$ for all the barycentric coordinates a_i. The *boundary* $\text{bd}(S)$ is defined to be $S \setminus \text{int}(S)$. (See Figure 1.43.)

It is straightforward to check that for any n-simplex S, there are homeomorphisms

$$\text{bd}(S) \cong S^{n-1} \qquad \text{and} \qquad S \cong D^{n+1}.$$

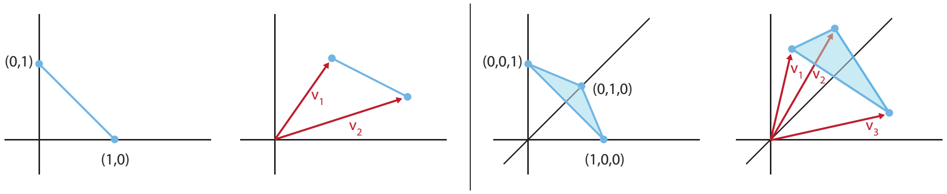

Figure 1.42 Geometric simplices specified by a set of vectors (including 0). On the left, the simplices are determined by the standard axial unit vectors; on the right, they are specified by the indicated vectors.

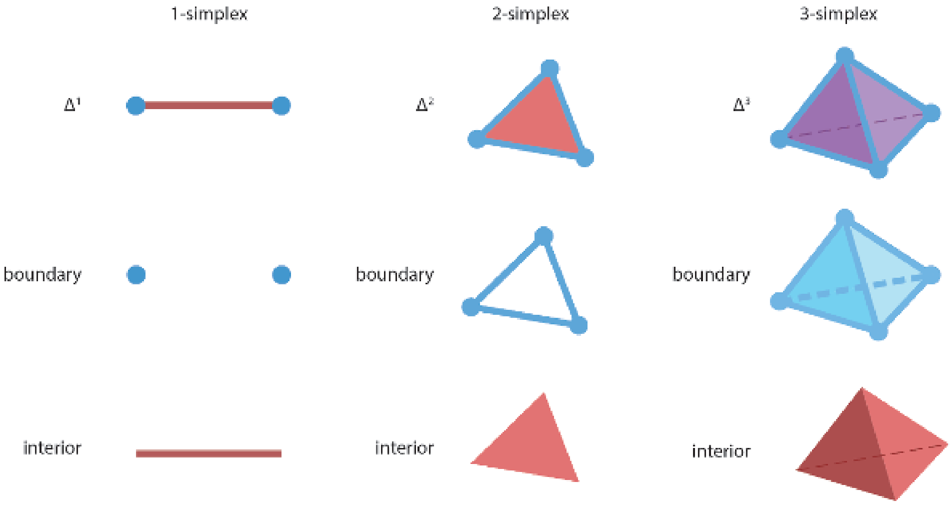

Figure 1.43 The boundary of a standard simplex is a combinatorial sphere; the interior is an open disk.

Therefore, there is a close analogy between gluing together simplices and building CW complexes. The advantage of working with simplices rather than CW complexes is that the boundaries of a simplex decompose into unions of simplices; we will be able to use a very restricted universe of attaching maps.

Definition 1.8.4. For a simplex S spanned by the points $P = \{x_0, \ldots, x_k\}$, a *face* of S refers to any simplex spanned by a subset of P.

Example 1.8.5.

1. There are no non-empty faces of a 0-simplex.
2. The non-empty faces of a 1-simplex determined by the points x_0 and x_1 are the two 0-simplices spanned by $\{x_0\}$ and $\{x_1\}$ respectively.
3. The non-empty faces of a 2-simplex determined by the points $\{x_0, x_1, x_2\}$ are the edges of the triangle and the vertices, the three 1-simplices determined by $\{x_0, x_1\}$, $\{x_1, x_2\}$, and $\{x_2, x_0\}$ and the three 0-simplices $\{x_0\}$, $\{x_1\}$, and $\{x_2\}$.

The following lemma is the key observation that allows us to glue together simplices in a simple way (Figure 1.44).

Lemma 1.8.6. *Let S be a simplex. The union of all of the faces of S is* $\mathrm{bd}(S)$.

We now define the notion of a simplicial complex.

Definition 1.8.7. A *simplicial complex X* in $\mathbb{R}^n$ is a set of simplices in $\mathbb{R}^n$ such that:

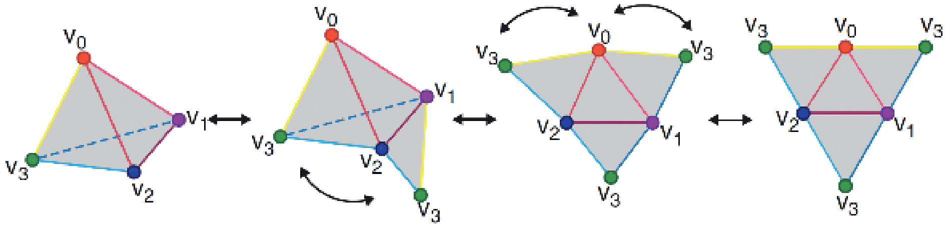

Figure 1.44 The boundary of the standard 3-simplex is a hollow pyramid; unfolding it makes clear how the 2-simplices that form the faces are glued along edges and vertices.

1. every face of a simplex in X is also a simplex in X, and
2. the intersection of two simplices in X is a face of each of them.

The zero simplices of a simplicial complex are referred to as the *vertices*. More generally, the collection of simplices of dimension at most k is referred to as the k-skeleton of the simplicial complex; we will denote the k-skeleton by X_k. For simplicity, we will restrict attention to simplicial complexes with finitely many simplices, referred to as *finite simplicial complexes*.

Definition 1.8.8. The *geometric realization* $|X|$ of a finite simplicial complex X is the topological space given by the union of simplices, given the subspace topology. (Here we regard the union as a subspace of $\mathbb{R}^n$.)

The geometric realization of a simplicial complex can be given the structure of a CW complex, where the cells correspond to the simplices and the attaching maps are determined by the faces.

Example 1.8.9. A circle can be given the structure of a simplicial complex (up to homeomorphism) in $\mathbb{R}^2$ where the 0-simplices are the points $(0, 0)$, $(1, 0)$, and $(1, 1)$ and the 1-simplices are the line segments specified by the equations

$$x + 0y = 1, \quad 0x + y = 0, \quad \text{and } x + y = 1,$$

where $x, y \in [0, 1]$. (In fact, as explained in Example 1.8.21, we can analogously model the circle with n 0-simplices and n 1-simplices connecting them for any n. See also Figure 1.45)

Remark 1.8.10. As with infinite CW complexes (recall Remark 1.5.8), we can make sense of the geometric realization of an infinite simplicial complex,

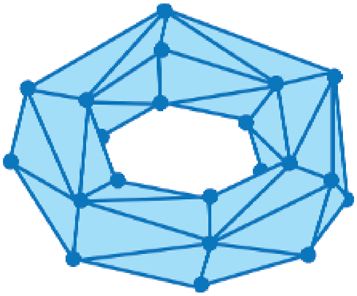

Figure 1.45 Up to homeomorphism, a torus can be triangulated as a simplicial complex.

but describing the topology is somewhat more complicated. However, all of the examples we consider in this book will be finite.

In a precise sense, a simplicial complex can be thought of as a higher dimensional generalization of a graph.

Example 1.8.11. A simplicial complex that has only 0-simplices and 1-simplices represents a graph embedded in Euclidean space, where the 0-simplices are the vertices and the 1-simplices are the edges.

We can assemble simplicial complexes into a category; for this purpose, we need an analogue of a continuous map.

Definition 1.8.12. Let X and Y be simplicial complexes. A *simplicial map* $f\colon X \to Y$ is specified by a map $X_0 \to Y_0$ such that whenever

$$\{z_0, \ldots, z_k\} \subset X_0$$

span a simplex of X,

$$\{f(z_0), f(z_1), \ldots, f(z_k)\}$$

span a simplex of Y.

Therefore, we can form a category with objects the simplicial complexes and morphisms the simplicial maps. It is useful to characterize the isomorphisms in this category.

Definition 1.8.13. Let X and Y be simplicial complexes. An *isomorphism of simplicial complexes* is a simplicial map $f\colon X \to Y$ that is a bijection on 0-simplices and such that for any $k > 1$, a collection of vertices $\{x_1, \ldots, x_k\}$ specifies a simplex of X if and only if $\{f(x_1), \ldots, f(x_k)\}$ is a simplex of Y.

