Introduction to Applied Algebraic Topology

Tom Needham

Last Updated: December 10, 2017

These are lecture notes for the course MATH 4570 at the Ohio State University. They are a work in progress and certainly contain mistakes/typos. Please contact needham.71@osu.edu to report any errors or to make comments.

Work on these notes was supported by the NSF RTG grant Algebraic Topology and Its Applications, # 1547357.

Contents

1 Preliminaries

In this chapter we review some basic notions of set theory and equivalence relations. The reader is presumably familiar with these concepts, so this chapter should be treated mainly as a refresher and to fix notation.

1.1 Basic Set Theory

1.1.1 Set Theoretic Notation

A *set* is a collection of elements. We are taking the naive view of set theory in assuming that such a definition is intuitively clear and proceeding from there.

We generally use capital letters A, B, X, Y, etc. to denote sets and lower case letters a, b, x, y, etc. to denote their elements. We use $a \in A$ to denote that the element a belongs to the set A. The expression $a \notin A$ means that a is *not* an element of A. Two sets A and B are called *equal* if they contain exactly the same elements, in which case we write $A = B$. The contents of a set are specified by listing them or using *set builder notation*, as in the following examples.

Example 1.1.1.

1. $A = \{a, b, c\}$ denotes the set with three elements a, b and c.
2. $\mathbb{Z} = \{\ldots, -3, -2, -1, 0, 1, 2, 3, \ldots\}$ denotes the set of integers.
3. $B = \{b \in \mathbb{Z} \mid b \text{ is even}\} = \{b \in \mathbb{Z} \mid b = 2k \text{ for some } k \in \mathbb{Z}\}$ denotes the set of even integers.
4. $\varnothing$ denotes the *empty set* containing no elements.

We note that set notation doesn't account for multiplicity; that is, a set should not include more than one copy of any particular element, e.g., $\{a, a, b, c\}$. The notion of a set which respects multiplicity is called a *multiset*. Such objects will appear naturally later in the text, so we will treat them when they arise. Also note that sets are not ordered; for example, the set $\{a, b, c\}$ is equal to the set $\{b, c, a\}$.

1.1.2 Combining Sets

Let A and B be sets. The *union* of A and B is the set

$$A \cup B = \{x \mid x \in A \text{ or } x \in B\}.$$

The use of "or" here is non-exclusive. This means that the defining condition of $A \cup B$ can be read as "$x \in A$ or $x \in B$ or x is in both A and B".

The *intersection* of A and B is the set

$$A \cap B = \{x \mid x \in A \text{ and } x \in B\}.$$

The *difference* of A and B is the set

$$A \backslash B = \{x \mid x \in A \text{ and } x \notin B\}.$$

The *product* of A and B is the set of ordered pairs (x, y) such that $x \in A$ and $y \in B$, denoted

$$A \times B = \{(x, y) \mid x \in A \text{ and } y \in B\}.$$

Example 1.1.2. Let $A = \{a, b, c, d\}$ and $B = \{b, d, x, y\}$. Then

1. $A \cup B = \{a, b, c, d, x, y\}$ (note that b and d are not included twice!),
2. $A \cap B = \{b, d\}$,
3. $A \backslash B = \{a, c\}$,
4. $A \times B = \{(a, b), (a, d), (a, x), (a, y), (b, b), (b, d), (b, x), (b, y), \ldots\}$ (there are $16 = 4 \times 4$ elements total in the set).

Example 1.1.3. For any set A,

1. $A \cup \varnothing = A$,
2. $A \backslash \varnothing = A$,
3. $A \cap \varnothing = \varnothing$,
4. $A \times \varnothing = \varnothing$.

The set A is called a *subset* of B if $a \in A$ implies $a \in B$. This is denoted $A \subset B$.

Example 1.1.4. Let $A = \{a, b\}$, $B = \{a, b, c, d\}$ and $C = \{b\}$. Then $A \subset B$ but $A \not\subset B$ (this should be read as "A is not a subset of B") because $a \in A$ but $a \notin C$, so the defining implication fails.

1.1.3 Sets of Sets

The elements of a set can themselves be sets!

Example 1.1.5. Let $B = \{\{a\}, \{b\}, x\}$. Then the elements of B are $\{a\}$, $\{b\}$ and x. Let $A = \{a, b\}$. Then

$$A \cup B = \{a, b, \{a\}, \{b\}, x\}.$$

Note that this doesn't contradict our convention that a set can't contain multiple copies of the same element, since a and $\{a\}$ represent different objects. We also have

$$A \cap B = \varnothing,$$

since, for example, a and $\{a\}$ are different elements.

The *power set* of A, denoted $\mathcal{P}(A)$ is the set of all subsets of A.

Example 1.1.6. Let $A = \{a, b, c\}$. Then

$$\mathcal{P}(A) = \{\varnothing, A, \{a\}, \{b\}, \{c\}, \{a, b\}, \{a, c\}, \{b, c\}\}.$$

The *cardinality* of a set A is the number of elements in A. If A contains infinitely many elements, we say that its cardinality is infinity and that A is an infinite set. Otherwise we say that A is a finite set. We denote the cardinality of A by $|A|$.

Proposition 1.1.1. *If A is a finite set, then so is $\mathcal{P}(A)$ and*

$$|\mathcal{P}(A)| = 2^{|A|}.$$

Proof. Let $A = \{a_1, \ldots, a_n\}$, so that $|A| = n$. To form a subset B of A, we have to make the binary choice of whether or not to include each a_j in B. There are n such choices to make and they are independent of one another, so

$$|\mathcal{P}(A)| = 2^n = 2^{|A|}.$$

$\square$

1.1.4 Functions on Sets

A *function* from the set A to the set B is a subset $f \subset A \times B$ such that each $a \in A$ appears in exactly one ordered pair $(a, b) \in f$. The more typical notation used for a function is $f : A \to B$, with $f(a) = b$ denoting $(a, b) \in f$. This is a precise way to say that f maps each element of A to exactly one element of B. The set A is called the *domain* of f and B is called the *range* B of f.

Example 1.1.7.

1. All of the usual functions from Calculus are functions in this sense. For example the function $f(x) = x^2$ should be thought of as the function $f : \mathbb{R} \to \mathbb{R}$ with $(x, x^2) \in f \subset \mathbb{R} \times \mathbb{R}$.

2. Let $A = \{a, b, c\}$ and $B = \{x, y\}$. Then the set $f = \{(a, x), (b, x), (c, y)\} \subset A \times B$ defines a function $f : A \to B$. In this case, we would write $f(a) = x$, $f(b) = x$ and $f(c) = y$.

A function $f : A \to B$ is called *injective* (or *one-to-one*) if $f(a) = f(a')$ if and only if $a = a'$. It is called *surjective* (or *onto*) if for all $b \in B$ there exists $a \in A$ such that $f(a) = b$. The function is called *bijective* if it is both injective and surjective. An *inverse* to f is a function $f^{-1} : B \to A$ such that $f^{-1}(f(a)) = a$ for all $a \in A$ and $f(f^{-1}(b)) = b$ for all $b \in B$.

Proposition 1.1.2. *A function $f : A \to B$ has an inverse if and only if f is a bijection. If an inverse for f exists, then it is unique.*

Proof. If f is a bijection, define $f^{-1} : B \to A$ by $f^{-1}(b) = a$, where $a \in A$ satisfies $f(a) = b$. Such an a must exist by surjectivity and it is unique by injectivity. In fact, if an inverse for f exists, then it must be exactly of this form and this shows that inverses are unique. Conversely, suppose that f^{-1} exists. Let $a, a' \in A$ satisfy $f(a) = f(a')$. Then $f^{-1}(f(a)) = f^{-1}(f(a'))$ and this implies $a = a'$, thus f is injective. To show surjectivity, let $b \in B$. Then $a = f^{-1}(b)$ satisfies $f(a) = f(f^{-1}(b)) = b$. □

1.2 Infinite Sets

1.2.1 Countable and Uncountable Sets

Let $\mathbb{R}$ denote the set of real numbers. The sets $\mathbb{Z}$ and $\mathbb{R}$ both have cardinality infinity, but they feel different in the sense that $\mathbb{Z}$ is "discrete" while $\mathbb{R}$ is "continuous". We will make this difference precise in this section.

Let $\mathbb{N} = \{1, 2, 3, \ldots\}$ denote the *natural numbers* . A set A is called *countable* if there exists an injective function $f : A \to \mathbb{N}$ and *countably infinite* if it is countable and has infinite cardinality. If a set is not countable, then we call it *uncountable* .

Example 1.2.1. 1. $\mathbb{Z}$ is countably infinite. To see this, define $f : \mathbb{Z} \to \mathbb{N}$ by

$$f(k) = \begin{cases} 2k & \text{if } k > 0 \\ -2k+1 & \text{if } k \leqslant 0. \end{cases}$$

Then it is easy to check that f is injective.

2. The set $\mathbb{Q}$ of rational numbers is countably infinite. We leave this as an exercise.

Lemma 1.2.1. *If a set A is countable, then there exists a surjective map $g : \mathbb{N} \to A$.*

Proof. Assume that A is countably infinite (if $|A|$ is finite, the existence of such a map g is obvious). Let $f : A \to \mathbb{N}$ be an injection and let $B = \{f(a) \mid a \in A\}$ denote the image of f. For each $b \in B$, there exists a unique (by injectivity) element $a \in A$ such that $f(a) = b$; we denote this element by $f^{-1}(b)$. We fix any element $b_0 \in B$ and define a map $g : \mathbb{N} \to A$ by

$$g(k) = \begin{cases} f^{-1}(k) & \text{if } k \in B \\ f^{-1}(b_0) & \text{if } k \notin B. \end{cases}$$

Then g is a well-defined, surjective function. □

Remark 1.2.2. *The converse of this lemma is also true, but requires the Axiom of Choice. We wish to avoid treating the Axiom of Choice for now, but the interested reader is invited to read the Appendix for a short discussion of one of its important consequences.*

Let S denote the set of ordered sequences of 1's and 0's. That is, elements of S are of the form $(1, 1, 1, 1, \ldots)$ or $(0, 0, 0, 0, \ldots)$ or $(1, 0, 1, 1, 0, 0, 0, 1, 0, 1, \ldots)$, etc.

Theorem 1.2.3. *The set S is uncountable.*

The proof of the theorem uses a technique called *Cantor's Diagonal Argument.*

Proof. Let $g : \mathbb{N} \to S$ be any function. We wish to show that g is not surjective. Since g was arbitrary, it follows from Lemma 1.2.1 that S is uncountable. We list the values of g as

$$\begin{aligned} g(1) &= (a_1^1, a_2^1, a_3^1, \ldots), \\ g(2) &= (a_1^2, a_2^2, a_3^2, \ldots), \\ g(3) &= (a_1^3, a_2^3, a_3^3, \ldots), \\ &\vdots \end{aligned}$$

with each $a_j^k \in \{0, 1\}$. We define an element $s \in S$ by

$$s = (a_1^1 + 1, a_2^2 + 1, a_3^3 + 1, \ldots),$$

where we add mod 2, i.e., $0 + 1 = 1$ and $1 + 1 = 0$. Then for all $k \in \mathbb{N}$,

$$g(k) = (a_1^k, a_2^k, \ldots, a_{k-1}^k, a_k^k, a_{k+1}^k, \ldots) \neq (a_1^1+1, a_2^2+1, \ldots, a_{k-1}^{k-1}+1, a_k^k+1, a_{k+1}^{k+1}+1, \ldots) = s,$$

because $g(k)$ and s differ in their k-th entry. It follows that g is not surjective. $\square$

An essentially straightforward corollary is left to the reader:

Corollary 1.2.4. *The set $\mathbb{R}$ of real numbers is uncountable.*

1.2.2 Arbitrary Unions and Intersections

We will frequently need to consider infinite collections of sets. We use the notation

$$\mathcal{U} = \{U_\alpha\}_{\alpha \in \mathcal{A}}.$$

Each U_α is a set, α is an *index* for the set, and $\mathcal{A}$ is an *indexing set*. We can consider unions and intersections of sets in this collection, which are denoted, respectively, by

$$\bigcup_{\alpha \in \mathcal{A}} U_\alpha \quad \text{and} \quad \bigcap_{\alpha \in \mathcal{A}} U_\alpha.$$

Example 1.2.2. Let U_n denote the interval $(1/n, 1] \subset \mathbb{R}$, where n is a natural number. We can consider the collection

$$\mathcal{U} = \{U_n\}_{n \in \mathbb{N}}.$$

The union of the elements of this collection is

$$\bigcup_{n \in \mathbb{N}} U_n = (0, 1].$$

To see this, note that any element r of the union must be an element of some $(1/n, 1] \subset (0, 1]$, so the union is a subset of $(0, 1]$. On the other hand, for every $r \in (0, 1]$, there exists some $n \in \mathbb{N}$ such that $1/n < r$ and it follows that $r \in U_n$, so that r is an element of the union.

The intersection of the elements of this collection is

$$\bigcap_{n\in\mathbb{N}} U_n = \{1\}.$$

1.3 Equivalence Relations

Let S be a set. A *binary relation* on S is a subset $R \subset S \times S$. We typically use the notation $x \sim x'$ or $x \sim_R x'$ to indicate that $(x, x') \in R$. A binary relation R is an *equivalence relation* if the following conditions hold:

1. (Reflexivity) $s \sim s$ for all $s \in S$
2. (Symmetry) $s \sim s'$ if and only if $s' \sim s$
3. (Transitivity) if $s \sim s'$ and $s' \sim s''$, then $s \sim s''$.

Example 1.3.1. Consider the set $\mathbb{R}$ of real numbers. The most obvious equivalence relation on $\mathbb{R}$ is equality; that is $x \sim y$ if and only if $x = y$. Another equivalence relation $\sim_2$ is defined by $x \sim_2 y$ if and only if $x - y$ is an integer multiple of 2. We can define a similar equivalence relation $\sim_r$ for any fixed $r \in \mathbb{R}$. You will examine some other equivalence relations on $\mathbb{R}$ in the exercises.

Let $\sim$ be a fixed equivalence relation on a set S. The *equivalence class* of $s \in S$, denoted $[s]$, is the set

$$\{s' \in S \mid s' \sim s\}.$$

We denote the set of all equivalence classes of S by $S/\sim$. That is,

$$S/\sim = \{[s] \mid s \in S\}.$$

Example 1.3.2. Consider the equivalence relation $\sim_2$ restricted to the set of integers $\mathbb{Z}$; that is, integers a and b satisfy $a \sim_2 b$ if and only if $a - b$ is an integer multiple of 2. Then the set of equivalence classes $\mathbb{Z}/\sim_2$ contains exactly two elements $[0]$ and $[1]$. Indeed, for any even integer $2k$, $2k \sim_2 0$ so that $[2k] = [0]$. Likewise, for any odd integer $2k + 1$, $[2k + 1] = [1]$.

For the equivalence relation $\sim_2$ on all of $\mathbb{R}$, the set of equivalence classes $\mathbb{R}/\sim_2$ is in bijective correspondence with the interval $[0, 2)$. This is the case because for any real number x, there is a unique $y \in [0, 2)$ such that $x - y$ is an integer multiple of 2. To see this, note that the set $\cup_{k\in Z}[0 + k, 2 + k)$ is a partition of $\mathbb{R}$, so there exists a unique $k \in Z$ such that $x \in [0 + k, 2 + k)$, and we define $y = x - 2k$.

1.4 Exercises

1. Show that the set of rational numbers $\mathbb{Q}$ is uncountable by finding an injective map $\mathbb{Q} \to \mathbb{N}$.

2. Show that a subset of a countable set must be countable.
3. Show that if there is a bijection between sets A and B, then A is countable if and only if B is countable.
4. Prove that $\mathbb{R}$ is uncountable. One suggested strategy is to show that there is a bijection from the set S (from Theorem 1.2.3) to the interval $(0, 1)$ and to then apply Theorem 1.2.3 and the results of the previous two exercises.
5. Let $r \in \mathbb{R}$ be a fixed real number. Define a binary relation $\sim_r$ on $\mathbb{R}$ by declaring $x \sim_r y$ if and only if $x - y$ is an integer multiple of r. Show that $\sim_r$ is an equivalence relation.
6. Show that $\leqslant$ is *not* and equivalence relation on $\mathbb{R}$.
7. Let M_n denote the set of $n \times n$ matrices with real entries. Define a binary relation $\sim$ on M_n by declaring $A \sim B$ if and only if $A = B^T$, where the superscript denotes matrix transpose. Show that $\sim$ is an equivalence relation.

2 Review of Linear Algebra

In this chapter we review the basic concepts of linear algebra. A strong grasp of abstract linear algebra will be essential for the latter material in these notes. For a more in-depth treatment of linear algebra, see, for example, [3, 4].

2.1 Abstract Vector Spaces

2.1.1 Vector Spaces over $\mathbb{R}$

Definition of a Vector Space

A *vector space over* $\mathbb{R}$ is a set V together with an operation

$$\begin{aligned} + : V \times V &\to V \\ (v_1, v_2) &\mapsto v_1 + v_2 \end{aligned}$$

called *vector addition* and an operation

$$\begin{aligned} \cdot : \mathbb{R} \times V &\to V \\ (\lambda, v) &\mapsto \lambda \cdot v \end{aligned}$$

called *scalar multiplication* such that the following axioms are satisfied:

1. (Additive Associativity) For any elements v_1, v_2 and v_3 in V,

$$(v_1 + v_2) + v_3 = v_1 + (v_2 + v_3).$$

2. (Scalar Multiple Associativity) For any λ_1 and λ_2 in $\mathbb{R}$ and any v in V,

$$\lambda_1 \cdot (\lambda_2 \cdot v) = (\lambda_1 \lambda_2) \cdot v.$$

3. (Additive Commutativity) For any v_1 and v_2 in V,

$$v_1 + v_2 = v_2 + v_1.$$

4. (Additive Identity) There exists an element $0_V \in V$ called the *additive identity* such that for any $v \in V$,

$$v + 0_V = v.$$

5. (Additive Inverse) For any $v \in V$, there exists an element $-v \in V$ called the *additive inverse of* v such that
$$v + (-v) = 0_V.$$

6. (Distributive Law I) For any $\lambda \in \mathbb{R}$ and v_1 and v_2 in V,
$$\lambda \cdot (v_1 + v_2) = \lambda \cdot v_1 + \lambda \cdot v_2.$$

7. (Distributive Law II) For any λ_1 and λ_2 in $\mathbb{R}$ and any $v \in V$,
$$(\lambda_1 + \lambda_2) \cdot v = \lambda_1 \cdot v + \lambda_2 \cdot v.$$

8. (Scalar Multiple Identity) For any v in V,
$$1 \cdot v = v.$$

Examples of Vector Spaces over $\mathbb{R}$

Example 2.1.1. The example of a vector space over $\mathbb{R}$ that you are probably most familiar working with is $\mathbb{R}^n$ (for some positive integer n). That is, $\mathbb{R}^n$ is the set of n-tuples of real numbers $(x_1, x_2, \ldots, x_n)$, vector addition is given by

$$(x_1, x_2, \ldots, x_n) + (y_1, y_2, \ldots, y_n) = (x_1 + y_1, x_2 + y_2, \ldots, x_n + y_n)$$

and scalar multiplication is given by

$$\lambda \cdot (x_1, x_2, \ldots, x_n) = (\lambda x_1, \lambda x_2, \ldots, \lambda x_n).$$

It is an easy but instructive exercise to check that these operations satisfy the axioms of a vector space over $\mathbb{R}$.

Familiarity with $\mathbb{R}^n$ gives good intuition for working with abstract vector spaces, but the material for this course will require us to work with more exotic vector spaces.

Example 2.1.2. Let $P_n(\mathbb{R})$ denote the set of degree-n polynomials with real coefficients. Any element of $P_n(\mathbb{R})$ can be written in the form $a_n x^n + a_{n_1} x^{n-1} + \cdots + a_1 x + a_0$ for some $a_i \in \mathbb{R}$. The set $P_n(\mathbb{R})$ forms a vector space with vector addition

$$\begin{aligned}(a_n x^n + a_{n-1} x^{n-1} + \cdots + a_0) &+ (b_n x^n + b_{n-1} x^{n-1} + \cdots + b_0) \\ &= (a_n + b_n) x^n + (a_{n-1} + b_{n-1}) x^{n-1} + \cdots + (a_0 + b_0)\end{aligned}$$

and scalar multiplication

$$\lambda \cdot (a_n x^n + a_{n-1} x^{n-1} + \cdots + a_0) = (\lambda a_n) x^n + (\lambda a_{n-1}) x^{n-1} + \cdots + (\lambda a_0).$$

In one of the exercises you will show that these operations on $P_n(\mathbb{R})$ satisfy vector space axioms. You may notice that the operations of $P_n(\mathbb{R})$ share some similarity with those of

R^{n+1}. Indeed, we will see later that $P_n(\mathbb{R})$ and $\mathbb{R}^{n+1}$ are actually "equivalent" as vector spaces in a precise sense (to be defined in Section 2.3.1).

Example 2.1.3. As one last even more exotic example, consider the set of differentiable functions

$$C^\infty([0,1],\mathbb{R}) = \{f : [0,1] \to \mathbb{R} \mid f \text{ is infinitely differentiable}\}.$$

We claim that this set forms a vector space with *pointwise addition* and *scalar multiplication*. That is, for functions (vectors) f and g in $C^\infty([0,1],\mathbb{R})$ and a scalar λ, we define the functions (vectors) $f+g$ and λf by

$$(f+g)(t) = f(t) + g(t)$$

and

$$(\lambda \cdot f)(t) = \lambda f(t),$$

respectively. While $\mathbb{R}^n$ and P_{n-1} (from the previous example) appear to be equivalent in some way, $C^\infty([0,1],\mathbb{R})$ should feel "different". In fact, R^n and P_{n-1} are n-dimensional and $C^\infty([0,1],\mathbb{R})$ is infinite-dimensional (dimension will be defined precisely in a couple of sections), so the vector spaces are quite different. In this course we will primarily be concerned with finite-dimensional vector spaces, but it is good to keep infinite-dimensional spaces in mind (at least as motivation for the necessity of an abstract definition of *vector space*!).

2.1.2 Vector Spaces over Arbitrary Fields

We will have need to consider a slightly more general notion of a vector space, where scalars are not required to be elements of $\mathbb{R}$, but of some field $\mathbb{F}$.

Fields

A *field* is a set $\mathbb{F}$ endowed with operations $\bullet$ and $+$ called *multiplication* and *addition*, respectively, satisfying the following axioms for all $a, b, c \in \mathbb{F}$:

1. (Identities) There exists an *additive identity* denoted $0_\mathbb{F}$ such that $a + 0_\mathbb{F} = a$. There also exists a *multiplicative identity* denoted $1_\mathbb{F}$ such that $1_\mathbb{F} \bullet a = a$.

3. (Associativity) Addition and multiplication are associative:

$$(a+b)+c = a+(b+c)$$
$$(a \bullet b) \bullet c = a \bullet (b \bullet c).$$

4. (Commutativity) Addition and multiplication are commutative:

$$a+b = b+a$$
$$a \bullet b = b \bullet a.$$

5. (Inverses) Each $a \in \mathbb{F}$ has *additive inverse* denoted $-a$ such that $a + (-a) = 0_{\mathbb{F}}$. Each $a \in F$ besides $0_{\mathbb{F}}$ also has a *multiplicative inverse* denoted a^{-1} such that $a \bullet a^{-1} = 1_{\mathbb{F}}$.

6. (Distributivity) Multiplication distributes over addition:

$$a \bullet (b + c) = (a \bullet b) + (a \bullet c).$$

We have the following standard examples of fields.

Example 2.1.4. The real numbers $\mathbb{R}$ form a field with the obvious multiplication and addition operations.

Example 2.1.5. The complex numbers $\mathbb{C}$ form a field with complex multiplication and addition.

Our last basic example of a field will play a very important role in later chapters.

Example 2.1.6. Let F_2 denote the *field with two elements*. As a set, $F_2 = \{0, 1\}$. The addition and multiplication operations are described by the following tables.

$+$	0	1
0	0	1
1	1	0

$\bullet$	0	1
0	0	0
1	0	1

You will verify that F_2 is indeed a field in the exercises.

Vector Spaces over $\mathbb{F}$

We then modify the definition of a vector space over $\mathbb{R}$ to obtain the more general notion of a vector space over a field $\mathbb{F}$. To be precise, a *vector field over* $\mathbb{F}$ is a set V together with a vector addition operation and a scalar multiplication operation

$$\begin{aligned} \cdot : \mathbb{F} \times V &\to V \\ (\lambda, v) &\mapsto \lambda \cdot v \end{aligned}$$

satisfying the axioms:

1. (Additive Associativity) For any elements v_1, v_2 and v_3 in V,

$$(v_1 + v_2) + v_3 = v_1 + (v_2 + v_3).$$

2. (Scalar Multiple Associativity) For any λ_1 and λ_2 in $\mathbb{K}$ and any v in V,

$$\lambda_1 \cdot (\lambda_2 \cdot v) = (\lambda_1 \lambda_2) \cdot v.$$

3. (Additive Commutativity) For any v_1 and v_2 in V,
$$v_1 + v_2 = v_2 + v_1.$$
4. (Additive Identity) There exists an element $0_V \in V$ called the *additive identity* such that for any $v \in V$,
$$v + 0_V = v.$$
5. (Additive Inverse) For any $v \in V$, there exists an element $-v \in V$ called the *additive inverse of* v such that
$$v + (-v) = 0_V.$$
6. (Distributive Law I) For any $\lambda \in \mathbb{F}$ and v_1 and v_2 in V,
$$\lambda \cdot (v_1 + v_2) = \lambda \cdot v_1 + \lambda \cdot v_2.$$
7. (Distributive Law II) For any λ_1 and λ_2 in $\mathbb{F}$ and any $v \in V$,
$$(\lambda_1 + \lambda_2) \cdot v = \lambda_1 \cdot v + \lambda_2 \cdot v.$$
8. (Scalar Multiple Identity) For any v in V,
$$1_{\mathbb{F}} \cdot v = v.$$

Example 2.1.7. We can modify our examples of vector spaces over $\mathbb{R}$ to get examples of vector spaces over $\mathbb{F}$. For example, $\mathbb{F}^n$ is the collection of n-tuples of elements of $\mathbb{F}$. We can similarly define $P_n(\mathbb{F})$ to be the collection of degree-n polynomials with coefficients in $\mathbb{F}$. For $\mathbb{F} = \mathbb{C}$ we can likewise define an "infinite-dimensional" vector space $C^\infty([0,1], \mathbb{C})$ over $\mathbb{C}$.

2.2 Basis and Dimension

The vector space $\mathbb{R}^n$ of n-tuples of real numbers comes with a way to decompose its elements in a canonical way. Let $e_j \in \mathbb{R}^n$ denote the n-tuple with a 1 in the jth entry and zeros elsewhere; e.g., $e_1 = (1, 0, 0, \ldots, 0)$. Then any element $(x_1, \ldots, x_n) \in \mathbb{R}^n$ can be decomposed as
$$(x_1, \ldots, x_n) = x_1 \cdot e_1 + x_2 \cdot e_2 + \cdots + x_n \cdot e_n.$$
The set $\{e_1, \ldots, e_n\}$ is called a *basis* for $\mathbb{R}^n$. In this section we develop the notion of a basis for an abstract vector space over a field $\mathbb{F}$.

2.2.1 Basis of a Vector Space

Let $S \subset V$ be a subset of a vector space V over the field $\mathbb{F}$. A *linear combination* of elements of S is an expression of the form

$$\sum_{s \in S} \lambda_s s,$$

where each $\lambda_s \in \mathbb{F}$ and only finitely many of them are nonzero. The *span* of S, denoted $\mathrm{span}_{\mathbb{F}}(S)$ (or simply $\mathrm{span}(S)$ when the field $\mathbb{F}$ is clear), is the set of all linear combinations of elements of S. The set S is called a *spanning set* for V if $\mathrm{span}(S) = V$. The set is called *linearly dependent* if there exist scalars $\lambda_s \in \mathbb{F}$, not all zero, such that

$$\sum_{s \in S} \lambda_s s = 0_V.$$

If no such collection of scalars exist, the set is called *linearly independent* . A *basis* for V is a linearly independent spanning set.

We have the following fundamental theorem.

Theorem 2.2.1. *Every vector space V admits a basis. Moreover, any linearly independent set $S \subset V$ can be extended to a basis for V.*

The proof of the theorem requires the Axiom of Choice. We will skip it for now, but the interested reader is invited to read a proof sketch in Section 9.1.