Moreover, a simplicial map can be extended to a continuous map $f\colon |X| \to |Y|$ by linear interpolation:

$$f\left(\sum_{i=0}^{n} a_i x_i\right) = \sum_{i=0}^{n} a_i f(x_i).$$

Put another way, geometric realization is a functor.

Lemma 1.8.14. *Geometric realization specifies a functor from the category of simplicial complexes and simplicial maps to the category of topological spaces and continuous maps.*

One inconvenience with working with simplicial complexes as specified in Definition 1.8.7 is the dependence on a choice of embedding in some ambient Euclidean space $\mathbb{R}^n$. For example, ensuring that simplices intersect properly can require solving equations. Fortunately, it turns out that the data of a simplicial complex can be abstracted even further; all that is really important is the data of how many simplices there are and which faces they are glued along.

Definition 1.8.15. An *abstract simplicial complex* is a set X of finite non-empty sets such that if A is an element of X then so is every non-empty subset of A.

1. Each element of X represents a simplex; we refer to elements of X as (abstract) simplices.
2. The dimension of an abstract simplex A is $|A| - 1$, where here $| - |$ denotes the number of elements of a set.
3. Any non-empty subset of a simplex A is a face of A.
4. The vertices of X are the one-point sets in X. (Notice that any simplex of X is a union of vertices.)
5. More generally, we will denote the subset of X consisting of sets of cardinality $\leq k + 1$ as X_k, the k-skeleton.

We have a natural generalization of Definition 1.8.12 to the setting of abstract simplicial complexes.

Definition 1.8.16. A *map of abstract simplicial complexes* $f\colon X \to Y$ is specified by a map of sets $f\colon X_0 \to Y_0$ with the property that for any element $\{x_0, \ldots, x_k\}$ in X, $\{f(x_0), \ldots, f(x_k)\}$ is an element of Y.

Therefore, we have a category with objects the abstract simplicial complexes and morphisms the simplicial maps.

Definition 1.8.17. Let X and Y be abstract simplicial complexes. A simplicial map $f\colon X \to Y$ is an isomorphism if f is a bijection on 0-simplices and $\{x_0, \ldots, x_k\}$ is an element of X if and only if $\{f(x_0), \ldots, f(x_k)\}$ is an element of Y.

We now explain the relationship between abstract simplicial complexes and the simplicial complexes of Definition 1.8.7, which to be clear we will refer to as *geometric* simplicial complexes.

Lemma 1.8.18. *Let X be a geometric simplicial complex spanned by the points $x_0, \ldots, x_k \subseteq \mathbb{R}^n$. Then there is an associated abstract simplicial complex specified by the collection of subsets of the vertices of X which span a simplex in X.*

Two geometric simplicial complexes are isomorphic if and only if their associated abstract simplicial complexes are isomorphic. Moreover, every abstract simplicial complex can be uniquely associated to a geometric simplicial complex.

Theorem 1.8.19. *For every abstract simplicial complex S, there exists a geometric simplicial complex $\tilde{S}$ such that S is associated to $\tilde{S}$.*

The preceding theorem allows us to define the geometric realization of an abstract simplicial complex in terms of the geometric realization of the associated geometric simplicial complex. Once again, geometric realization is a functor.

Lemma 1.8.20. *The geometric realization of the associated simplicial complex specifies a functor $|-|$ from the category of abstract simplicial complexes and simplicial maps to the category of topological spaces and continuous maps.*

Example 1.8.21.

1. The abstract simplicial complex

$$\{\{v_0\}, \{v_1\}, \{v_2\}, \{v_0, v_1\}, \{v_1, v_2\}, \{v_2, v_0\}, \{v_0, v_1, v_2\}\}$$

describes the 2-simplex and its faces; the geometric realization has the homotopy type of a disk in $\mathbb{R}^2$.

2. Removing the interior from the previous example, the abstract simplicial complex

$$\{\{v_0\}, \{v_1\}, \{v_2\}, \{v_0, v_1\}, \{v_1, v_2\}, \{v_2, v_0\}\}$$

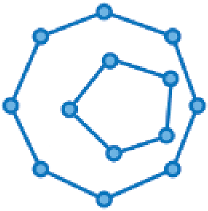

Figure 1.46 Two different models of the simplicial circle.

describes the boundary of the 2-simplex; the geometric realization has the homotopy type of a circle in $\mathbb{R}^2$. (In fact, Example 1.8.9 is homeomorphic to the geometric realization of this complex.)

3. More generally, we can make an abstract simplicial complex which models the circle using n vertices

$$\{v_0, v_1, \ldots, v_{n-1}\}$$

and n 1-simplices

$$\{\{v_0, v_1\}, \{v_1, v_2\}, \ldots, \{v_{n-2}, v_{n-1}\}, \{v_{n-1}, v_0\}\}.$$

(See Figure 1.46 for examples of this.)

4. The previous examples are all of two kinds; we can form the standard simplex Δ^n by taking a single n-simplex $[v_0, \ldots, v_n]$ and all of its subsets. The boundary $\partial\Delta^n$ is given by removing the n-simplex from the complex Δ^n.
5. Although computationally tractable, simplicial complexes describing even relatively simple surfaces can be large; see Figure 1.47 for a representation of a complex modeling a torus.

A next question one might wonder about is whether every topological space is homeomorphic or at least homotopy equivalent to a simplicial complex. In the case of homeomorphism, this kind of question turns out to be very difficult to answer. But for homotopy equivalence, there is a simple and satisfying criterion.

Proposition 1.8.22. *Let X be an abstract simplicial complex. The geometric realization |X| is a CW complex with an n-cell for each n-simplex of X.*

Proposition 1.8.23. *Let X be a CW complex. Then X is homotopy equivalent to the geometric realization of a simplicial complex K. Moreover, if X is a finite CW complex, then K can be taken to be a finite simplicial complex.*

Thus a topological space is homotopy equivalent to the geometric realization of a simplicial complex if and only if it is homotopy equivalent to a CW complex.

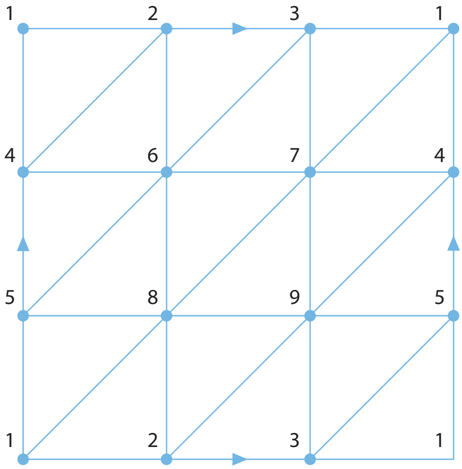

Figure 1.47 This diagram represents the vertices, 1-simplices, and 2-simplices of an abstract simplicial complex with realization homeomorphic to the torus. (Note that we identify the top edge with the bottom edge and the left edge with the right edge.)

But what about the morphisms? That is, can every continuous map $|X| \to |Y|$ be described as the geometric realization of a simplicial map? To be precise, we might ask the following question.

Question 1.8.24. Let X and Y be abstract simplicial complexes. Is every continuous map $|X| \to |Y|$ homotopic to the geometric realization of a simplicial map $X \to Y$?

As the question is posed, the answer is no.

Example 1.8.25. Let S_1 be the minimal abstract simplicial complex that models the circle; S_1 has vertices x_0, x_1, and x_2 and 1-simplices $\{x_0, x_1\}$, $\{x_1, x_2\}$, and $\{x_2, x_0\}$. If we consider simplicial maps from $S_1 \to S_1$, it is clear that there is no way to model the continuous maps $S^1 \to S^1$ given by $t \mapsto e^{kt(2\pi i)}$ for $k > 1$. That is, we cannot represent homotopy classes of maps that wrap the circle around itself more than once.

However, this deficiency can be repaired. The counterexample in Example 1.8.25 works because the "feature scale" of the domain is not fine enough. We can improve the situation using the notion of *subdivision*. In this case, if we use a model of the circle with n vertices and $n-1$ 1-simplices, as n increases we can represent maps which wrap around the circle more and more. More generally, we can subdivide any simplicial complex by dividing the simplices into unions of smaller simplices.