2.2.2 Dimension of a Vector Space

Let $B \subset V$ be a basis for a vector space V. We define the *dimension* of V to be the number of elements in B. If B contains a finite number of elements n, we say that V is *n-dimensional* and otherwise we say that V is *infinite-dimensional*—most of the vector spaces that we will see in this course are finite-dimensional. We will use the notation $\dim(V)$ for the dimension of the vector space V. In Proposition 2.2.4 below we will show that our definition of dimension actually makes sense; that is, if we choose two different bases for V then they will always have the same number of elements. To prove it, we will need some lemmas. The first lemma tells us that if B is a basis for V, then any element of V can be written as a linear combination of elements of B in a *unique way.*

Lemma 2.2.2. *Let S be a linearly independent set. For any $v \in V$, there exists at most one way to write v as a linear combination of elements of S.*

Proof. Assume that for some $v \in V$ there exist sets of scalars λ_s and μ_s such that

$$v = \sum_s \lambda_s s = \sum_s \mu_s s.$$

This implies that

$$0_V = v - v = \sum_s \lambda_s s - \sum_s \mu_s s = \sum_s (\lambda_s - \mu_s) s.$$

Since S is linearly independent, this implies $\lambda_s - \mu_s = 0_{\mathbb{F}}$ for all s, which implies in turn that $\lambda_s = \mu_s$. We conclude that the two representations of v as a linear combination of elements of S were actually the same to begin with. $\square$

Lemma 2.2.3. *Let $B \subset V$ be a basis for V. Then for any nonzero $v \in V \backslash B$, the set $B \cup \{v\}$ is linearly dependent.*

Proof. We can express v uniquely as a linear combination

$$v = \sum_{b \in B} \lambda_b b$$

for some scalars λ_b, not all of which are zero. Then we have a linear combination

$$(-1)v + \sum_{b \in B} \lambda_b b = 0_V,$$

with not all coefficients equal to zero. It follows that $B \cup \{v\}$ is linearly dependent. $\square$

Proposition 2.2.4. *The dimension of V is independent of choice of basis.*

Proof. Let B and B' be bases for V. Our goal is to show that $|B| = |B'|$. If $|B| = |B'| = \infty$ then we are done, so let's assume by way of obtaining a contradiction (and without loss of generality) that $|B| = m$ and $|B| < |B'|$. Write $B = \{b_1, \ldots, b_m\}$ and let b_0' denote some distinguished element of B' which is not an element of B (using our assumption on the cardinalities of the sets). By Lemma 2.2.3, the set $B \cup \{b_0'\}$ is linearly dependent, so there exist scalars μ_j, $j = 1, \ldots, m$, and $\mu_{b_0'}$ such that not all of them are zero and

$$\mu_{b_0'} b_0' + \sum_j \mu_j b_j = 0_V. \tag{2.1}$$

Moreover, the linear independence of B implies that $\mu_{b_0'} \neq 0_{\mathbb{F}}$.

Since B' is a basis, there exist scalars $\lambda_{b'}^j$ such that

$$b_j = \sum_{b' \in B'} \lambda_{b'}^j b'$$

for all j. Plugging this into (2.1), we have

$$0_V = \mu_{b_0'} b_0' + \sum_j \mu_j \left(\sum_{b' \in B'} \lambda_{b'}^j \right) b'.$$

Collecting terms, this can be rewritten as

$$0_V = \mu_{b_0'} b_0' + \sum_{b' \in B'} \left(\sum_j \mu_j \lambda_{b'}^j \right) b' = \left(\mu_{b_0'} + \sum_j \mu_j \lambda_{b_0'}^j \right) b_0' + \sum_{b' \in B' \backslash \{b_0'\}} \left(\sum_j \mu_j \lambda_{b'}^j \right) b'.$$

This gives a linear combination of the elements of B' which is equal to 0_V such that not all coefficients in the combination are zero. This is a contradiction to the assumption that B' is a basis for V. □

Example 2.2.1.

1. The set $\{e_1, e_2, \ldots, e_n\} \subset \mathbb{R}^n$ defined at the beginning of this section is a basis for $\mathbb{R}^n$, and is frequently referred to as the *canonical basis* for $\mathbb{R}^n$. The dimension of $\mathbb{R}^n$ is therefore equal to n.
2. In general $\dim(\mathbb{F}^n) = n$, as we would hope.
3. The vector space $C^\infty([0,1], \mathbb{R})$ is infinite-dimensional (see the exercises).
4. The vector space of polynomials $P_n(\mathbb{R})$ defined in Example 2.1.2 has a basis given by monomials $\{1 = x^0, x = x^1, x^2, x^3, \ldots, x^n\}$. It follows that the dimension of $P_n(\mathbb{R})$ is $n+1$.

2.3 Linear Transformations

2.3.1 Abstract Linear Transformations

Linear Transformations

Let V and W be vector spaces over the same field $\mathbb{F}$. A *linear transformation* (also called a *linear map*) from V to W is a function $L : V \to W$ with the properties

1. $L(v_1 + v_2) = L(v_1) + L(v_2)$ for all $v_1, v_2 \in V$,
2. $L(\lambda v) = \lambda L(v)$ for all $v \in V$ and $\lambda \in \mathbb{F}$.

Put more simply, a linear map is just a map between vector spaces which preserves vector space structure; that is, it takes addition to addition and scalar multiplication to scalar multiplication.

Example 2.3.1. Let e_1, e_2, e_3 denote the standard basis for $\mathbb{R}^3$. Consider the linear map L defined by

$$L(e_1) = 2e_2 + e_3, \quad L(e_2) = e_1, \quad L(e_3) = 0.$$

Note that we have only defined L explicitly on 3 elements of the vector space $\mathbb{R}^3$. The linear structure of L and the fact that the e_j determine a basis for $\mathbb{R}^3$ allow us to *extend* the map to all of $\mathbb{R}^3$. Indeed, an arbitrary element $v \in \mathbb{R}^3$ can be expressed as a sum

$$v = \lambda_1 e_1 + \lambda_2 e_2 + \lambda_3 e_3$$

for some scalars λ_j. The linear structure of L allows us to evaluate $L(v)$ as

$$L(v) = \lambda_1 L(e_1) + \lambda_2 L(e_2) + \lambda_3 L(e_3) = 2\lambda_1 e_2 + \lambda_1 e_3 + \lambda_2 e_1.$$

As a concrete example, the vector $v = (1,2,3) = e_1 + 2e_2 + 3e_2$ takes the value

$$L(v) = e_2 + e_3 + 2e_1 = (2,1,1).$$

Linear Extensions

Let us expand on the observation in Example 2.3.1 that linear maps can be defined by defining their values on basis elements.

Proposition 2.3.1. *Let $e_1, \ldots, e_n$ be a fixed basis for the vector space V and let W be an arbitrary vector space. Then for any choices of images $L(e_i) \in W$, there exists a unique linear map $L : V \to W$ which takes these values on the basis.*

Proof. For any $v \in V$, there is a unique representation of v as a linear combination

$$v = \lambda_1 e_1 + \lambda_2 e_2 + \cdots + \lambda_n e_n$$

for some scalars λ_j. We then define $L(v)$ as

$$L(v) = \lambda_1 L(e_1) + \lambda_2 L(e_2) + \cdots + \lambda_n L(e_n).$$

On the other hand, if L is linear then it follows from the defining properties of a linear map that L must take this value on v, and we see that this is the unique linear map taking the prescribed values on the basis. □

The process of defining a linear map from its values on basis vectors is called *extending linearly*.

Linear Isomorphisms

A linear transformation which is a bijection is called a *linear isomorphism* . A pair of vector spaces are called *isomorphic* if there is a linear isomorphism between them. In this case we write $V \approx W$.

Proposition 2.3.2. *Let $L : V \to W$ be a linear isomorphism. The inverse function $L^{-1} : W \to V$ is a linear map.*

Proof. Let $w, w' \in W$ and $\lambda \in \mathbb{F}$. Since L is a bijection, there exist unique $v, v' \in V$ such that $L(v) = w$ and $L(v') = w'$. We can see that L^{-1} satisfies the conditions making it linear map by direct calculation:

$$L^{-1}(w + w') = L^{-1}(L(v) + L(v')) = L^{-1}(L(v + v')) = v + v' = L^{-1}(w) + L^{-1}(w')$$

and

$$L^{-1}(\lambda w) = L^{-1}(\lambda L(v)) = L^{-1}(L(\lambda v)) = \lambda v = \lambda L^{-1}(w).$$

□

We can now see that finite-dimensional vector spaces have a simple classification up to isomorphism. It requires the following simple but useful lemmas, which hold even for infinite-dimensional vector spaces.

Lemma 2.3.3. *A linear map $L : V \to W$ is injective if and only if $L(v) = 0_W$ implies $v = 0_V$.*

Proof. Any linear map $L : V \to W$ satisfies $L(0_V) = 0_W$. If L is injective, it follows that $L(v) = 0_W$ implies $v = 0_V$. On the other hand, assume that the only element of V which maps to 0_W is 0_V. Then for any $v, v' \in V$ with $v \neq v'$,

$$v \neq v' \Rightarrow v - v' \neq 0_v \Rightarrow L(v - v') \neq 0_W \Rightarrow L(v) - L(v') \neq 0_W \Rightarrow L(v) \neq L(v')$$

and it follows that L is injective. □

Lemma 2.3.4. *An injective linear map takes linearly independent sets to linearly independent sets. A surjective linear map takes spanning sets to spanning sets.*

Proof. Let $L : V \to W$ be a linear map. Assuming that L is injective, let $S \subset V$ be a linearly independent set. Because L is injective, any element w in the image of S under L can be written uniquely as $L(s)$ for some $s \in S$. A linear combination of elements of the image of S then satisfies

$$0_W = \sum_{s \in S} \lambda_s L(s) = \sum_{s \in S} L(\lambda_s s) = L(\sum_{s \in S} \lambda_s s)$$

only if $\sum \lambda_s s = 0_V$ by Lemma 2.3.3. The independence of S then implies that $\lambda_s = 0$ for all s and it follows that the image of S under L is linearly independent.

Now assume that L is surjective and let S be a spanning set. We wish to show that the image of S is spanning. For any $w \in W$, the surjectivity of L implies that there exists $v \in V$ with $L(v) = w$ (although such a v is not necessarily unique). Since S is spanning, there exist coefficients λ_s such that $v = \sum \lambda_s s$ and it follows that $w = \sum \lambda_s L(s)$. □

We have the following immediate corollary.

Corollary 2.3.5. *Let $L : V \to W$ be a linear transformation of finite-dimensional vector spaces of the same dimension. If L is injective then it is an isomorphism. Likewise, if L is surjective then it is an isomorphism.*

Proof. If L is injective, choose a basis B for V. The image $L(B)$ of this basis is linearly independent in W, and since the dimension of W is the same as the dimension of V, it follows that L is surjective as well. A similar argument works in the case that L is surjective. □

Finally, we have the following classification result for finite-dimensional vector spaces.

Theorem 2.3.6. *Let V and W be finite-dimensional vector spaces over $\mathbb{F}$. Then $V \approx W$ if and only if $\dim(V) = \dim(W)$.*

Proof. Let $\{v_1, \ldots, v_n\}$ and $\{w_1, \ldots, w_m\}$ be bases for V and W, respectively. If $V \approx W$, there exists a linear isomorphism $L : V \to W$. By Lemma 2.3.4, the injectivity of L implies that the image of the basis for V is linearly independent in W, while the surjectivity of L implies that the image of the basis for V is spanning. Therefore $n = m$ and V and W have the same dimension.

Conversely, suppose that $n = m$. We define a linear map $L : V \to W$ by defining it on the basis by $L(v_j) = w_j$ and extending. This is clearly an isomorphism. □

Example 2.3.2. It follows from Example 2.2.1 and Proposition 2.3.6 that the spaces $P_n(\mathbb{R})$ and $\mathbb{R}^{n+1}$ are isomorphic.

2.3.2 Linear Transformations of Finite-Dimensional Vector Spaces

Matrix Representations: An Example

You are probably used to writing linear maps between finite-dimensional vector spaces in terms of matrices, as in the following example.

Example 2.3.3. Consider the linear map $L : \mathbb{R}^3 \to \mathbb{R}^3$ from Example 2.3.1. Using the standard column vector notation

$$\lambda_1 e_1 + \lambda_2 e_2 + \lambda_3 e_3 \leftrightarrow \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \end{pmatrix},$$

the map can be written as matrix multiplication:

$$L(\lambda_1 e_1 + \lambda_2 e_2 + \lambda_3 e_3) = \begin{pmatrix} 0 & 1 & 0 \\ 2 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \end{pmatrix} = \begin{pmatrix} \lambda_2 \\ 2\lambda_1 \\ \lambda_1 \end{pmatrix} \leftrightarrow \lambda_2 e_1 + 2\lambda_1 e_2 + \lambda_1 e_3.$$

We will see in a moment that linear maps in finite dimensions can always be expressed as matrices, but that this representation depends on choices of bases. This choice is sometimes unnatural, so it is important to understand the abstract definition of a linear map. To further convince you, the next example gives a linear map between infinite-dimensional vector spaces, where there is no hope to represent it using a matrix.

Example 2.3.4. Consider the map $D : C^\infty([0,1], \mathbb{R}) \to C^\infty([0,1], \mathbb{R})$, where $D(f)$ is defined at each $x \in [0,1]$ by

$$D(f)(x) = f'(x).$$

You will show that this is a linear map of vector spaces in the exercises.

Matrix Representations: Formal Theory

We now turn to the matrix representation of a linear map for finite-dimensional vector spaces. Let $L : V \to W$ be a linear map. Abstractly this just means that it satisfies certain properties which mean that L respects the vector space structures of V and W. However, if we fix ordered bases $e_1, \ldots, e_n$ for V and $f_1, \ldots, f_m$ for W, we can represent L by a size $m \times n$ matrix as follows. For each e_j, we can write

$$L(e_j) = \lambda_{1j} f_1 + \lambda_{2j} f_2 + \cdots + \lambda_{mj} f_m$$

for some scalars λ_{ij}. Cycling through the n basis vectors of V, we obtain $m \cdot n$ such scalars, and our matrix representation of L (with respect to these ordered bases basis) is given by

$$\begin{pmatrix} \lambda_{11} & \lambda_{12} & \cdots & \lambda_{1n} \\ \lambda_{21} & \lambda_{22} & \cdots & \lambda_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ \lambda_{m1} & \lambda_{m2} & \cdots & \lambda_{mn} \end{pmatrix} =: (\lambda_{ij})_{ij}$$

This is called the *matrix representation* of L with respect to the chosen bases.

To do calculations in matrix representations, note that we can also express an arbitrary $v \in V$ as a matrix with respect to this basis. Namely, as the $n \times 1$ matrix $(\lambda_1, \ldots, \lambda_n)^T$ (superscript T denotes matrix transpose), where $v = \lambda_1 e_1 + \cdots + \lambda_n e_n$. To evaluate the linear map on the vector, we just perform the matrix multiplication

$$\begin{pmatrix} \lambda_{11} & \lambda_{12} & \cdots & \lambda_{1n} \\ \lambda_{21} & \lambda_{22} & \cdots & \lambda_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ \lambda_{m1} & \lambda_{m2} & \cdots & \lambda_{mn} \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_n \end{pmatrix}.$$

The resulting $m \times 1$ matrix is interpreted as a list of coefficients for $L(v)$ in the fixed basis for W.

2.3.3 Determinants

Let $L : V \to W$ be a linear transformation of finite-dimensional vector spaces of the same dimension. There is an algorithmic way to tell when L is an isomorphism from any of its matrix representation.

From the previous section, we see that once we have chosen bases for V and W, the linear map L can be represented as matrix multiplication by an $n \times n$ matrix A, where $n = \dim(V)$. Let $A = (\vec{a}_1\ \vec{a}_2 \cdots \vec{a}_n)$, where the $\vec{a}_j$ are column vectors representing the columns of A. The *determinant* is a function det from the set of $n \times n$ matrices over $\mathbb{F}$ to $\mathbb{F}$ which satisfies three properties:

1. For any scalars $\lambda_1, \lambda_2 \in \mathbb{F}$, any $j = 1, \ldots, n$ and any column vector $\vec{v}$,
$$\det(\vec{a}_1 \cdots \lambda_1\vec{a}_j + \lambda_2\vec{v} \cdots \vec{a}_n) = \lambda_1 \cdot \det(\vec{a}_1 \cdots \vec{a}_j \cdots \vec{a}_n) + \lambda_2 \cdot \det(\vec{a}_1 \cdots \vec{v} \cdots \vec{a}_n).$$
2. For every $j = 1, \ldots, n$,
$$\det(\vec{a}_1 \cdots \vec{a}_j\ \vec{a}_{j+1} \cdots \vec{a}_n) = -\det(\vec{a}_1 \cdots \vec{a}_{j+1}\ \vec{a}_j \cdots \vec{a}_n).$$
3. $\det(I_n) = 1$.

Theorem 2.3.7. *There is a unique map* det *satisfying the properties of a determinant. If A is a square matrix representing a linear map $L : V \to W$, then* $\det(A) \neq 0$ *if and only if L is an isomorphism.*

To prove the theorem, let us first examine the $n = 2$ case. Let

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix} = \left(\begin{pmatrix} a \\ c \end{pmatrix}, \begin{pmatrix} b \\ d \end{pmatrix} \right)$$

be an arbitrary 2×2 matrix with $a, b, c, d \in \mathbb{F}$. Using the Property 1 of a determinant, we have

$$\det(A) = \det\left(a\begin{pmatrix} 1 \\ 0 \end{pmatrix} + c\begin{pmatrix} 0 \\ 1 \end{pmatrix}, \begin{pmatrix} b \\ d \end{pmatrix}\right) = a \cdot \det\begin{pmatrix} 1 & b \\ 0 & d \end{pmatrix} + c \cdot \det\begin{pmatrix} 0 & b \\ 1 & d \end{pmatrix}.$$

Continuing with a similar calculation shows

$$\begin{aligned} \det(A) = a\left(b \cdot \det\begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix} + d \cdot \det\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right) \\ + c\left(b \cdot \det\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} + d \cdot \det\begin{pmatrix} 0 & 0 \\ 1 & 1 \end{pmatrix}\right). \end{aligned} \tag{2.2}$$

Property 3 of a determinant tells us that $\det(I_2) = 1$. Moreover, Property 2 says that

$$\det\begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix} = -\det\begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix},$$

since the matrix is unchanged by switching the order of the columns. This implies that

$$\det\begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix} = 0$$

and the same reasoning yields

$$\det\begin{pmatrix} 0 & 0 \\ 1 & 1 \end{pmatrix} = 0.$$

Finally, Property 2 implies

$$\det\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} = -\det(I_2) = -1.$$

Putting all of this into (2.2), we have

$$\det(A) = ad - bc.$$

This shows that in the case of $n = 2$, a determinant map exists. Moreover, the map is unique since our formula was forced by the properties of a determinant.

It is also easy to see that A is an isomorphism if and only if $\det(A) \neq 0$. Indeed, $\det(A) = ad - bc = 0$ if and only if the columns of A are linearly dependent. To see this, assume some entry of A is nonzero (otherwise we are done). Without loss of generality, say $c \neq 0$. Then the columns of A satisfy

$$\begin{pmatrix} b \\ d \end{pmatrix} = \frac{d}{c} \cdot \begin{pmatrix} a \\ c \end{pmatrix}.$$

Having linearly dependent columns is equivalent to the failure of A to be the matrix representation of an isomorphism. This completes the proof of the theorem in the $n = 2$ case.

For the general case, we build the determinant by induction. For a 3×3 matrix, we define

$$\det \begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix} = a \cdot \det \begin{pmatrix} e & f \\ h & i \end{pmatrix} + b \cdot \det \begin{pmatrix} d & f \\ g & i \end{pmatrix} + c \cdot \det \begin{pmatrix} d & e \\ g & h \end{pmatrix}$$

Inductively, for an $n \times n$ matrix A we define the j-th $(n-1) \times (n-1)$ minor A_j by deleting the first row and the j-th column of A. Then

$$\det(A) = a_{11} \cdot \det(A_1) + a_{12} \cdot \det(A_2) + \cdots + a_{1n} \cdot \det(A_n).$$

It is easy to check that this map satisfied the desired properties. To show that this map is unique is straightforward, but somewhat time consuming. A full proof can be found in any linear algebra textbook, e.g. Chapter 5 of [3]. A proof that $\det(A) \neq 0$ if and only if A is an isomorphism can be found there as well.

2.4 Vector Space Constructions

2.4.1 Subspaces

Let V be a vector space over $\mathbb{F}$. A subset $U \subset V$ is a *vector subspace* (also called a *linear subspace* or simply *subspace*) of V if it is itself a vector space with respect to operations obtained by restricting the vector space operations of V. More concretely, U is a vector subspace if and only if:

1. (Closure Under Addition) for all $u, v \in U$, $u + v \in U$
2. (Closure Under Scalar Multiplication) for all $u \in U$ and $\lambda \in \mathbb{F}$, $\lambda u \in U$.

The *dimension* of a vector subspace is just its dimension as a vector space, using the usual definition.

Example 2.4.1. For an n-dimensional vector space V, the vector subspaces of V take one of the following forms:

1. the subset containing only the zero vector $\{0_V\}$
2. spans of collections of linearly independent vectors; for $v_1, \ldots, v_m \in V$ linearly independent, the set
$$\mathrm{span}(\{v_1, \ldots, v_m\})$$
is an m-dimensional vector subspace
3. the full space V.

Example 2.4.2. As you might expect, subspaces of infinite-dimensional vector spaces can be much more exotic. For example, the set

$$\{f \in C^\infty([0,1],\mathbb{R}) \mid f(0) = 0\}$$

is a vector subspace of $C^\infty([0,1],\mathbb{R})$.

2.4.2 Special Subspaces Associated to a Linear Transformation

Kernel

Let $L : V \to W$ be a linear transformation. We define the *kernel* of V to be the set

$$\ker(L) = \{v \in V \mid L(v) = 0_W\}.$$

Proposition 2.4.1. *The kernel of a linear transformation $L : V \to W$ is a vector subspace of V.*

Proof. Let $u, v \in \ker(L)$ and λ a scalar. We need to show that $u + v$ and λu are elements of $\ker(L)$. Indeed,

$$L(u + v) = L(u) + L(v) = 0_W + 0_W = 0_W$$

and

$$L(\lambda u) = \lambda L(u) = \lambda 0_W = 0_W.$$

□

Image

The *image* of the linear map $L : V \to W$ is the set

$$\text{image}(L) = \{w \in W \mid w = L(v) \text{ for some } v \in V\}.$$

Proposition 2.4.2. *The image of a linear transformation $L : V \to W$ is a vector subspace of W.*

We leave the proof of this proposition as an exercise.

2.4.3 Rank and Nullity

Let $L : V \to W$ be a linear map of finite-dimensional vector spaces. We define the *rank* of L to be the dimension of image(L). We define the *nullity* of L to be the dimension of $\ker(L)$. These quantities are denoted rank(L) and null(L), respectively. We have the following fundamental theorem.

Theorem 2.4.3 (Rank-Nullity Theorem)**.** *For a linear map of $L : V \to W$ of finite-dimensional vector spaces,*

$$\text{rank}(L) + \text{null}(L) = \dim(V).$$

Proof. Let $\dim(V) = n$. Let $\{v_1, \ldots, v_k\}$ be a basis for $\ker(L) \subset V$, so that $\mathrm{null}(L) = k$. By Theorem 2.2.1, the set $\{v_1, \ldots, v_k\}$ can be extended to a basis $\{v_1, \ldots, v_k, w_1, \ldots, w_{n-k}\}$ for V. We claim that the set $B = \{L(w_1), \ldots, L(w_{n-k})\}$ forms a basis for $\mathrm{image}(L)$, and this will complete the proof of the theorem.

To see that B is a spanning set for $\mathrm{image}(L)$, let $w \in \mathrm{image}(L)$. Then there exists $v \in V$ with $L(v) = w$. There exist unique scalars $\lambda_1, \ldots, \lambda_k, \nu_1, \ldots, \nu_{n-k}$ such that $v = \sum \lambda_j v_j + \sum \nu_\ell w_\ell$. Because the v_j lie in the kernel of L, it follows that

$$\begin{aligned} w = L(v) &= L\left(\sum_{j=1}^{k} \lambda_j v_j + \sum_{\ell=1}^{n-k} \nu_\ell w_\ell\right) \\ &= \sum \lambda_j L(v_j) + \sum \nu_\ell L(w_\ell) \\ &= \sum \nu_\ell L(w_\ell), \end{aligned}$$

and this shows that B is spanning.

To see that B is linearly independent, suppose that

$$0_W = \sum_{\ell=1}^{n-k} \nu_\ell L(w_\ell) = L\left(\sum \nu_\ell w_\ell\right).$$

Since L is injective on $\mathrm{span}\{w_1, \ldots, w_\ell\}$, it follows from Lemma 2.3.3 that $\sum \nu_\ell w_\ell = 0_V$. Since the w_ℓ are linearly independent, this implies that all $\nu_\ell = 0$. □

2.4.4 Direct Sums

Given two vector spaces V and W over $\mathbb{F}$, we define the *direct sum* to be the vector space $V \oplus W$ with $V \oplus W = V \times W$ as a set, addition defined by

$$(v_1, w_1) + (v_2, w_2) = (v_1 + v_2, w_1 + w_2)$$

and scalar multiplication defined by

$$\lambda \cdot (v, w) = (\lambda \cdot v, \lambda \cdot w).$$

Proposition 2.4.4. *The dimension of $V \oplus W$ is $\dim(V) + \dim(W)$.*

Proof. If either V or W is infinite-dimensional, then so is $V \oplus W$. Indeed, assume without loss of generality that $\dim(V) = \infty$. Let $S \subset V$ be a linearly independent set containing infinitely many elements. Then for any $w \in W$, the set $\{(s, w) \mid s \in S\}$ is an infinite linearly independent subset of $V \oplus W$.

On the other hand, if V and W are both finite-dimensional, let $\{v_1, \ldots, v_n\}$ be a basis for V and $\{w_1, \ldots, w_m\}$ a basis for W. Then it is easy to check that

$$\{(v_j, 0)\} \cup \{(0, w_k)\}$$

gives a basis for $V \oplus W$. It follows that $\dim(V \oplus W) = n + m = \dim(V) + \dim(W)$. □

2.4.5 Quotient Spaces

Let U be a vector subspace of V. We define an equivalence relation $\sim_U$ on V by $v \sim_U w$ if and only if $v - w \in U$. As usual, we denote by $[v]$ the *equivalence class* of $v \in V$,

$$[v] = \{w \in V \mid w \sim_U v\} = \{w \in V \mid w - v \in U\}.$$

The collection of equivalence classes is called the *quotient of* V *by* U and is denoted V/U .

Proposition 2.4.5. *The quotient space V/U has a natural vector space structure.*

Proof. We define the zero vector to be

$$0_{V/U} = [0_V] = U,$$

we define vector addition by the formula

$$[u] + [v] = [u + v]$$

and we define scalar multiplication by the formula

$$\lambda[u] = [\lambda u].$$

We leave it as an exercise to show that the vector space axioms are satisfied with respect to these operations. □

Let V be finite-dimensional. Then the dimension of V/U is readily computable.

Proposition 2.4.6. *The dimension of V/U is* $\dim(V) - \dim(U)$.

Proof. Let B' be a basis for U and let B denote its completion to a basis for V (which exists by Theorem 2.2.1). We claim that

$$\{[b] \mid b \in B \backslash B'\} \tag{2.3}$$

is a basis for V/W. Indeed, this set is spanning, since any $[v] \in V/U$ can be written as

$$[v] = \left[\sum_{b \in B} \lambda_b b\right] = \sum_{b \in B} \lambda_b [b] = \sum_{b \in B \backslash B'} \lambda_b [b].$$

The existence of the coefficients λ_b comes from the fact that B is a basis for V, the first equality follows by the definition of the vector space structure of V/U, and the last equality follows because $[b] = [0]$ for any $b \in B'$. Moreover, the set (2.3) is linearly independent, as

$$[0] = \sum_{b \in B \backslash B'} \lambda_b [b] = \left[\sum_{b \in B \backslash B'} \lambda_b b\right]$$

implies that $\sum_{b \in B \setminus B'} \lambda_b b \in B'$ and this can only be the case if all $\lambda_b = 0$ by the linear independence of $B \setminus B' \subset B$.

Finally, we claim that the set (2.3) contains $|B| - |B'|$ distinct elements. Its cardinality is certainly bounded above by this number, so we need to check that if $b_1, b_2 \in B \setminus B'$ satisfy $b_1 \neq b_2$, then $[b_1] \neq [b_2]$. This holds because $[b_1] = [b_2]$ if and only if $b_1 - b_2 \in B'$, which is impossible by linear independence. □

2.4.6 Row and Column Operations

Let $L : V \to W$ be a linear map between finite-dimensional vector spaces and let A denote the matrix representation of L with respect to some fixed choices of ordered bases for V and W. We have the following *row and column operations* on A:

1. Multiply all entries in a row/column by the same nonzero element of F,
2. Permute two rows/columns,
3. Add a nonzero multiple of a row/column to another row/column.

Each row/column operation corresponds to matrix multiplication:

1. To multiply the jth row of A by nonzero $a \in \mathbb{F}$, we multiply on the left by the square matrix of size $\dim(W)$ which is diagonal with entries $a_{ii} = 1$ for all $i \neq j$ and $a_j j = a$. To multiply a column of A a constant, we multiply on the right by a similar matrix with size $\dim(V)$.
2. To permute the i and j row of A, we multiply on the left by the matrix obtained by form the identity (of dimension $\dim(W)$) by switching its ith and jth row. A similar trick, using right multiplication, can be used to permute columns of A.
3. To add a times the ith row of A to its jth row, we left multiply by the matrix which is equal to the identity except for the entry $a_{ij} = a$. Once again, a similar procedure using right multiplication gives the column operation.