The resulting complex has geometric realization homeomorphic to the original one. Since we do not need these results we do not discuss them further here, but in fact there is a fundamental result (the simplicial subdivision theorem) that guarantees that any homotopy class of maps $|X| \to |Y|$ can be represented by a simplicial map from some subdivision of X to Y.

We now turn to the discussion of algebraic invariants of topological spaces that can be computed in terms of combinatorial operations on simplicial complexes. The oldest and simplest example of such an invariant is the Euler characteristic.

1.9 The Euler Characteristic

A basic and classical combinatorial invariant associated to a CW complex or an abstract simplicial complex is the Euler characteristic.

Definition 1.9.1. Let X be a finite CW complex, with cells of dimension at most n. The *Euler characteristic* of X is defined to be the alternating sum

$$\chi(X) = \sum_{i=0}^{n} (-1)^i k_i,$$

where k_i denotes the number of i-cells.

Equivalently, we can define the Euler characteristic of a finite simplicial complex directly.

Definition 1.9.2. Let X be a finite simplicial complex, with simplices of dimension at most n. The Euler characteristic of X is defined to be the alternating sum

$$\chi(X) = \sum_{i=0}^{n} (-1)^i k_i$$

where k_i denotes the number of i-simplices (see Figure 1.48).

It is straightforward to verify that these two notions are consistent under geometric realization.

The Euler characteristic is a very appealing invariant insofar as it does not depend on any information about the way in which cells or simplices are glued together, just their counts. As a consequence, it is very easy to compute. However, it is not completely clear from the definition what sorts of equivalences the Euler characteristic is preserved by. It is easy to see that χ is an isomorphism invariant for simplicial complexes.

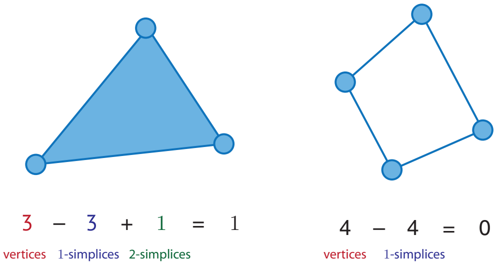

Figure 1.48 The Euler characteristic of a finite simplicial complex is computed as the alternating sum of the counts of simplices.

Lemma 1.9.3. *Let $f\colon X \to Y$ be an isomorphism of simplicial complexes. Then $\chi(X) = \chi(Y)$.*

But this is not tremendously useful; as we have seen in Example 1.8.21, there are many non-isomorphic models for the circle S^1. We would like there to be a well-defined Euler characteristic for "the circle" that does not depend on the simplicial model. Direct computation is encouraging, however – all the models of the circle have n vertices and n 1-simplices, and therefore have Euler characteristic 0. It turns out that $\chi(X)$ is a homotopy invariant for CW complexes.

Another concern about the Euler characteristic is that it does not reflect simplicial maps. The issue is simply that numbers are not rich enough to support functoriality. A central motivation for constructing invariants of topological spaces that land in algebraic categories (e.g., groups or vector spaces) is to provide enough structure for them to be functors.

1.10 Simplicial Homology

In this section, we finally develop the central invariant that we will use in topological data analysis, the homology groups. The homology groups will be a collection of functors indexed on the natural numbers

$$H_n\colon \mathrm{Simp} \to \mathrm{Vect}_{\mathbb{F}}, \quad n \geq 0.$$

Let X be an abstract simplicial complex. Roughly speaking, the homology groups of X are going to encode information about the way in which the simplices in successive dimensions are glued together. For the definition, we will need to pick an *orientation* for the simplices – in the case of a 1-simplex, this amounts to picking a direction for the line segment connecting the two vertices.

Let X be an abstract simplicial complex and σ a simplex. We will pick an ordering for the set of vertices in σ. Consider the case of a 2-simplex $[v_0, v_1, v_2]$.

Then there are six possible orderings: (v_0, v_1, v_2), (v_0, v_2, v_1), (v_1, v_0, v_2), (v_1, v_2, v_0), (v_2, v_0, v_1), and (v_2, v_1, v_0). However, we want to regard the possible choices of orientation for this 2-simplex as twofold, either clockwise or counterclockwise. We can express this by identifying orderings that are given by "rotations" of the vertices.

Definition 1.10.1. An *orientation* of the vertices of a simplex σ is an equivalence class of orderings of the vertices under the equivalence relation that two orderings are the same if they differ by an even permutation. (Recall that an even permutation is one that can be written as the composite of an even number of transpositions.)

Each k-simplex can be given one of two possible orientations for $k > 0$; there is only a single orientation for a vertex. We now assume that we have chosen orientations for the k-simplices of X; this can be done arbitrarily. We let $[v_0, \ldots, v_k]$ denote the oriented simplex specified by the vertices $\{v_0, \ldots, v_k\}$, where the orientation is specified by the ordering of the vertices.

1.10.1 Chains and Boundaries

We now explain the building blocks for the homology groups, the chain groups and the boundary homomorphism. These provide algebraic encodings of the combinatorial information of a simplicial complex. We start with the case of coefficients in a field $\mathbb{F}$, as this is most relevant for topological data analysis.

Definition 1.10.2. The *k-chains* $C_k(X; \mathbb{F})$ is the vector space with basis the set of oriented k-simplices. That is, elements of $C_k(X; \mathbb{F})$ are linear combinations of generators $\{g_\sigma\}$, where σ varies over the oriented k-simplices of X.

Example 1.10.3. Consider the abstract simplicial complex

$$X = \{[v_0], [v_1], [v_2], [v_0, v_1], [v_1, v_2]\}.$$

1. The space of 0-chains $C_0(X; \mathbb{F})$ for X is a vector space which is isomorphic to $\mathbb{F} \oplus \mathbb{F} \oplus \mathbb{F}$. We think of $C_0(X; \mathbb{F})$ as having elements of the form

$$a_0 v_0 + a_1 v_1 + a_2 v_2, \quad a_0, a_1, a_2 \in \mathbb{F},$$

   where generators correspond to the vertices v_0, v_1, and v_2 respectively.
2. The space of 1-chains $C_1(X; \mathbb{F})$ for X is a vector space which is isomorphic to $\mathbb{F} \oplus \mathbb{F}$. We think of $C_1(X; \mathbb{F})$ as having elements of the form

$$a_0 g_{01} + a_1 g_{12}, \quad a_0, a_1 \in \mathbb{F},$$

   where g_{01} and g_{12} are generators corresponding to the two 1-simplices of X.

3. The space of 2-chains $C_2(X;\mathbb{F})$ (and all higher chain groups) is the trivial vector space $\{0\}$ since there are no k-simplices for $k > 1$.

We now define a linear transformation $\partial_k \colon C_k(X;\mathbb{F}) \to C_{k-1}(X;\mathbb{F})$, the *boundary map*. As we will see, this is an algebraic way to encode the boundary of a simplex.

Definition 1.10.4. The linear transformation

$$\partial_k \colon C_k(X;\mathbb{F}) \to C_{k-1}(X;\mathbb{F})$$

is specified on the generators as

$$\partial_n([v_0,\ldots,v_k]) \mapsto \sum_{i=0}^{k} (-1)^i [v_0,\ldots,\hat{v}_i,\ldots,v_k]$$

where the $\hat{v}_i$ notation means we delete that vertex. The homomorphism is then specified by extending linearly to all of $C_k(X;\mathbb{F})$. (The orientation of the image is determined by the ordering of the vertices.)

Notice that this expression has a clear geometric interpretation: the boundary map applied to a simplex is precisely the alternating sum over the faces that make up the boundary of the simplex. (See Figure 1.49.)

Example 1.10.5.

1. The boundary of the 1-simplex $[v_0, v_1]$ is $v_1 - v_0$.
2. The boundary of the 2-simplex $[v_0, v_1, v_2]$ is $[v_1, v_2] - [v_0, v_2] + [v_0, v_1]$.

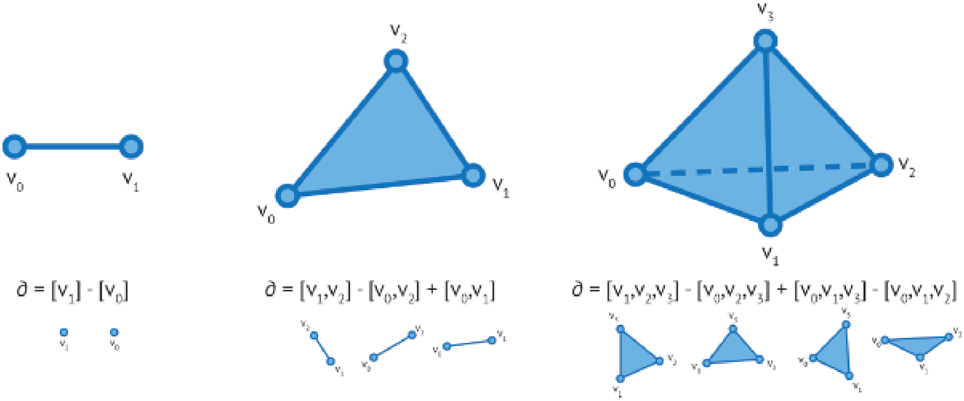
v0
v1
v2
v1
v0
v3
v0
v2
v1
∂ = [v1] - [v0]
∂ = [v1,v2] - [v0,v2] + [v0,v1]
∂ = [v1,v2,v3] - [v0,v2,v3] + [v0,v1,v3] - [v0,v1,v2]

Figure 1.49 The boundary map applied to a simplex is the alternating sum of the simplices along the boundary.

The boundary map has the special property that applying it twice is 0; "the boundary of a boundary is 0."

Lemma 1.10.6. *The composite $\partial_k \circ \partial_{k+1} = 0$.*

Checking this is an easy algebraic argument; the alternating signs result in cancellation.

Example 1.10.7. We compute $\partial_1 \circ \partial_2$ applied to the 2-simplex $[v_0, v_1, v_2]$. As in Example 1.10.5 above,

$$\partial_2([v_0, v_1, v_2]) = [v_1, v_2] - [v_0, v_2] + [v_0, v_1],$$

and applying ∂_1 we obtain

$$\begin{aligned}\partial_1\partial_2([v_0, v_1, v_2]) &= \partial_1([v_1, v_2]) - \partial_1([v_0, v_2]) + \partial_1([v_0, v_1]) \\ &= (v_2 - v_1) - (v_2 - v_0) + (v_1 - v_0) \\ &= v_2 - v_1 - v_2 + v_0 + v_1 - v_0 \\ &= 0.\end{aligned}$$

As an immediate corollary, we have the following.

Corollary 1.10.8. *For any simplicial complex X and natural number k,*

$$\operatorname{im}(\partial_{k+1}) \subseteq \ker(\partial_k).$$

1.10.2 Homology Groups

We now define the homology groups associated to the simplicial complex; the kth homology group H_k measures the failure of the inclusion of $\operatorname{im}(\partial_{k+1})$ in $\ker(\partial_k)$ to be an isomorphism. The idea of the homology groups is to take the subgroup of $C_k(X)$ of *cycles*, i.e., $\ker(\partial_k)$, and impose the equivalence relation that two chains c_1 and c_2 are *homologous* if their difference $c_1 - c_2$ is a *boundary*, i.e., if $c_1 - c_2$ is an element of $\operatorname{im}(\partial_{k+1})$.

Definition 1.10.9. The kth *homology group* with $\mathbb{F}$-coefficients $H_k(X; \mathbb{F})$ is defined to be the quotient group $\ker(\partial_k)/\operatorname{im}(\partial_{k+1})$. (In fact, this quotient group inherits the structure of a vector space.)

The zeroth homology group has a very natural interpretation.

Theorem 1.10.10. *Let X be an abstract simplicial complex. The homology group $H_0(X; \mathbb{F})$ is a vector space on generators in bijection with the path components of X.*

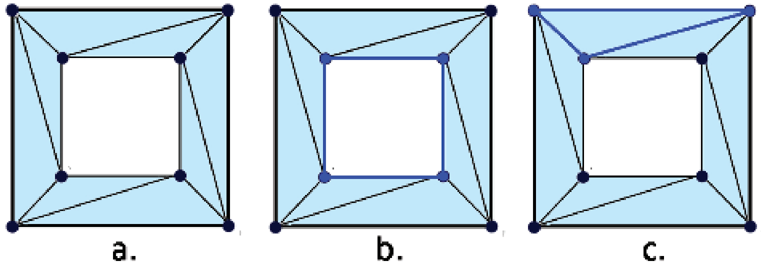

Figure 1.50 (a) gives a simplicial complex for an annulus. The blue paths in pictures (b) and (c) are examples of cycles in the complex. The cycle in picture (b) is not the boundary of any collection of simplices in the complex; it represents a non-zero class in the first homology group. In contrast, (c) is the boundary of a simplex and therefore is 0 in the homology group.

As we make precise below in Theorem 1.10.29, the first homology group is closely related to the fundamental group and hence to loops in X (see Figure 1.50).

Crudely, we can think of homology groups as the set of cycles in $C_k(X;\mathbb{F})$ that *are not* the boundaries of elements of $C_{k+1}(X;\mathbb{F})$. Roughly speaking, the fact that an element γ in $C_k(X;\mathbb{F})$ is a cycle means that it encloses a k-dimensional region, and the fact that γ is not a boundary means that the interior of the region is not part of the space X.

More precisely, consider the simplicial complex $\partial\Delta_k$, consisting of the boundary of the standard k-simplex. There is a cycle consisting of the alternating sum of the $(k-1)$-simplices; this is the boundary of the (missing) k-simplex. But this cycle cannot be a boundary, since there are no k-simplices. Thus, it specifies a class in the homology group H_{k-1}; this class detects the "hole." But if we fill the hole in, we get the standard simplex Δ_k, and now this cycle is clearly in the image of δ_k, and so vanishes in homology. More generally, given a simplicial complex X that contains $\partial\Delta_k$ but not the k-simplex, there will be a homology class representing that hole. Of course, this analysis does not directly apply to "larger" holes (with boundaries that are the union of many $(k-1)$-simplices), but a similar analysis does apply.

Summarizing:

1. H_0 a measure of path components of X,
2. H_1 is a measure of the one dimensional "holes" in X, and
3. more generally, H_k is a measure of k-dimensional geometric features of X, specifically, a count of the number of k-dimensional "holes" in X.

One of the advantages of simplicial homology is that it is easily computable given the data of an abstract simplicial complex. We illustrate this with some examples below.

Example 1.10.11.

1. Let S be the abstract simplicial complex $\{[v_0], [v_1], [v_0, v_1]\}$; this represents the interval. Then

$$\begin{cases} C_0(S\,;\mathbb{F}) \cong \mathbb{F} \oplus \mathbb{F}, \\ C_1(S\,;\mathbb{F}) \cong \mathbb{F}, \\ C_i(S\,;\mathbb{F}) = 0, \quad i > 1. \end{cases}$$

The boundary map $\partial_1 \colon C_1(S\,;\mathbb{F}) \to C_0(S\,;\mathbb{F})$ is specified by

$$1 \in \mathbb{F} \mapsto (1, -1) \in \mathbb{F} \oplus \mathbb{F}.$$

Then $H_0(S) = \mathbb{F}$, since $\ker(\partial_0)$ is all of $C_0(S\,;\mathbb{F})$ and the image of ∂_1 is $\mathbb{F}$. $H_1(S\,;\mathbb{F}) = 0$, as the kernel of ∂_1 is 0. And all $H_i(S\,;\mathbb{F}) = 0$ for $i > 1$.

Interpreting geometrically, this answer tells us that S represents a topological space that has one path component and no holes.