Proposition 2.4.7. *Let $L : V \to W$ be a linear map between finite-dimensional vector spaces of the same dimension and let A be a matrix representation of L with respect to some choices of bases for V and W. Let $M : V \to V$ and $N : W \to W$ be matrices corresponding to row and column operations on A. Then M and N are isomorphisms.*

Proof. The matrices corresponding to row and column operations have nonzero determinant, so the claim follows by Proposition 2.3.7. □

2.4.7 Vector Space Associated to a Linear Map

Let $L : V \to W$ be a linear map of finite-dimensional vector spaces. There is a vector space associated to L, denoted $\theta(L)$ and defined as

$$\theta(L) = W/\mathrm{image}(V).$$

We refer to the map taking L to $\theta(L)$ as the *θ-correspondence* .

Let A be its matrix representation with respect to some choice of ordered bases for V and W. We have the following characterization of the circumstances where vector spaces obtained via the θ-correspondence are isomorphic.

Proposition 2.4.8. *Let $M : V \to V$ and $N : W \to W$ be linear isomorphisms. Then $\theta(NLM) \approx \theta(L)$. It follows that if a matrix A' is obtained from A by row and column operations, then the linear transformation L' associated to A' satisfies $\theta(L') \approx \theta(L)$.*

Proof. Since M and N are isomorphisms, image(NLM) has the same dimension as image(L) and it follows from Proposition 2.3.6 and Proposition 2.4.6 that $\theta(L) \approx \theta(NLM)$. The second part of the proposition follows from Proposition 2.4.7. $\square$

2.5 Structures on Vector Spaces

In this section we introduce some extra structures on vector spaces. Let V denote a vector space over $\mathbb{R}$ throughout this section.

2.5.1 Inner Products

An *inner product* on V is a map

$$\begin{aligned} \langle \cdot, \cdot \rangle : V \times V &\to \mathbb{R} \\ (v_1, v_2) &\mapsto \langle v_1, v_2 \rangle \end{aligned}$$

with the following properties for all $u, v, w \in V$ and scalars $\lambda \in \mathbb{R}$:

1. (Positive-Definititeness) $\langle v, w \rangle \geqslant 0$ and equality holds if and only if v or w is equal to 0_V,

2. (Symmetry) $\langle v, w \rangle = \langle w, v \rangle$,

3. (Bilinearity) $\langle u + v, w \rangle = \langle u, w \rangle + \langle v, w \rangle$ and $\langle \lambda v, w \rangle = \lambda \langle v, w \rangle$. It follows from the symmetry property that $\langle v, u + w \rangle = \langle v, u \rangle + \langle v, w \rangle$ and $\langle v, \lambda w \rangle = \lambda \langle v, w \rangle$.

Example 2.5.1. It is a useful exercise to verify that the standard dot product on $\mathbb{R}^n$

$$(a_1, a_2, \ldots, a_n) \bullet (b_1, b_2, \ldots, b_n) = a_1 \cdot b_1 + a_2 \cdot b_2 + \ldots + a_n \cdot b_n$$

defines an inner product.

Example 2.5.2. Notice that the dot product can be expressed in matrix form as vw^T for any $v, w \in \mathbb{R}^n$, where the superscript T denotes matrix transpose. More generally, let M be an $n \times n$ matrix with real entries. Then the map

$$(v, w) \mapsto v \cdot M \cdot w^T$$

defines an inner product on $\mathbb{R}^n$ provided:

1. M is *symmetric*; i.e., $M^T = M$,
2. M is *positive-definite*; i.e. $v \cdot M \cdot v^T > 0$ for all $v \neq \vec{0}$.

This gives a large collection of examples of inner products on $\mathbb{R}^n$ which can be easily generalized to any finite-dimensional vector space over $\mathbb{R}$.

Example 2.5.3. As a more exotic example, consider the vector space $C^\infty([0,1], \mathbb{R})$ with the map $\langle \cdot, \cdot \rangle_{L^2}$ defined for functions f and g by

$$\langle f, g \rangle_{L^2} = \int_0^1 f(t) \cdot g(t) \, \mathrm{d}t.$$

You will show in the exercises that this map defines an inner product.

A pair $(V, \langle \cdot, \cdot \rangle)$ consisting of a vector space together with a choice of inner product is called an *inner product space.*

Remark 2.5.1. *One can similarly define an inner product on a vector space over $\mathbb{C}$ with a slight change to the axioms. In this case, the definition is meant to be a generalization of the map on $\mathbb{C}^n \times \mathbb{C}^n$ defined by*

$$((z_1, z_2, \ldots, z_n), (w_1, w_2, \ldots, w_n)) \mapsto z_1 \cdot \overline{w_1} + z_2 \cdot \overline{w_2} + \ldots z_n \cdot \overline{w_n},$$

where $\overline{w}$ denotes the complex conjugate of w. Can you guess what needs to be changed in the definition of an inner product in this case?

2.5.2 Norms

A *norm* on V is a map

$$\begin{aligned} \| \cdot \| : V &\to \mathbb{R} \\ v &\mapsto \|v\| \end{aligned}$$

with the following properties for all $u, v \in V$ and scalars $\lambda \in R$:

1. (Positive-Definiteness) $\|v\| \geqslant 0$ and equality holds if and only if $v = 0_V$,
2. (Linearity Over Scalar Multiplication) $\|\lambda v\| = |\lambda| \cdot \|v\|$,
3. (Triangle Inequality) $\|u + v\| \leqslant \|u\| + \|v\|$.

There is one immediate source of norms on V.

Proposition 2.5.2. *Any inner product $\langle \cdot, \cdot \rangle$ on V determines a norm on V.*

To prove the proposition, we need to make use of a famous lemma.

Lemma 2.5.3 (Cauchy-Schwarz Inequality)**.** *For any inner product $\langle \cdot, \cdot \rangle$ and any $u, v \in V$,*

$$| \langle u, v \rangle | \leqslant \langle u, u \rangle \langle v, v \rangle .$$

Proof. If $v = 0_V$ we are done, so suppose not. Define $\lambda = \langle u, v\rangle / \langle v, v\rangle^2$. Then positive-definiteness and bilinearity of the inner product implies

$$\begin{aligned} 0 &\leqslant \langle u - \lambda v, u - \lambda v\rangle \\ &= \langle u, u\rangle^2 + \lambda^2 \langle v, v\rangle^2 - 2\lambda \langle u, v\rangle \\ &= \frac{\langle u, u\rangle^2 \langle v, v\rangle^2}{\langle v, v\rangle^2} + \frac{\langle u, v\rangle^2}{\langle v, v\rangle^2} - 2\frac{\langle u, v\rangle^2}{\langle v, v\rangle^2}. \end{aligned}$$

Rearranging the terms and taking a square root proves the claim. □

We can now prove the proposition.

Proof. We define a candidate for a norm $\|\cdot\|$ on V by the formula

$$\|v\| = \sqrt{\langle v, v\rangle}.$$

We need to check that this definition satisfies the definition of a norm. Positive-definiteness and linearity over scalar multiplication follow immediately from the corresponding properties of $\langle \cdot, \cdot\rangle$. It remains to check the triangle inequality.

Let $u, v \in V$. Then the bilinearity of the inner product and the Cauchy-Schwarz Inequality imply

$$\begin{aligned} \|u + v\|^2 &= \langle u + v, u + v\rangle \\ &= \langle u, u\rangle^2 + \langle v, v\rangle^2 + 2\langle u, v\rangle \\ &\leqslant \|u\|^2 + \|v\|^2 + 2\|u\|\|v\| \\ &= (\|u\| + \|v\|)^2, \end{aligned}$$

and taking square roots proves the result. □

Example 2.5.4. An important family of examples of norms on $\mathbb{R}^n$ are the ℓ_p-norms , defined as follows. For each $1 \leqslant p < \infty$, define the norm $\|\cdot\|_p$ on $v = (a_1, \ldots, a_n) \in \mathbb{R}^n$ by the formula

$$\|v\|_p = (|v_1|^p + \cdots + |v_n|^p)^{1/p}.$$

For $p = \infty$, define

$$\|v\|_\infty = \max_i |v_i|.$$

Clearly, $\|v\|_2$ is the standard norm on $\mathbb{R}^n$, which can be written (as in the proposition) in the form

$$\|v\|_2 = \langle v, v\rangle.$$

Perhaps surprisingly, is a fact that none of the other ℓ_p norms are induced by inner products!

A pair $(V, \|\cdot\|)$ consisting of a vector space together with a choice of norm is called a *normed vector space*.

2.6 Exercises

1. Show that in any vector space V over $\mathbb{F}$, $0_{\mathbb{F}}v = 0_V$ for any $v \in V$.

2. Show that for the vector space $\mathbb{R}^n$, the additive inverse of $v \in \mathbb{R}^n$ is given by

$$-v = -1 \cdot v.$$

3. Show that $P_n(\mathbb{R})$ is a vector space over $\mathbb{R}$. (See Example 2.1.2.)

4. Show that $C^\infty([0,1],\mathbb{R})$ is a vector space over $\mathbb{R}$ (See Example 2.1.3.)

5. Show that F_2 is a field. (See Example 2.1.6.)

6. Show that $C^\infty([0,1],\mathbb{C})$ is a vector space over $\mathbb{C}$. (See Example 2.1.7.)

7. Show that the derivative map defined in Exercise 2.3.4 is a linear transformation between vector spaces.

8. Let $C^0([0,1],\mathbb{R})$ denote the set of continuous functions $f : [0,1] \to \mathbb{R}$. Show that $C^0([0,1],\mathbb{R})$ is a vector space over $\mathbb{R}$ and then show that $C^\infty([0,1],\mathbb{R})$ is a vector subspace of $C^0([0,1],\mathbb{R})$.

9. Prove that $C^\infty([0,1],\mathbb{R})$ is infinite-dimensional. Hint: Find a countably infinite collection of functions which you can prove are linearly independent.

10. Prove Proposition 2.4.2.

11. Complete the proof of Proposition 2.4.5.

12. Show that the $\langle \cdot, \cdot \rangle_{L^2}$ defines an inner product on $C^\infty([0,1],\mathbb{R})$ (see Example 2.5.3).

3 Metric Space Topology

In applications, we often think of data as some collection of datapoints; i.e., the data forms a set. Realistically, the set of datapoints usually comes with the extra structure of a notion of distance between the points. For example, if each datapoint is a vector of (real) numbers, then we can think of the data set as a collection of points in a vector space. There is a natural notion of distance between $v, w \in \mathbb{R}^n$ given by $\|v - w\|$, where $\| \cdot \|$ is any choice of norm on $\mathbb{R}^n$.

It is easy to imagine that the situation of the previous example can generalized to more exotic structures on the dataset. Perhaps the datapoints actually all lie on or near a sphere (or more complicated surface) inside of $\mathbb{R}^n$. Perhaps the points are more naturally represented as the nodes of some graph.

The correct abstract version of this idea is to represent the dataset as a metric space. A metric space is simply a set X together with a choice of distance function d on X. The distance function is an abstract function $d : X \times X \to \mathbb{R}$ which satisfies some natural axioms (see the following section).

We will see in this chapter that the simple idea of treating sets with distance functions abstractly produces a very rich theory. The study of metric spaces is a subfield of *topology*. We will also introduce some basic ideas from topology in this section. For more in depth coverage of metric spaces and more general topological spaces, a standard reference is Munkres' textbook [6].

3.1 Metric Spaces

3.1.1 Definition of a Metric Space

Let X be a set. A *metric* (or *distance function*) on X is a map

$$d : X \times X \to \mathbb{R}$$

satisfying the following properties for all elements x, y and z of X:

1. (Positive Definite) $d(x, y) \geqslant 0$ and $d(x, y) = 0$ if and only if $x = y$,
2. (Symmetry) $d(x, y) = d(y, x)$,
3. (Triangle Inequality) for any elements x, y and z of the set X,

$$d(x, z) \leqslant d(x, y) + d(y, z).$$

A set together with a choice of metric (X, d) is called a *metric space*.

A *metric subspace* of a metric space (X, d) is a metric space (Y, d_Y), where $Y \subset X$ and $d_Y = d|_{Y\times Y}$; that is d_Y is obtained by restricting the function d to $Y \times Y \subset X \times X$. We will frequently abuse notation and continue to denote the restricted metric by d.

3.1.2 Examples of Metric Spaces

Basic Examples

Example 3.1.1. As a very basic example, consider the set of real numbers $\mathbb{R}$ together with the metric

$$d(x, y) = |x - y|.$$

This is called the *standard metric on* $\mathbb{R}$.

Example 3.1.2. More generally, for any normed vector space $(V, \|\cdot\|)$ we define a metric by the formula

$$d(v, w) = \|v - w\|.$$

It follows immediately from the definition of a norm that this function satisfies the properties required for it to be a metric.

Example 3.1.3. Let X be any nonempty set. Define a metric d on X by setting

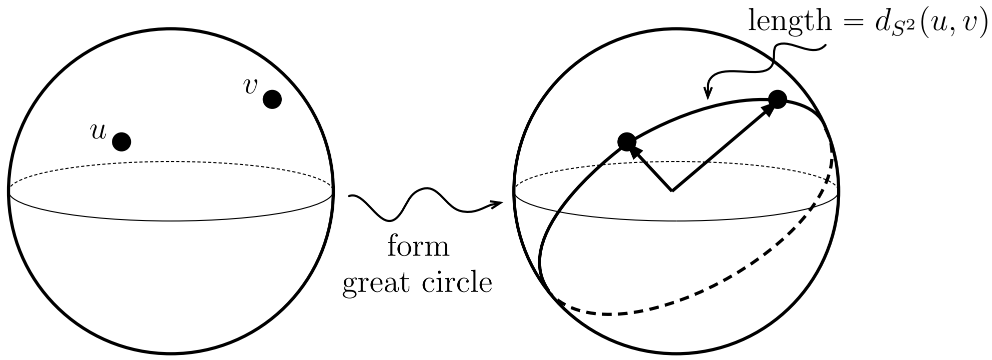
$$d(x, y) = \begin{cases} 1 & \text{if } x \neq y \\ 0 & \text{if } x = y. \end{cases}$$

Let's check that this formula really defines a metric. It is clear that d is positive-definite and symmetric, so we just need to show that it satisfies the triangle inequality. Let $x, y, z \in X$. There are five cases to consider:

1. $x = y = z$: Then $d(x, z) = 0 \leqslant 0 + 0 = d(x, y) + d(y, z)$.
2. $x = y$, $y \neq z$: Then $d(x, z) = 1 \leqslant 0 + 1 = d(x, y) + d(y, z)$.
3. $x \neq y$, $y = z$: Then $d(x, z) = 1 \leqslant 1 + 0 = d(x, y) + d(y, z)$.
4. $x \neq y$, $y \neq z$, $z = x$: Then $d(x, z) = 0 \leqslant 1 + 1 = d(x, y) + d(y, z)$.
5. $x \neq y$, $y \neq z$, $z \neq x$: Then $d(x, z) = 1 \leqslant 1 + 1 = d(x, y) + d(y, z)$.