2. Let S be the abstract simplicial complex $\partial\Delta^2$, with vertices

$$\{[v_0], [v_1], [v_2]\}$$

and 1-simplices

$$\{[v_0, v_1], [v_1, v_2], [v_2, v_0]\}.$$

This complex is a model for the circle. Then

$$\begin{cases} C_0(S\,;\mathbb{F}) \cong \mathbb{F} \oplus \mathbb{F} \oplus \mathbb{F}, \\ C_1(S\,;\mathbb{F}) \cong \mathbb{F} \oplus \mathbb{F} \oplus \mathbb{F}, \\ C_2(S\,;\mathbb{F}) = 0, \quad i > 1. \end{cases}$$

Since $\partial_1([v_0, v_1]) = v_1 - v_0$, $\partial_1([v_1, v_2]) = v_2 - v_1$, and $\partial_1([v_2, v_0]) = v_0 - v_2$, it is straightforward to check that $\ker(\partial_1)$ is $\mathbb{F}$ with generator $[v_0, v_1] + [v_1, v_2] - [v_2, v_0]$. Therefore, $H_1(S) \cong \mathbb{F}$ and a similar argument shows that $H_0(S) \cong \mathbb{F}$. Specifically, $\ker(\partial_0)$ must be all of $C_0(S) \cong \mathbb{F} \oplus \mathbb{F} \oplus \mathbb{F}$. The computations of the image of ∂_1 above imply that in the quotient by $\mathrm{im}(\partial_1)$, we have that $v_0 = v_1$ since $v_0 + \partial_1([v_0, v_1]) = v_1$. Similarly, $v_1 = v_2$. Therefore, the quotient must be $\mathbb{F}$, generated by the coincident coset of v_0, v_1, and v_2. Interpreting geometrically, this example tells us that S represents a topological space that has one path component and one one-dimensional hole (see Figure 1.51).

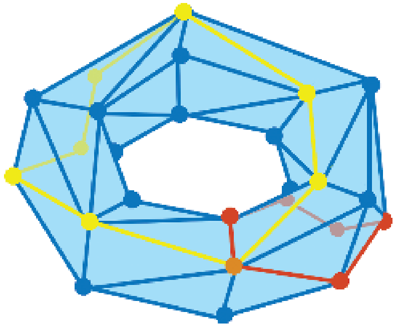

Figure 1.51 The red and yellow paths indicate representative generators for H_1 of the torus, which is $\mathbb{F} \oplus \mathbb{F}$.

3. More generally, for the simplicial complex Δ^{n+1} modeling S^n (i.e., the boundary of the standard $(n+1)$-simplex), we compute the answer

$$\begin{cases} H_0(S^n;\mathbb{F}) \cong \mathbb{F}, \\ H_n(S^n;\mathbb{F}) \cong \mathbb{F}, \\ H_k(S^n;\mathbb{F}) = 0, \quad k \neq 0, n. \end{cases}$$

This computation makes precise the sense in which we can think of the nth homology group as capturing information about n-dimensional holes.

Of particular relevance for topological data analysis is the fact that simplicial homology is algorithmically tractable; ∂_k can be expressed as a matrix where each column specifies the image in $C_{k-1}(S;\mathbb{F})$ of a generator of $C_k(S;\mathbb{F})$. We can then compute the image and kernel using linear algebra manipulations. Specifically, using Gaussian elimination we put ∂_k and ∂_{k+1} into Smith normal form; the rank of the homology group can then be computed in terms of the ranks of ∂_k and ∂_{k+1}.

Theorem 1.10.12. *Given a simplicial complex, there exists an algorithm to compute $H_k(-;\mathbb{F})$ whose running time is polynomial (cubic) in the total number of $(k+1)$-simplices, k-simplices, and $(k-1)$-simplices.*

1.10.3 Homology of Chain Complexes

The impressionistic description of the homology groups as computing information about k-dimensional holes strongly suggests that the groups H_k are homotopy invariants. To provide context for stating this kind of invariance result, it is useful to describe the homology groups as functors. As we have emphasized, much of the power of the invariants of algebraic topology comes because they are functorial. We can check directly from the definition that homology is in fact a functor

$$H_n \colon \mathrm{Simp} \to \mathrm{Vect}_{\mathbb{F}}.$$

Theorem 1.10.13. *Let X and Y be abstract simplicial complexes and let $f \colon X \to Y$ be a simplicial map. Then for each $k \geq 0$ there is an induced group homomorphism*

$$f_* \colon H_k(X; \mathbb{F}) \to H_k(Y; \mathbb{F}).$$

To explain this result, we provide an algebraic category to abstract the construction underlying homology. To this end, we now define the category $\mathrm{Ch}(\mathrm{Vect}_{\mathbb{F}})$ of chain complexes of $\mathbb{F}$-vector spaces.

Definition 1.10.14. A *chain complex of vector spaces $A_\bullet$* is a collection of vector spaces $\{\mathcal{A}_n\}$, for $n \in \mathbb{Z}$, and linear transformations

$$\partial_n \colon A_n \to A_{n-1}$$

such that $\partial_{n-1} \circ \partial_n = 0$. More succinctly, a chain complex is a functor $\mathbb{Z}^{\mathrm{op}} \to \mathrm{Vect}_{\mathbb{F}}$ satisfying the condition above on the successive composites of maps.

Having specified the objects of $\mathrm{Ch}(\mathrm{Vect}_{\mathbb{F}})$, we now need to explain the morphisms.

Definition 1.10.15. A *map of chain complexes $f \colon A_\bullet \to B_\bullet$* is a collection of linear transformations $f_n \colon A_n \to B_n$ for each $n \in \mathbb{Z}$ such that $f_{n-1} \circ \partial^A_n = \partial^B_n \circ f_n$, i.e., such that the diagrams

$$\begin{tikzcd}
\vdots \arrow[d, "\partial^A_{n+1}"'] & \vdots \arrow[d, "\partial^B_{n+1}"] \\
A_n \arrow[r, "f_n"] \arrow[d, "\partial^A_n"'] & B_n \arrow[d, "\partial^B_n"] \\
A_{n-1} \arrow[r, "f_{n-1}"] \arrow[d, "\partial^A_{n-1}"'] & B_{n-1} \arrow[d, "\partial^B_{n-1}"] \\
\vdots & \vdots
\end{tikzcd}$$

commute.

There is a natural functor $\mathrm{Vect}_{\mathbb{F}} \to \mathrm{Ch}(\mathrm{Vect}_{\mathbb{F}})$ that takes a vector space V to the chain complex

$$A_\bullet = \ldots \to 0 \to 0 \to V \to 0 \to 0 \to \ldots$$

where $A_0 = V$ and $A_i = 0$ for $i \neq 0$.

As we have seen, the category of topological spaces has several useful notions of equivalence: homeomorphisms, which are categorical isomorphisms, as well as homotopy equivalences and weak equivalences. In contrast, the algebraic category $\mathrm{Vect}_{\mathbb{F}}$ does not have a good analogue of the notion of homotopy equivalence. One of the advantages of $\mathrm{Ch}(\mathrm{Vect}_{\mathbb{F}})$ is precisely that is an algebraic category that enlarges $\mathrm{Vect}_{\mathbb{F}}$ enough to have a notion of homotopy equivalence, which is called *quasi-isomorphism*.

To explain a quasi-isomorphism of chain complexes, we need to observe that the definition of homology makes sense for arbitrary chain complexes. Notice that since by definition $\partial_n \circ \partial_{n+1} = 0$, we have the evident inclusion of groups

$$\mathrm{im}(\partial_{n+1}) \subseteq \ker(\partial_n).$$

We have the following general analogue of Definition 1.10.9.

Definition 1.10.16. For a chain complex $A_\bullet$, the nth *homology group* H_n is defined as the quotient

$$H_n(A_\bullet) = \ker(\partial_n)/\,\mathrm{im}(\partial_{n+1}).$$

The construction of homology is functorial.

Lemma 1.10.17. *A map $f\colon A_\bullet \to B_\bullet$ of chain complexes induces a linear transformation of vector spaces $H_n(A_\bullet) \to H_n(B_\bullet)$. Moreover, H_n specifies a functor from the category of chain complexes to the category of vector spaces.*

We think of the homology groups of a chain complex as akin to the homotopy groups of a space, and this leads to the following definition.

Definition 1.10.18. A map $f\colon A_\bullet \to B_\bullet$ of chain complexes is a quasi-isomorphism when each induced map $H_n(A_\bullet) \to H_n(B_\bullet)$ is an isomorphism.

Of course, if each map f_n is an isomorphism, then f is a quasi-isomorphism. But there are many examples of quasi-isomorphisms that are not isomorphisms.

Example 1.10.19. Consider the chain complex where $C_3 = \mathbb{F}$, $C_2 = \mathbb{F}$, all other $C_i = 0$, and $\partial_3 = \mathrm{id}$. Then the homology is zero for all n; this chain complex is quasi-isomorphic to the zero complex (i.e., the complex where $C_i = 0$ for all i).

For our purposes, the most interesting examples of chain complexes come from the construction of the simplicial chains. This assignment is functorial.

Lemma 1.10.20. *For a simplicial complex X, the chains $C_\bullet(X;\mathbb{F})$ form a chain complex of vector spaces. A simplicial map of simplicial complexes $f\colon X \to Y$ induces a chain map $C_\bullet(X;\mathbb{F}) \to C_\bullet(Y;\mathbb{F})$. That is, passage to simplicial chains induces a functor*

$$C_\bullet(-)\colon \mathrm{Simp} \to \mathrm{Ch}(\mathrm{Vect}_{\mathbb{F}}).$$

We can immediately deduce the functoriality of homology from this construction. An isomorphism of simplicial complexes clearly induces a quasi-isomorphism of chains; in fact, so does a homeomorphism of the associated topological spaces. But the power of homology arises because it is in fact a homotopy invariant. To explain this, we need to consider the question of when two maps $f, g\colon A_\bullet \to B_\bullet$ induce the same map on homology.

Definition 1.10.21. We say that two maps of chain complexes $f, g\colon A_\bullet \to B_\bullet$ are *chain homotopic* if there exist maps $h\colon A_n \to B_{n+1}$ such that $f_n - g_n = \partial_{n+1} \circ h_n - h_{n-1} \circ \partial_n$.

The definition of chain homotopy is a precise analogue of the notion of homotopy of maps of spaces.

Theorem 1.10.22. *If $f, g\colon X \to Y$ are simplicial maps of abstract simplicial complexes such that $|f|, |g|\colon |X| \to |Y|$ are homotopic, then the induced maps $f, g\colon C_\bullet(X;\mathbb{F}) \to C_\bullet(Y;\mathbb{F})$ are chain homotopic.*

In fact, Definition 1.10.21 can be derived by considering the chain complexes $C_\bullet(X \times [0,1];\mathbb{F})$ and $C_\bullet(Y;\mathbb{F})$ and the conditions imposed by the existence of a homotopy $h\colon X \times I \to Y$.

For our purposes, the most important fact about chain homotopic maps is the following result:

Proposition 1.10.23. *If two maps $f, g\colon A_\bullet \to B_\bullet$ of chain complexes are chain homotopic, then they induce the same map on homology.*

The point is simply that

$$\partial_n \circ (f_n - g_n) = \partial_n \circ \partial_{n+1} \circ h_n - \partial_n \circ h_{n-1} \circ \partial_n = \partial_n \circ h_{n-1} \circ \partial_n,$$

i.e., the difference between f_n and g_n is a boundary.

Corollary 1.10.24. *If $f\colon X \to Y$ is a simplicial map of abstract simplicial complexes such that $|f|\colon |X| \to |Y|$ is a homotopy equivalence, then f induces an isomorphism on homology.*

Put another way, we really have a functor

$$H_n\colon \mathrm{Ho(Simp)} \to \mathrm{Vect}_{\mathbb{F}},$$

where we define Ho(Simp) to be the category with objects abstract simplicial complexes and morphisms from X to Y specified by the homotopy classes of maps $|X| \to |Y|$.

Remark 1.10.25. Typically, this fact is proved using a related homology theory called singular homology, which coincides with simplicial homology but is (by definition) independent of the simplicial structure. In addition, as we mentioned in Remark 1.10.31, one can also define homology directly for CW complexes. In light of this menagerie of definitions, a basic consistency question arises: given an abstract simplicial complex X, do all the possible ways of defining its homology agree? Direct comparisons are possible, but it turns out that the collection of homology functors $H_n\colon \mathrm{Top} \to \mathrm{Vect}_{\mathbb{F}}$ can be axiomatically characterized in terms of a very simple set of axioms, the Eilenberg-Steenrod axioms. Roughly speaking, these axioms describe families of functors that have prescribed behavior on the spheres S^n and satisfy certain gluing relationships; the proof that this suffices to characterize the theories amounts to induction over a CW structure.

1.10.4 Simplicial Homology with Coefficients in an Abelian Group

In fact, simplicial homology can take values in the category of abelian groups instead of vector spaces. We consider the case of $\mathbb{Z}$ for clarity. Definitions 1.10.14 and 1.10.15 generalize immediately to the category Ch(Ab) of chain complexes of abelian groups.

Definition 1.10.26. A *chain complex of abelian groups* $A_\bullet$ is a collection of abelian groups $\{A_n\}$, for $n \in \mathbb{Z}$, and homomorphisms

$$\partial_n\colon A_n \to A_{n-1}$$

such that $\partial_{n-1} \circ \partial_n = 0$. More succinctly, a chain complex is a functor $\mathbb{Z} \to \mathrm{Ab}$ satisfying the condition above on the successive composites of maps. The morphisms are the maps of chain complexes, i.e., the collections of homomorphisms $f_n\colon A_n \to B_n$ such that $f_{n-1} \circ \partial_n^A = \partial_n^B \circ f_n$.

We can build the simplicial chains by working with the free abelian group generated by the simplices. Specifically, we have the following definition.

Definition 1.10.27. The group of m-chains $C_m(X; \mathbb{Z})$ is the free abelian group with basis elements in bijection with the oriented m-simplices. That is, elements of $C_m(X; \mathbb{Z})$ are linear combinations (with coefficients in $\mathbb{Z}$) of generators $\{g_\sigma\}$, where σ varies over the oriented m-simplices of X.

Lemma 1.10.28. *A map of simplicial complexes $S \to S'$ determines a chain map $C_\bullet(S; \mathbb{Z}) \to C_\bullet(S'; \mathbb{Z})$. That is, passage to simplicial chains induces a functor*

$$C_\bullet(-; \mathbb{Z})\colon \mathrm{Simp} \to \mathrm{Ch(Ab)}.$$

Applying homology, we get a composite functor

$$H_n(-; \mathbb{Z})\colon \mathrm{Simp} \to \mathrm{Ch(Ab)} \to \mathrm{Ab}.$$

We refer to this as homology with coefficients in the group $\mathbb{Z}$. As in Proposition 1.10.23, homotopy classes of maps induce the same map on homology and quasi-isomorphisms of chain complexes induce isomorphisms.

In this context, the first homology group can be described in terms of something we have already seen.

Theorem 1.10.29. *Let X be an abstract simplicial complex that is connected. The homology group $H_1(X; \mathbb{Z})$ is the abelianization of the fundamental group $\pi_1(X, x)$, where here the abelianization of a group is the quotient by the subgroup generated by terms of the form $xyx^{-1}y^{-1}$.*

The advantage of working with $\mathbb{Z}$ coefficients is that the homology captures more information about the space X. More generally, it is possible to consider homology with coefficients in any ring R; the situation for $\mathbb{Z}$ is a special case. However, when working with topological data analysis, only the cases of homology with field coefficients tend to be used. We explain the reason for this in Section 2.3. Just as for the case of field coefficients, there is an efficient algorithm for computing homology with coefficients in $\mathbb{Z}$.

Theorem 1.10.30. *Given a simplicial complex, there exists an algorithm to compute $H_k(-; \mathbb{Z})$ whose running time is polynomial (cubic) in the total number of $(k+1)$-simplices, k-simplices, and $(k-1)$-simplices.*

Remark 1.10.31. Because of our focus on algorithmic methods, we have discussed simplicial homology in this section. As we mentioned in Remark 1.10.25, there are a number of other candidate constructions of the homology of a space; for example, one can define homology using calculus (in terms of differential forms) or infinite-dimensional functions (singular homology). Most notably, in the spirit of our discussion, it is possible to give a definition of homology that works directly from the CW complex structure on a space; one begins with a chain complex defined where $C_k(X)$ is the free abelian group on the k-cells. However, the boundary map is considerably more complicated in this case. Nonetheless, this approach has been the basis for computational work in discrete Morse theory and computational cubical homology (e.g., see [228, 280]).

1.11 Manifolds

The definition of a topological space is very general; an arbitrary topological space can be extremely complicated and have a very non-geometric flavor. For example, the Cantor set, constructed by removing the middle third from the interval $[0, 1]$, then the middle third from each of the resulting intervals, and so on (i.e., the subset of $[0, 1]$ consisting of elements whose ternary expansion does not contain 1) is an exotic topological space. When we work up to weak homotopy equivalence, we can restrict attention to simplicial complexes, which are a much nicer collection of spaces. Nonetheless, simplicial complexes still admit a very wide collection of examples with complicated local geometry.