This shows that the triangle inequality is always satisfied. The reader should check that there are no other possibilities to consider! This metric is called the *discrete metric* on X and it is useful to keep in mind when you are trying to think of counterexamples.

Example 3.1.4. Consider the standard unit sphere $S^2 \subset \mathbb{R}^3$. We define a metric on S^2 as follows. Let $u, v \in S^2$ (i.e., u and v are unit vectors in $\mathbb{R}^3$). The intersection of the plane spanned by u and v with the sphere S^2 is called the *great circle* associated to u and v. There are two segments along the great circle joining u to v. We define the metric d_{S^2} by taking $d_{S^2}(u, v)$ to be the length of the shorter of these two segments. A similar construction works for spheres of all dimensions $S^{n-1} \subset \mathbb{R}^n$. You will show that d_{S^2} is really a metric in the exercises.

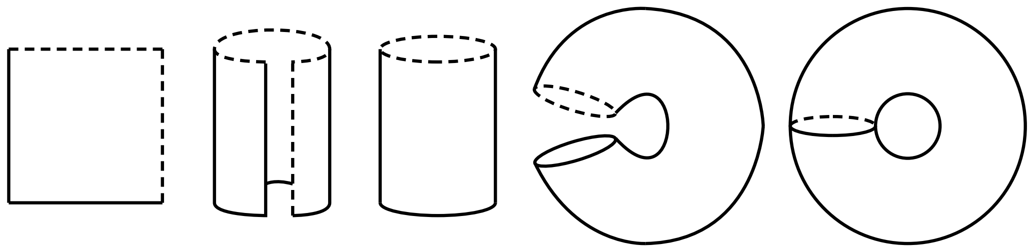


Example 3.1.5. Consider the square $T := [0, 1) \times [0, 1) \subset \mathbb{R}^2$. We define the distance d_T between points $p_1 = (x_1, y_1)$ and $p_2 = (x_2, y_2)$ in T by the formula

$$d_T(p_1, p_2) = \min_{k,\ell \in \mathbb{Z}} \|(x_1, y_1) - (x_2 + k, y_2 + \ell)\|,$$

where the norm is the standard one on $\mathbb{R}^2$. This is called the *torus metric* on T and a similar construction works for cubes of all dimensions $[0, 1]^n \subset \mathbb{R}^n$.

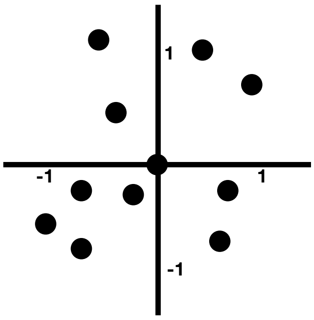
A *torus* is the geometric shape formed by the surface of a donut. An explanation of this name for d_T is given by the figure below. In the figure we form a donut shape by identifying edges of the square which have "distance zero".

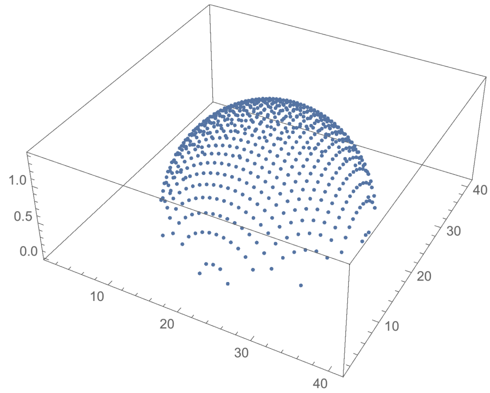


You will show that d_T is really a metric in the exercises.

Common Examples Arising in Data Analysis

Example 3.1.6. A *point cloud* in a metric space (X, d) is a metric subspace $(Y, d|_{Y \times Y})$, where Y is some finite set. The figure below shows some examples of point clouds. The figure on the left shows a simple point cloud in $\mathbb{R}^2$. The figure on the fight shows a more complicated point cloud which appears to lie along the surface of a sphere. The point clouds that we are interested in—those coming from real-world data—typically have a large number of points and exhibit some underlying structure. The tools that we develop will help us to discern this structure!



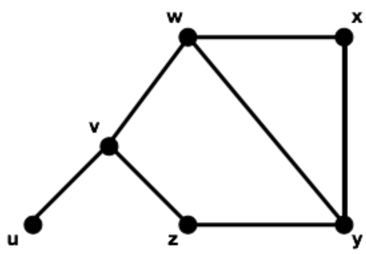


Real-world data is often naturally represented as a point cloud in some metric space. For example, consider customer records for a movie streaming service. Say the service offers streaming for n titles (n some large integer). Then, as a vastly simplified model, the record of a single customer could consist of a sequence of 0's (for movies that have not been watched) and 1's (for movies that have been watched). This record can then be represented a vector in $\mathbb{R}^n$. For two customer records v and w in $\mathbb{R}^n$, the number $\|v - w\|$ (i.e. the distance between v and w in $\mathbb{R}^n$) represents the similarity in viewing patterns between the two customers. The collection of all customer records therefore forms a point cloud in $\mathbb{R}^n$.

Of course, the customer records of any streaming service are much more detailed and include information such as when titles were viewed and what ratings the customer assigned. Thus the vectors of information can live in a space with much higher dimension and can contain numbers besides 0's and 1's. When comparing the viewing patterns between two customers, different types of information should potentially be weighted differently. This can be interpreted as the statement that the vector space containing the pointcloud should be endowed with a more complicated metric!

Example 3.1.7. Data frequently has the structure of a graph. A *graph* $G = (V, E)$ consists of a set of points V called *vertices* and a set E of *edges* $e = \{v, w\}$, where $v, w \in V$. Graphs are realized geometrically by drawing the vertex set and joining vertices v and w by a line segment when $\{v, w\} \in E$. The figure below shows a realization of the graph

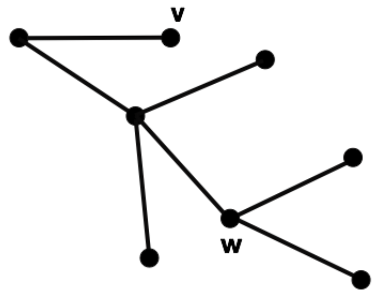
$$G = (V = \{u, v, w, x, y, z\}, E = \{\{u, v\}, \{v, w\}, \{v, z\}, \{w, x\}, \{w, y\}, \{x, y\}, \{y, z\}\}).$$



Graphs are a convenient representation of data which describes relationships between points; for example, one could take as a vertex set the members of a social media platform with a connection between vertices when the corresponding members are "friends".

A graph G defines a metric space as follows. We define the *graph distance* between vertices v, w, $d_G(v, w)$, to be the length of the shortest path joining v to w in G. Here *length* means number of edges along the path. In the above example, $d_G(v, y) = 2$, because we could take either path $\{v, w\}, \{w, y\}$ or $\{v, z\}, \{z, y\}$ to join the vertices and there is no shorter path.

A graph in which there is a unique path joining any two points is called a *tree* . Consider the tree T pictured below. For the particular v and w marked in the figure, $d_T(v, w) = 3$.



3.1.3 Open and Closed Sets

An *open metric ball* in a metric space (X, d) is defined for a *center point* $x \in X$ and a *radius* $r > 0$ to be the set

$$B(x, r) = \{y \in X \mid d(x, y) < r\}.$$

The notation will sometimes be decorated. For example, we may use $B_d(x, r)$ to emphasize the choice of metric or $B_X(x, r)$ to emphasize the set when the choice of metric is clear from context.

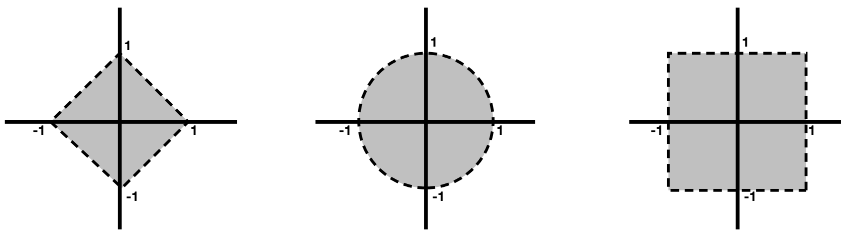
Example 3.1.8. For $\mathbb{R}$ with its standard metric, the open metric balls are open intervals.

Example 3.1.9. For a set X with the discrete metric, the open metric balls are of the form

$$B(x, r) = \begin{cases} x & \text{if } r < 1 \\ X & \text{if } r \geqslant 1. \end{cases}$$

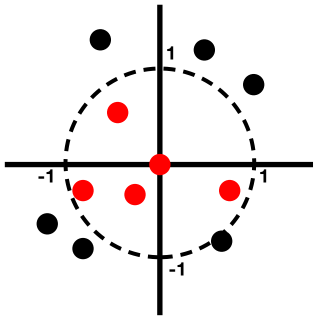
for all $x \in X$.

Example 3.1.10. In the figure below we show metric balls in $\mathbb{R}^2$ of radius 1 with the metrics induced by the ℓ_1, ℓ_2 and ℓ_∞ norms respectively.

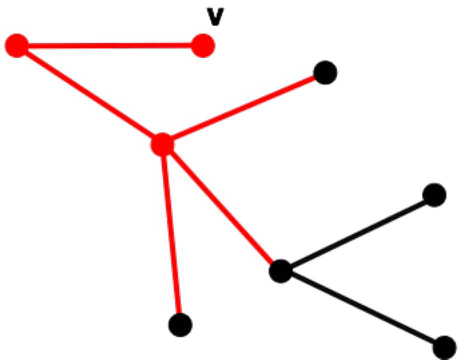


$B_{\ell_1}(\vec{0}, 1)$ $B_{\ell_2}(\vec{0}, 1)$ $B_{\ell_\infty}(\vec{0}, 1)$

Example 3.1.11. The next figure shows a point cloud of 11 points in $\mathbb{R}^2$. The 5 red points comprise the radius-1 open metric ball centered at the point $\vec{0}$ with respect to the subspace metric induced by the ℓ_2 norm.



Example 3.1.12. The next example returns to the tree metric space T of Example 3.1.7. The figure shows the open metric ball $B(v, 3)$ highlighted in red. Notice that the far endpoints are not included in the ball, since the definition $B(v, 3) = \{v \in T \mid d(v, w) < 3\}$ uses a strict inequality.



The following proposition characterizes the open metric balls of a metric subspace. The proof is left as an exercise.

Proposition 3.1.1. *Let (X, d) be a metric space and let $Y \subset X$ be endowed with the subspace metric. Then the metric open balls $B_Y(y, r)$ are of the form*

$$B_Y(y, r) = B_X(y, r) \cap Y,$$

where $B_X(y, r)$ is the metric open ball in X.

A subset $U \subset X$ of a metric space is called *open* if for all $x \in U$ there exists $r > 0$ such that $B(x, r) \subset U$. A subset $C \subset X$ is called *closed* if it can be expressed as the complement of an open set; that is,

$$C = X \backslash U = \{x \in X \mid x \notin U\}$$

for some open subset of X.

Proposition 3.1.2. *Open sets have the following properties:*

1. *X and $\varnothing$ are open;*
2. *for any collection $\mathcal{U}$ of open sets, the set*

$$\bigcup_{U \in \mathcal{U}} U$$

   is also open;
3. *for any finite collection $\mathcal{U} = \{U_1, \dots, U_n\}$ of open sets, the set*

$$\bigcap_{i=1}^{n} U_i$$

   is open.

Proof. For the first point, note that for any $x \in X$, any $r > 0$ satisfies $B(x, r) \subset X$. Thus X is open. Moreover, the statement that $\varnothing$ is open is vacuously true. To show that arbitrary unions of open sets are open, let $x \in \bigcup_{U \in \mathcal{U}} U$. Then $x \in U$ for some element of the collection, so there is $r > 0$ such that $B(x, r) \subset U \subset \bigcup_{U \in \mathcal{U}} U$. Finally, for finite intersections, let $x \in \bigcap_{i=1}^{n} U_i$. Then for each $i = 1, \dots, n$ there exists $r_i > 0$ such that $B(x, r_i) \subset U_i$. Let r be the minimum of the r_i's. Then $B(x, r) \subset \bigcap U_i$, and this completes the proof. □

Example 3.1.13. Note that an arbitrary union of open sets is open, while the corresponding statement for intersections only concerns finite collections. Indeed, it is easy to find infinite collections of open sets whose intersection is not open. For example, let $\mathcal{U} = \{U_n\}_{n=1,2,3,\dots}$ be a collection of open subsets of $\mathbb{R}$, where

$$U_n = (-1/n, 1/n)$$

is an open interval. Each set is open, but the intersection

$$\bigcap_{n=1}^{\infty} U_n = \{0\}$$

is not.

We have a similar proposition for closed sets, whose proof we leave as an exercise.

Proposition 3.1.3. *Closed sets have the following properties:*

1. *X and $\varnothing$ are closed;*
2. *for any collection $\mathcal{C}$ of closed sets, the set*

$$\bigcap_{C \in \mathcal{C}} C$$

   is also open;

3. *for any finite collection* $\mathcal{C} = \{C_1, \ldots, C_n\}$ *of closed sets, the set*

$$\bigcup_{i=1}^{n} C_i$$

is closed.

We remark that a subset Y of a metric space (X, d), can be open, closed, both open *and* closed, or neither open *nor* closed.

Example 3.1.14. Consider $\mathbb{R}$ with its standard metric induced by $|\cdot|$. The following are examples of open sets:

- $\mathbb{R}$
- (a, b) for any $a < b$
- $(0, 1) \cup (2, 3)$
- $(0, 1) \cup (2, 3) \cup (4, 5) \cup \cdots = \bigcup_{i=0}^{\infty}(2i, 2i + 1)$.

The following are examples of closed sets:

- $\mathbb{R}$
- $[a, b]$ for any $a < b$
- $\{1\}$
- $[0, 1] \cup [2, 3] \cup [4, 5] \cup \cdots = \bigcup_{i=0}^{\infty}[2i, 2i + 1]$. Note that the union of infinitely many closed sets is *not* necessarily closed; this example is just a special case.