However, in many applications (e.g., computer vision, medical imaging, physics), particularly nice examples of topological spaces tend to arise; these are spaces which admit Euclidean coordinates, at least locally, and permit the definition of a precise generalization of classical calculus. Such a topological space is called a *manifold*. A wonderful introduction to smooth manifolds is given by Milnor's classic book [355]; for more on Riemannian manifolds see [96].

In order to define a manifold, we need to explain what we mean by coordinates.

Definition 1.11.1. Let X be a topological space. Given an open set $U \subseteq X$, we say that a *chart* is a homeomorphism $\theta\colon U \to V$, where V is an open subset of $\mathbb{R}^n$. The inverse θ^{-1} equips U with a coordinate system. (See Figure 1.52 for examples of charts.)

An *atlas* for X is a collection of charts such that the $\{U_i\}$ cover X. The composites

$$\theta_\alpha \theta_\beta^{-1}\colon \theta_\beta(U_\alpha \cap U_\beta) \to \theta_\alpha(U_\alpha \cap U_\beta)$$

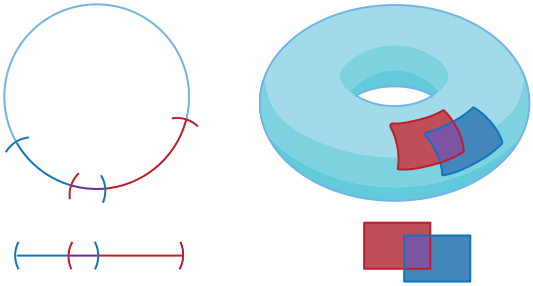

Figure 1.52 Left: Two overlapping charts on a circle. Right: Two overlapping charts on a torus. Each chart gives a little coordinate system, and transition functions connect these coordinates on the overlaps.

are referred to as transition functions. These explain how coordinates change as we move between different charts.

Definition 1.11.2. An n-dimensional *topological manifold* X is a second-countable, Hausdorff topological space equipped with an atlas where the charts are all subsets of $\mathbb{R}^n$.

(Here recall that second-countable means that the topological space has a countable base and Hausdorff means that any pair of points can be separated by enclosing open sets.)

It is often the case that examples have additional smoothness which permits the use of the methods of calculus. Since the transition functions involve maps from subsets of Euclidean space to itself, we can ask about their continuity and derivatives using the standard techniques of multivariable calculus.

Definition 1.11.3. An n-dimensional *smooth manifold* is a topological manifold where the transition functions are continuous and infinitely differentiable.

Many of the most familiar examples of topological spaces are manifolds.

Example 1.11.4.

1. Any Euclidean space $\mathbb{R}^n$ is a manifold, covered by a single chart.
2. The space $S^1 = \{(x, y) \subset \mathbb{R}^2 \mid x^2 + y^2 = 1\}$ is a manifold, covered by two charts, one covering points with $y > \frac{1}{2} - \epsilon$ and one covering points with $y < \frac{1}{2} + \epsilon$. (Here we can choose any $\epsilon > 0$.)
3. More generally, any sphere $S^n = \{(x_1, x_2, \ldots, x_{n+1}) \in \mathbb{R}^{n+1} \mid \sum_i x_i^2 = 1\}$ is a manifold covered by two charts.
4. The torus is a manifold; charts can be provided by considering a covering of the torus by little overlapping squares, for instance.

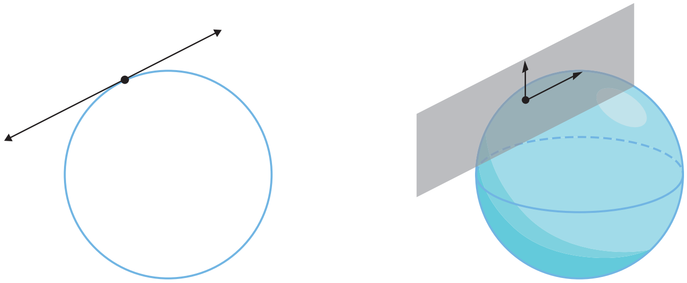

Figure 1.53 The tangent space at a point is all the directions in which a derivative of a curve could point; equivalently, it is the plane perpendicular to the normal or "outward" pointing direction.

Calculus on manifolds is expressed in terms of the notion of tangent spaces. At each point x of a manifold M, the *tangent space* T_xM is simply a vector space in which the tangent vectors (i.e., derivatives) to curves through that point can lie. The derivative of a function $f\colon M \to \mathbb{R}$ at a point $x \in M$ is a vector which lies in T_xM.

Example 1.11.5.

1. The tangent space $T_x\mathbb{R}^n$ of Euclidean space $\mathbb{R}^n$ at any point $x \in \mathbb{R}^n$ is isomorphic to $\mathbb{R}^n$.
2. The tangent space T_xS^1 at a point $x \in S^1$ is isomorphic to $\mathbb{R}^1$; the tangent space can be viewed as the tangent line to the circle.
3. The tangent space to T_xS^n to a sphere at a point $x \in S^n$ is a plane $\mathbb{R}^n$.

(See Figure 1.53 for a representation of the tangent space of spheres.)

For particularly nice manifolds (including the examples we have discussed above), the tangent spaces T_xM admit inner products that vary smoothly as we move around on M. Recall that an inner product (sometimes referred to as a dot product) is a pairing of the following form.

Definition 1.11.6. An *inner product* on a vector space V over the field $\mathbb{R}$ is a function

$$\langle -,-\rangle\colon V \times V \to \mathbb{R}$$

such that

1. $\langle x, x\rangle \geq 0$,
2. $\langle x, y\rangle = \langle y, x\rangle$, and
3. $\langle x + y, z\rangle = \langle x, z\rangle + \langle y, z\rangle$.

The significance of an inner product is that it allows us to define the length of a vector as the *norm* $\|x\| = \sqrt{\langle x, x\rangle}$ and the (cosine of the) angle between two vectors as being proportional to their inner product. That is, manifolds with inner products on the tangent spaces admit nice notions of area and angles; such manifolds are referred to as *Riemannian manifolds*. Riemannian manifolds have a number of rich geometric properties.

1. A path-connected Riemannian manifold has a metric; a path γ in M has a length computed by integrating the norms of the tangent vectors along γ. The distance between two points p and q is computed by taking the infimum (recall Definition 1.2.17) of the lengths of all paths joining them.
2. A Riemannian manifold has a notion of area or volume of regions on the manifold, referred to as the *volume form*, coming from the determinant in the tangent spaces.
3. A Riemannian manifold M has a notion of *curvature*, which can be described in terms of the divergence of paths following the tangent vectors at a point. For example, the standard sphere has curvature 1, Euclidean space has curvature 0, and hyperbolic space has curvature -1. (Here recall that hyperbolic space is a description of the geometry that arises when Euclid's parallel postulate is modified to allow infinitely many distinct parallel lines between two points.) We will say more about this below in Section 4.7.3.

Such manifolds allow a theory of integration and sampling, and although one does not expect data to lie on such manifolds, these provide a vital source of intuition and theoretical backing for the behavior of topological data analysis algorithms; such examples play an important motivating role, as we will see in Chapters 2 and 3.

Despite their rigidity, there are an enormous number of possible manifold topologies as the dimension increases; easy estimates show the number of homeomorphism classes of manifolds grows faster than exponentially as the dimension increases [533]. We can classify manifolds in low dimensions, however.

Example 1.11.7.

1. In dimension 0, the only manifolds are disjoint unions of points.
2. In dimension 1, the manifolds are homeomorphic to disjoint unions of circles and copies of $\mathbb{R}$. For example, a compact manifold with a single path component must be a circle.
3. In dimension 2, the classification of surfaces is an early and important theorem in topology; compact manifolds can be completely described as either a sphere or a manifold classified by a pair of natural numbers, describing how the manifold is made by gluing two kinds of basic pieces (toruses and projective planes) together. (For a nice treatment, see [369, §12].)

Another important class of examples comes from matrix groups, which are examples of Lie groups.

Example 1.11.8. Roughly speaking, a Lie group is a group which is also a topological space that is a manifold, so that the group operations are continuous. For example, the circle S^1 can be given the structure of a Lie group where the group operation is specified by adding angles. As another important example, the set $\mathrm{GL}_n(\mathbb{R})$ of invertible matrices can be given the structure of a manifold; such manifold symmetry groups are ubiquitous in physical applications.

On the one hand, manifolds provide geometric intuition for many methods in computational topology, and provide a large and familiar class of topological spaces. On the other hand, in contrast to physics, in applications to biology and genomics we do not usually expect the metric spaces we encounter to come from Riemannian manifold structures. In many cases, we do not even expect them to come from continuous topological spaces, in the sense that for many biologically relevant metrics, there is a minimum bound such that any distance is larger than this bound – for example, the Hamming distance between strings has this property.

One potential compromise between manifolds and arbitrary topological spaces comes from the theory of *stratified spaces*. Although a precise definition is more technical than we require, roughly speaking a stratified space is a topological space that is the union of manifolds (of possibly different dimensions) that fit together nicely. (See [534] for a wide-ranging treatment.)

Example 1.11.9.

1. Any graph embedded in Euclidean space is a stratified space comprising zero dimensional manifolds (points) and one dimensional manifolds (open intervals). Notably, trees are stratified spaces.
2. The disjoint union of manifolds $\coprod_i M_i$ is a stratified space.

1.12 Morse Functions and Reeb Spaces

A natural question to ask about a manifold is whether we can endow it with a CW structure which reflects the geometric structure of the manifold. A classical answer to this question is provided by *Morse theory*. Morse theory starts by considering a manifold M along with a "height function"

$$h\colon M \to \mathbb{R}.$$

Example 1.12.1. Consider the standard sphere

$$S^2 = \{(x, y, z) \in \mathbb{R}^3 \mid x^2 + y^2 + z^2 = 1\}.$$

We think of this as sitting on the tangent plane $z = -1$, and we can define the height at a point (x, y, z) as simply $z + 1$. (Of course, there are many other reasonable choices of height functions.)

Given a height function, the approach of Morse theory is to study the information about M encoded in the inverse images $f^{-1}(k)$ as k varies; specifically, we consider the inverse images for $k \in \mathbb{R}$, or more generally in the inverse images $f^{-1}(I)$, where $I \subseteq \mathbb{R}$ is an open interval (a, b). The places where the inverse images change in interesting ways turn out to be precisely the critical points of the function h. That is, the idea of Morse theory is that a space can be characterized by the critical points of suitable continuous functions from $M \to \mathbb{R}$.

Example 1.12.2. A standard example to consider is the torus "stood on its end," where the bottom has height 0. As a varies, the inverse images $h^{-1}([0, a])$ start as a disk, then become a cylinder, then the torus with a disk cut out, and then finally become the entire torus. From the perspective of homotopy theory, the process described is precisely cell attachment in a CW structure! Attaching occurs as h passes through a critical point. (See Figure 1.54.)

We do not need the full generality of Morse theory to explain the techniques of topological data analysis, so we do not give precise statements of the main theorems; for a beautiful treatment, see [354]. However, constructions inspired by this approach, the *Reeb graph* and *Reeb space*, have turned out to be incredibly useful in topological data analysis and computational geometry. We now give a brief overview of these constructions; see [459] for a more in-depth exposition.

Suppose that we are given a topological space X (e.g., a CW complex or the geometric realization of a finite simplicial complex) along with a continuous map $h \colon X \to \mathbb{R}^n$.

Definition 1.12.3. We define an equivalence relation on X by stipulating that $p \simeq q$ if $p, q \in f^{-1}(k)$ for some k and moreover that p and q are in the same path component. The *Reeb space* of X is the quotient of X by this equivalence relation.

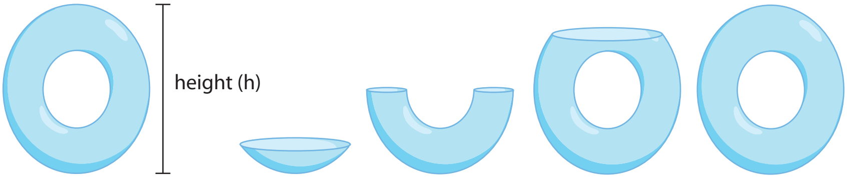

Figure 1.54 As the height increases, the inverse image includes more and more of the torus.

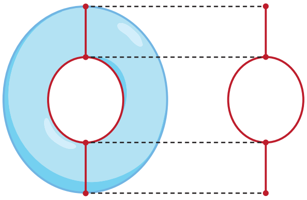

Figure 1.55 The Reeb graph of a torus.

When $n = 1$, Definition 1.12.3 yields a graph, referred to as the Reeb graph. The vertices of the Reeb graph correspond to components of the level sets, with edges connecting components that merge as k varies. See Figure 1.55 for the Reeb graph of the torus; notice the similarities to the Morse theory description above.

It is sometimes helpful in theoretical work to have a more general version of the Reeb space, referred to as the *categorical Reeb space*.

Definition 1.12.4. Given a topological space X equipped with a continuous function $h\colon X \to \mathbb{R}^n$, we specify the functor $R_{X,f}$ from the category of open sets of $\mathbb{R}^n$ with morphisms inclusions $U_1 \to U_2$ to the category of spaces as follows.

Let $R_{X,f}(I)$ be the space $f^{-1}(I)$, and let the induced map $R_{X,f}(I) \to R_{X,f}(J)$ be the evident inclusion $f^{-1}(I) \to f^{-1}(J)$.

Under good conditions, when $n = 1$ the Reeb graph of Definition 1.12.3 can be recovered from the categorical Reeb space of Definition 1.12.4 by applying π_0 to pass to components [459].

1.13 Summary

- Metric spaces, topological spaces, groups, and vector spaces are sets endowed with additional mathematical structure. These structures are the central objects upon which topological data analysis is built.
- A topological space $(X, \mathcal{U})$ is a set X endowed with a topology $\mathcal{U}$. We may describe the similarities of $(X, \mathcal{U})$ to other topological spaces by considering homeomorphisms and homotopy equivalences.
- We may construct topological spaces by gluing together simpler spaces such as cells (n-disks D^n) along their boundaries. Spaces produced in this way are called CW complexes. We may also create new topological spaces by considering the product of two or more smaller spaces (such as the torus in Example 1.5.11) or by collapsing subspaces of larger spaces, called a quotient (such as the cone in Figure 1.29).

- The fundamental group $\pi_1(X, x)$ of a topological space X is the set of homotopy classes of loops in X based at a fixed point $x \in X$. We may generalize the idea of the fundamental group to higher dimensions with the nth homotopy group $\pi_n(X, x)$ (see Definition 1.4.10). As the name suggests, $\pi_n(X, x)$ is a mathematical group under composition of rescaled maps (see Theorem 1.6.27). The homotopy groups of a space capture information about the space encoded in maps out of test spaces, namely spheres.
- Category theory provides a means of formalizing the notion of moving between different mathematical worlds. For example, we may use the language of category theory to restate the previous bulletpoint: $\pi_n(X, x)$ specifies a functor between the categories of based topological spaces and algebraic groups.
- Simplicial complexes provide a discrete, combinatorial framework for studying topological spaces. Many topological spaces arise as the geometric realizations of finite simplicial complexes.
- The combinatorial nature of simplicial complexes allows us to develop the idea of simplicial homology. For each $k \geq 0$, we may consider the group $C_k(X)$ of linear combinations of oriented k-simplices. Cycles in $C_k(X)$ may or may not form the boundaries of elements in $C_{k+1}(X)$. The kth homology group $H_k(X; \mathbb{Z})$ measures the size of the difference between $C_k(X)$ and the set of boundaries of elements in $C_{k+1}(X)$. That is, the kth homology group encodes information about the k-dimensional holes in X. The homology groups can be computed efficiently using linear algebra.
- Manifolds are topological spaces that are especially nice in that they admit Euclidean coordinates locally. Riemannian manifolds provide geometric structure: a metric, volume, and curvature.
- Given a function $f : X \to \mathbb{R}$, the Reeb space encodes information about topological changes in the level sets defined via inverse images.

1.14 Suggestions for Further Reading

The material we have covered in this section is standard, and in each of the previous sections, we have made suggestions about accessible treatments for readers who want more detail. For a reader who wants a geodesic path to the necessary background for topological data analysis, there are two sources to focus on: the first part of Munkres' book on simplicial homology [368], and Riehl's introductory textbook on category theory [428]. These are mostly self-contained (e.g., Munkres has a concise but detailed treatment of the required abstract algebra) and provide very lucid explanations.