The whole real line $\mathbb{R}$ and the empty set $\varnothing$ are examples of sets which are both open and closed. We will see in Section 3.3.2 below that these are the *only* subsets of $\mathbb{R}$ with this property. The following are examples of sets which are neither open nor closed:

- $[0, 1] \cup (2, 3)$
- $[0, 1)$.

It will be useful to characterize the open sets of a metric subspace.

Proposition 3.1.4. *Let* (X, d) *be a metric space and let* $Y \subset X$ *be endowed with the subspace metric. The open subsets of* Y *are of the form* $U \cap Y$*, where* U *is an open subset of* X*.*

Proof. Let U be an open set in X and let $y \in U \cap Y$. Then there exists an open metric ball $B_X(y, r)$ which is contained in U, and it follows that the metric open ball $B_Y(y, r) = B_X(y, r)$ is contained in $U \cap Y$. This shows that $U \cap Y$ is an open subset of Y.

Now we wish to show that *every* open subset of Y is of the form $U \cap Y$. Let V be an open subset of Y. For each $y \in V$ there exists $r(y) > 0$ such that $B_Y(y, r(y)) \subset V$. Now consider the set

$$U = \bigcup_{y \in V} B_X(y, r(y)).$$

This set is open in X (since it is the union of open sets) and has the property that $V = Y \cap U$. $\square$

3.1.4 Topological Spaces

While we are primarily concerned with metric spaces, it will occasionally be useful to use more general terminology. For some of the ideas about a metric space (X, d) that we will introduce, the metric d is auxillary, and we are really interested in the open sets of (X, d) (as defined in the last section). Based on the properties of open sets that we just derived, we make the following definition: a *topological space* is a set X together with a collection $\mathcal{T}$ of subsets of X satisfying the following axioms:

1. X and $\varnothing$ are in $\mathcal{T}$,
2. for any collection $\mathcal{U} \subset \mathcal{T}$ of elements of $\mathcal{T}$, the set
$$\bigcup_{U \in \mathcal{U}} U$$
is also in $\mathcal{T}$;
3. for any finite collection $\{U_1, \ldots, U_n\}$ of elements of $\mathcal{T}$, the set
$$\bigcap_{i=1}^{n} U_i$$
is in $\mathcal{T}$.

The collection $\mathcal{T}$ is called a *topology on* X and elements of $\mathcal{T}$ are called *open sets*.

Example 3.1.15. A metric space (X, d) is an example of a topological space. The topology $\mathcal{T}$ consists of the open subsets with respect to the metric, as we defined in the previous section. This topology is called the *metric topology* on (X, d).

All of the topological spaces that we will study will be metric spaces. However, many of the concepts that we will cover can be applied to arbitrary topological spaces; that is, they are defined in terms of topologies and the metric is of secondary importance. We refer the reader interested in studying general topological spaces to the excellent textbook [6].

The notion of a topological space is strictly more general than that of a metric space; that is, there exist topological spaces whose topologies are not induced by a metric. A simple example of such a space is given below.

Example 3.1.16. Let $X = \{a, b, c\}$. Define a topology on X to be the collection of sets $\mathcal{T} = \{\varnothing, X, \{a, b\}, \{b, c\}, \{b\}\}$. Note that there are no open sets U and V such that $a \in U$, $c \in V$ and $U \cap V = \varnothing$; in standard terminology, X is not a *Hausdorff space*. On the other hand, any metric space (Y, d) has the Hausdorff property: for any $x, y \in Y$ with $x \neq y$, the open sets $U = B_d(x, \epsilon/2)$ and $V = B_d(y, \epsilon/2)$, where $\epsilon = d(x, y)$, have the properties that $x \in U$, $y \in V$ and $U \cap V = \varnothing$ (i.e., any metric space is Hausdorff).

3.1.5 Limit Points

Let $Y \subset X$ be a subset of a metric space. The *interior* of Y is the set of points $y \in Y$ such that there exists $r > 0$ with $B(y, r) \subset Y$. The interior of Y is denoted $\mathrm{int}(Y)$.

A *limit point* of Y is a point $x \in X$ such that any open metric ball $B(x, r)$ intersects Y in some point besides x. In set notation, this condition is written

$$(B(x, r) \cap Y) \backslash \{x\} \neq \varnothing.$$

Proposition 3.1.5. *A subset $Y \subset X$ is closed if and only if it contains all of its limit points.*

Proof. First assume that Y is closed. Then $Y = X \backslash U$ for some open set $U \subset X$. For any $x \in U$ there exists $r > 0$ such that $B(x, r) \cap Y = \varnothing$ and this implies that x is not a limit point of Y. Therefore Y must contain all of its limit points.

Now assume that Y contains all of its limit points. We claim that $X \backslash Y$ is open, whence it follows that Y is closed. If $X \backslash Y = \varnothing$ we are done, so assume not and let $x \in X \backslash Y$. Then there exists $r > 0$ such that $B(x, r) \cap Y = \varnothing$; i.e., $B(x, r) \subset X \backslash Y$. Thus $X \backslash Y$ is open. $\square$

The *closure* of a subset Y is the set Y together with all limit points of Y and is denoted $\overline{Y}$. By the previous proposition, $\overline{Y}$ is a closed set. Moreover, $\overline{Y}$ is the "smallest" closed set containing Y in sense which is made precise by the following proposition.

Proposition 3.1.6. *The closure of a subset $Y \subset X$ can be characterized as*

$$\overline{Y} = \bigcap \{C \subset X \mid Y \subset C \text{ and } C \text{ is closed}\}.$$

Proof. To save space with notation, let

$$Z = \bigcap \{C \subset X \mid Y \subset C \text{ and } C \text{ is closed}\}.$$

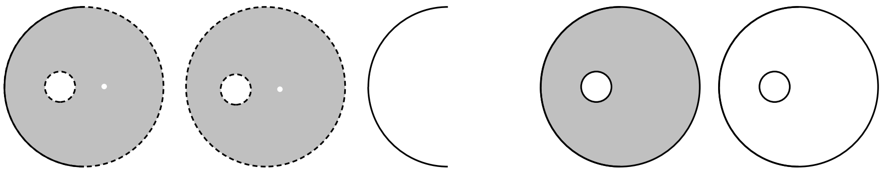
First note that $\overline{Y}$ is a closed set which contains Y, so it must be that $Z \subset \overline{Y}$. It remains to show that $\overline{Y} \subset Z$. Let $y \in \overline{Y}$. If $y \in Y$, then y is an element of each set in the intersection defining Z, so it is an element of Z and we are done. Assume that y is a limit point of Y such that $y \notin Y$ and let C be a closed set with $Y \subset C$. Since y is a limit point of Y it must also be a limit point of C and it follows from Proposition 3.1.5 that $y \in C$. Since C was arbitrary, it must be that $y \in Z$. $\square$

The *boundary* of a set $Y \subset X$ is the set

$$\partial Y := \overline{Y} \cap \overline{(X \backslash Y)}.$$

In the next example we give an example to demonstrate the intuitive meaning of the interior, boundary and closure of a given set. You will work out a similar example in the exercises.

Example 3.1.17. The figure below shows a set $Y \subset \mathbb{R}^2$ consisting of a disk with part of its boundary circle included, a closed disk removed and a point removed. The other figures show its interior $\operatorname{int}(Y)$, boundary ∂Y, closure $\overline{Y}$ and the boundary of its closure $\partial \overline{Y}$.



Y $\qquad$ $\operatorname{int}(Y)$ $\qquad$ ∂Y $\qquad$ $\overline{Y}$ $\qquad$ $\partial \overline{Y}$

3.2 Continuous Maps

The notion of a continuous map between metric spaces is of fundamental importance. Accordingly, it has several equivalent definitions which are useful in different contexts. The next proposition gives two of them. For a function $f : X \to Y$ between sets and a subset $Z \subset Y$, we use

$$f^{-1}(Z) = \{x \in X \mid f(x) \in Z\}$$

to denote the *preimage set of* Z.

Proposition 3.2.1. *Let* (X, d_X) *and* (Y, d_Y) *be metric spaces and let* $f : X \to Y$ *be a function. The following are equivalent:*

1. *for any open set* $U \subset Y$*, the preimage set* $f^{-1}(U)$ *is open in* X*;*
2. *for any* $\epsilon > 0$ *and any* $x \in X$*, there exists* $\delta > 0$ *such that* $d_Y(f(x), f(x')) < \epsilon$ *whenever* $d_X(x, x') < \delta$*.*

Proof. Assume that the first property holds and let $\epsilon > 0$ and $x \in X$. Consider the open metric ball $B_Y(f(x), \epsilon)$. By our assumption, the preimage set $U = f^{-1}(B_Y(f(x), \epsilon))$ is open. It certainly contains x, and by definition this means that there exists $\delta > 0$ such that $B_X(x, \delta) \subset U$. Then whenever $d_X(x, x') < \delta$, we have $x' \in B_X(x, \delta)$ which implies $x' \in U$ and this in turn implies that $d_Y(f(x), f(x')) < \epsilon$.

We now turn to the reverse implication. Assume that the second property holds and let $U \subset Y$ be an open set. We wish to show that $f^{-1}(U)$ is open. Assuming that the preimage set is nonempty (otherwise we are done), let $x \in f^{-1}(U)$. Then $f(x) \in U$ and since U is open this implies that there is an $\epsilon > 0$ such that $B_Y(f(x), \epsilon) \subset U$. By our assumption, we can choose $\delta > 0$ such that $B_X(x, \delta) \subset f^{-1}(B_Y(f(x), \epsilon) \subset f^{-1}(U)$, and this implies that $f^{-1}(U)$ is open. $\square$

If a function f satisfies the properties of the previous proposition, we say that it is *continuous*.

Example 3.2.1. Consider the metric space $(\mathbb{R}, |\cdot|)$; that is, use the metric on $\mathbb{R}$ induced by absolute value. Let $f : \mathbb{R} \to \mathbb{R}$ be a function. Then the second definition of continuity reads: f is continuous if for all $\epsilon > 0$ and all $x \in \mathbb{R}$, there exists $\delta > 0$ such that $|f(x) - f(x')| < \epsilon$ whenever $|x - x'| < \delta$. This is the usual definition of continuous that you have used since Calculus I! This means that all of the elementary functions (polynomials, exponentials, trigonometric functions with appropriately restricted domains) are continuous in the metric space sense.

The following lemma will be useful and we leave its proof as an exercise.

Lemma 3.2.2. *Let $f : X \to Y$ and $g : Y \to Z$ be continuous maps of metric spaces. Then $g \circ f : X \to Z$ is continuous as well.*

3.3 Topological Properties

Continuous maps are extremely important in the study of metric spaces, as they preserve the "large scale" metric structures of metric spaces. More generally, they preserve the open set structure of the metric spaces; that is, they preserve the metric topologies. Because of this, properties which are preserved by continuous maps are called *topological properties*. In the next two subsections we will introduce the two most basic topological properties.

3.3.1 Compactness

An *open cover* of a metric space (X, d) is a collection $\mathcal{U}$ of open sets such that $\cup_{U \in \mathcal{U}} U = X$. A *subcover* is a subset $\mathcal{U}' \subset \mathcal{U}$ which is still an open cover of X. The space is said to be *compact* if every open cover admits a finite subcover. We call a subset $Y \subset X$ *compact* if it is compact as a metric space with its subspace metric.

Example 3.3.1. A basic example of a compact space is a finite set of points $Y = \{y_1, \ldots, y_n\}$ in a metric space (X, d). For any open cover $\mathcal{U}$ of Y, there exists an open set U_j such that $x_j \in U_j$ for all j (this must be the case, since $\mathcal{U}$ covers Y!). Then $\{U_1, \ldots, U_n\}$ is a finite subcover of $\mathcal{U}$.

Example 3.3.2. The space $\mathbb{R}$ with its standard topology is not compact. To see this, consider the open cover $\mathcal{U} = \cup_{k \in Z}(-k, k)$. Any finite subcollection of elements of $\mathcal{U}$ is of the form $\{(-k_1, k_1), (-k_2, k_2), \ldots, (-k_n, k_n)\}$ for some positive integer n. Let k_M denote the maximum k_j. Then $\cup_j(-k_j, k_j) \subset (-k_M, k_M)$, and the point $k_M + 1 \in \mathbb{R}$ is not contained in the subcollection. Therefore the open cover does not admit a finite subcover.

Example 3.3.3. The subspace $(0, 1) \subset \mathbb{R}$ is not compact. To see this, consider the open cover

$$\mathcal{U} = \{(1/k, 1) \mid k \in \mathbb{Z} \text{ and } k > 0\}.$$

This is an open cover, since for any $x \in (0,1)$, there exists a positive integer k such that $1/k < x$. By an argument similar to the last example, any finite subcollection of elements of $\mathcal{U}$ will have its union contained in an interval $(1/k_M, 1)$. Then $1/(k_M+1)$ is not contained in the union of the subcollection, so that $\mathcal{U}$ contains no finite subcover.

We see from this example that it is fairly easy to show that the open interval $(0,1)$ is not compact. As you might guess, the closed interval $[0,1]$ is compact, but this takes much more work to prove. We omit the proof here, but include it for the interested reader in the appendix.

Theorem 3.3.1. *The closed interval* $[0,1]$ *is a compact subset of* $\mathbb{R}$ *with its standard metric.*

Proof. Let $\mathcal{A}$ be an open cover of $[0,1]$. Let

$$C = \{x \in [0,1] \mid [a,x] \text{ is covered by finitely many sets of } \mathcal{A}\}.$$

Then $0 \in C$, since $\mathcal{A}$ is an open cover. Then C is nonempty and bounded above (by 1), so it has a suprememum c. We wish to show that $c \in C$ and that $c = 1$, hence $\mathcal{A}$ admits a finite subcover of $[0,1]$.

We first note that $c > 0$. Indeed, since there is some open set $A \in \mathcal{A}$ with $0 \in A$, it must be that the whole half-interval $[0,\epsilon) \subset A$ for sufficiently small $\epsilon > 0$. Then $[0,\epsilon/2] \subset A \in \mathcal{A}$, and it follows that $c \geqslant \epsilon/2 > 0$.

Now we can show that $c \in C$. Certainly $c \in [0,1]$, so it must be contained in some open set $A \in \mathcal{A}$. Then $(c-\epsilon, c] \subset A$ for sufficiently small $\epsilon > 0$. Writing $[0,c] = [0, c-\epsilon] \cup (c-\epsilon, c]$, we have that $[0,c]$ is contained in a funite subcover of $\mathcal{A}$, so $c \in C$.

Finally, we show that $c = 1$. If not, $c < 1$. Since $c \in A \in \mathcal{A}$ for some open set A, it must be that $[c, c+\epsilon] \subset A$ for some small $\epsilon > 0$. Then $[0, c+\epsilon] = [0,c] \cup [c, c+\epsilon]$ is contained in a finite subcover of $\mathcal{A}$, contradicting the definition of c. Therefore $c = 1$, and this completes the proof of the theorem. □

Theorems About Compactness

This subsection includes some fundamental theorems about compact spaces. We leave most of the proofs as guided exercises.

Proposition 3.3.2. *Let* $f : X \to Y$ *be a continuous map of metric spaces. If* X *is compact then the image of* f *is also compact.*

Proof. Let $\mathcal{U}$ be an open cover of $f(X)$. We form an open cover of X by pulling back each open set $U \in \mathcal{U}$ to the open set $f^{-1}(U)$. The collection of these preimages forms an open cover of X and since X is compact there is a finite subcover $f^{-1}(U_1), \ldots, f^{-1}(U_n)$. Then $U_1, \ldots, U_n$ forms an open subcover of $f(X)$, and since $\mathcal{U}$ was arbitrary it follows that $f(X)$ is compact. □

An easy way to get examples of compact spaces is to take products of compact spaces.

Proposition 3.3.3. *A finite product of compact metric spaces is also compact.*

The Extreme Value Theorem is a theorem you learned in calculus which is of fundamental importance in optimization problems. The next theorem shows that it holds more generally, and is really a statement about topology.

Theorem 3.3.4 (Extreme Value Theorem)**.** *Let (X, d) be a compact metric space. Any continuous function $X \to \mathbb{R}$ achieves its maximum and minimum values.*

The definition of compactness for a general metric space is somewhat abstract. The next two theorems show that in special circumstances, we can replace it with a simpler definition.

Theorem 3.3.5 (Heine-Borel Theorem)**.** *A subset $A \subset \mathbb{R}^d$ is compact if and ony if it is closed and bounded with respect to the standard metric.*

Theorem 3.3.6 (Sequential Compactness Theorem)**.** *Let (X, d) be a metric space. Then X is compact if and only if every sequence in X has a convergent subsequence.*

3.3.2 Connectedness

A *separation* of a metric space (X, d) is a pair of nonempty open sets U and V such that $U \cup V = X$ and $U \cap V = \varnothing$. The metric space is called *connected* if it does not admit a separation. We call a subset $Y \subset X$ *connected* if it is connected as a metric space with its subspace metric.

Proposition 3.3.7. *Let $f : X \to Y$ be a continuous map between metric spaces. If X is connected then the image of f is connected as well.*

Proof. Let U and V be open subsets of Y such that $U \cup V = f(X)$ and $U \cap V = \varnothing$. Consider the preimages $f^{-1}(U)$ and $f^{-1}(V)$. Since f is continuous, the preimages are open. Moreover, it must be that $f^{-1}(U) \cup f^{-1}(V) = X$ and $f^{-1}(U) \cap f^{-1}(V) = \varnothing$. Since X is connected, this implies that one of the preimage sets is empty and it follows that one of the sets U or V is empty as well. Since U and V were arbitrary, it must be that no separation of $f(X)$ exists. $\square$

Theorem 3.3.8. *The subspace $[0, 1] \subset \mathbb{R}$ is connected.*

Proof. By way of obtaining a contradiction, assume that $U \cup V$ is a disconnection of $[0, 1]$. The sets U and V are closed and bounded subsets of $[0, 1]$, so they must be compact as well by the Heine-Borel theorem. Proposition 3.3.3 then implies that $U \times V$ is compact. The distance function $U \times V \to \mathbb{R}$ taking $(x, y) \in U \times V$ to $|x - y|$ is continuous, so it achieves its minimum value, by Proposition 3.3.4. Let $u \in U$ and $v \in V$ be points achieving this minimum, with $|u - v| = t$ and assume without loss of generality that $u \leqslant v$. It must be that $t > 0$, because otherwise U and V intersect. Let $x \in (u, v)$. If $x \in U$, then $|x - v| < t$ gives a contradiction to the assumption that (u, v) is a minimum of the distance function on $U \times V$. Likewise, if $x \in V$, then $|u - x| < t$ gives a contradiction. Therefore no such disconnection of $[0, 1]$ exists. $\square$

Corollary 3.3.9. *The metric space $(\mathbb{R}, |\cdot|)$ is connected.*

Proof. The proof of the previous theorem can be easily adapted to show that any closed interval $[a, b]$ is connected. To obtain a contradiction, suppose that $U \cup V$ is a disconnection of $\mathbb{R}$. Choose a closed interval $[a, b]$ such that $U \cap [a, b] \neq \varnothing$ and $V \cap [a, b] \neq \varnothing$. Then $(U \cap [a, b]) \cup (V \cap [a, b])$ forms a disconnection of $[a, b]$, giving us a contradiction. $\square$

The following is a useful alternate characterization of connectedness.

Proposition 3.3.10. *Let (X, d) be a connected metric space. Then the only subsets of X which are both open and closed are X and $\varnothing$.*

Proof. Let $Y \subset X$ be an arbitrary subset. If Y is both open and closed, then the pair of open sets Y and $X \backslash Y$ satisfies $Y \cup (X \backslash Y) = X$. Then the connectedness assumption implies that one of Y or $X \backslash Y$ is empty, hence $Y = \varnothing$ or $Y = X$. $\square$

Path Connectedness

A *path* in a metric space (X, d) is a continuous map from the interval $I = [0, 1]$ into X. The metric space is said to be *path-connected* if for any $x, y \in X$ there exists a path $\gamma : I \to X$ with $\gamma(0) = x$ and $\gamma(1) = y$.

Proposition 3.3.11. *If (X, d) is path-connected then it is connected.*

Proof. We prove the statement by contrapositive. Assume that X is not connected and let $U \cup V$ form a separation of X. Let $x \in U$ and $y \in V$. We claim that there is no path $\gamma : I \to X$ joining x and y. Indeed, since I is connected, its image under γ must be connected as well. But $(U \cap \gamma(I)) \cup (V \cap \gamma(I))$ would form a separation of its image. $\square$

3.4 Equivalence Relations

3.4.1 Isometry

Let (X, d_X) and (Y, d_Y) be metric spaces. We say that a function $f : X \to Y$ is an *isometry* if it is onto and $d_Y(f(x), f(x')) = d_X(x, x')$ for all $x, x' \in X$. We say that the metric spaces are *isometric* and write $(X, d_X) \sim_{iso} (Y, d_Y)$ if there exists an isometry between them. We will show that $\sim_{iso}$ is an equivalence relation momentarily, but first we need a lemma.

Lemma 3.4.1. *If $f : X \to Y$ is an isometry then it is invertible and its inverse is also an isometry.*

Proof. That f is onto is part of the definition, so we need to show that it is one-to-one. Indeed, if $f(x) = f(x')$, then $0 = d_Y(f(x), f(x')) = d_X(x, x')$ implies that $x = x'$. It follows that f is invertible and it remains to show that $f^{-1} : Y \to X$ is an isometry. Let $y, y' \in Y$. Then

$$d_X(f^{-1}(y), f^{-1}(y')) = d_Y(f(f^{-1}(y)), f(f^{-1}(y'))) = d_Y(y, y'),$$

where the first equality follows from the assumption that f is an isometry. $\square$

Proposition 3.4.2. *The relation $\sim_{iso}$ defines an equivalence relation on the set of metric spaces.*

Proof. For any metric space (X, d), the identity map defines an isometry of the space with itself and it follows that $\sim_{iso}$ is reflexive. The previous lemma shows that $\sim_{iso}$ is symmetric. To show that $\sim_{iso}$ is transitive, let $(X, d_X) \sim_{iso} (Y, d_Y)$ and $(Y, d_Y) \sim_{iso} (Z, d_Z)$. Let $f : X \to Y$ and $g : Y \to Z$ denote isometries. We claim that $g \circ f : X \to Z$ is also an isometry. Indeed, for any $x, x' \in X$,

$$d_Z(g \circ f(x), g \circ f(x')) = d_Y(f(x), f(x')) = d_X(x, x').$$

□

Example 3.4.1. Consider the unit disks $B((0,0),1) \subset \mathbb{R}^2$ and $B((1,0),1) \subset \mathbb{R}^2$, each endowed with subspace metrics for the standard metric on $\mathbb{R}^2$. These metric spaces are isometric, with the isometry $f : B((0,0),1) \to B((1,0),1)$ given by the translation map

$$(x, y) \mapsto (x + 1, y).$$

3.4.2 Homeomorphism

The equivalence relation $\sim_{iso}$ is very restrictive. For our purposes, we will typically want an equivalence relation which isn't required to completely preserve the metric structure, but instead preserves topological structure. Let (X, d_X) and (Y, d_Y) be metric spaces. A *homeomorphism* between them is a map $f : X \to Y$ which is a continuous bijection with continuous inverse. Metric spaces are called *homeomorphic* if there exists a homeomorphism between them. If X and Y are homemorphic metric spaces, we write $X \approx Y$.

Proposition 3.4.3. *The relation $\approx$ defines an equivalence relation on the set of metric spaces. If metric spaces X and Y are isometric, then they are homeomorphic.*

Proof. The proof follows easily from the definition of homeomorphism. The most interesting part of the first part of hte proposition is the transitivity of $\approx$, but this follows easily from Lemma 3.2.2. To see that $X \sim_{iso} Y$ implies $X \approx Y$, it suffices to show that an isomorphism $f : X \to Y$ is continuous and this follows immediately from the $\epsilon - \delta$ definition of continuity. □

Example 3.4.2. The converse of the second part of the proposition does not hold. To prove this, we need to find spaces X and Y which are homeomorphic but not isometric. Consider $X = [0,1]$ and $Y = [0,2]$, each endowed with the subspace metric from $\mathbb{R}$. Then the function $f : X \to Y$ given by $f(x) = 2x$ is a homeomorphism, but it is not an isometry because $d(0,1) \neq d(0,2)$.

Homeomorphism is a much weaker notion of equivalence than isometry. That two spaces are homeomorphic only depends on their underlying topological structure and does not reference distance at all, whereas isometry is defined exactly in terms of distance preservation. In fact, the definition of homeomorphism extends without change to topological spaces (which do not necessarily have a metric).

3.4.3 Homotopy Equivalence

We now arrive at our weakest form of equivalence for metric spaces, which also has the most involved definition. This notion of equivalence is called *homotopy equivalence*. We will not use it in practice too frequently, but it is a fundamental idea of topology and will be useful for describing invariance properties of homology.

Let (X, d_X) and (Y, d_Y) be metric spaces. Two continuous maps $f_0 : X \to Y$ and $f_1 : X \to Y$ are said to be *homotopic* if there exists a continuous map $F : [0,1] \times X \to Y$ such that $F(0,x) = f_0(x)$ and $F(1,x) = f_1(x)$ for all $x \in X$. Spaces (X, d_x) and (Y, d_Y) are said to be *homotopy equivalent* if there exist continuous maps $f : X \to Y$ and $g : Y \to X$ such that $f \circ g : Y \to Y$ is homotopic to the identity map on Y and $g \circ f : X \to X$ is homotopic to the identity map on X. In this case, we write $X \sim_{h.e.} Y$. We leave the proof of the following proposition to the reader.

Proposition 3.4.4. *Homotopy equivalence is an equivalence relation on the set of metric spaces. If spaces X and Y are homeomorphic, then they are homotopy equivalent.*

Example 3.4.3. Let $X = \mathbb{R}$ and $Y = \{0\} \subset \mathbb{R}$. We claim that X and Y are homotopy equivalent. Let $f : X \to Y$ be the constant map $x \mapsto 0$ and let $g : Y \to X$ be the inclusion map $0 \mapsto 0 \in \mathbb{R}$. Then $g \circ f$ is equal to the identity map on Y, so there is nothing to prove here. On the other hand $f \circ g : X \to X$ is the constant map $x \mapsto 0$, and we need to show that this is homotopic to the identity on X. To do so, define $F : [0,1] \times X \to X$ by

$$F(t,x) = t \cdot x.$$

Then F is continuous, $F(0,x) = 0 \cdot x = 0$ is the constant-zero map and $F(1,x) = 1 \cdot x = x$ is the identity map on X. This proves our claim.

On the other hand, X and Y are clearly not homeomorphic, since the sets have different cardinalities (i.e., Y is finite and X is uncountably infinite). This shows that the converse of the second part of the previous proposition does not hold in general.

Homotopy equivalence is a much weaker notion of equivalence than homeomorphic or isometric. Note that, like homeomorphism, homotopy equivalence is perfectly well-defined for general topological spaces.

3.5 Exercises

1. A pair of metrics d and d' on the same set X are said to be *equivalent* if for any $x \in X$ and $r > 0$, there exist positive numbers r' and r'' such that $B_{d'}(x, r') \subset B_d(x, r) \subset B_{d'}(x, r'')$. Show that metric equivalence is an equivalence relation on the set of metrics on a fixed set.

2. Prove Proposition 3.1.3.

3. Work out a more explicit representation of the function d_{S^2} defined in Example 3.1.4. Hint: try to write the distance $d_{S^2}(u, v)$ using the angle between the vectors $u, v \in \mathbb{R}^3$, then relate this to the formula for standard dot product.

4. Prove that the function d_{S^2} defined in Example 3.1.4 is a metric.

5. Prove that the function d_T defined in Example 3.1.5 is a metric.

6. For the set Y shown below, draw the interior $\mathrm{int}(Y)$, boundary ∂Y and closure $\overline{Y}$.

7. Consider $\mathbb{R}^2$ with its standard metric. Classify the following sets as open, closed, open and closed, or none of the above:
   a) $[0,1) \times [0,1]$
   b) $\mathbb{R}^2 \backslash \{(0,0)\}$
   c) $\{(a,a) \mid a \in \mathbb{R}\}$
   d) $([0,1] \times [0,1]) \backslash ((1/4, 3/4) \times (1/4, 3/4))$

8. Prove Lemma 3.2.2.

9. Prove Proposition 3.3.3.

10. Prove the Heine-Borel Theorem Add hints.

11. Prove the Sequential Compactness Theorem Add hints.

12. Prove Proposition 3.4.4.

13. Prove that the sets $\mathbb{R}^2$ and $D^2 = \{x \in \mathbb{R}^2 \mid \|x\| \leqslant 1\}$.

14. Prove that the sets $\mathbb{R}^2 \backslash \{0\}$ and $S^1 = \{x \in \mathbb{R}^2 \mid \|x\| = 1\}$ are homotopy equivalent.

4 Homology of Simplicial Complexes

In this chapter we begin to study a particular type of topological space called a *simplicial complex*. Our next major goal is to understand how to encode the topological features of a simplicial complex by a list of vector spaces called *homology groups*. The next few chapters follow quite closely the notation and exposition of the excellent survey paper [1].

4.1 Motivation: Distinguishing Topological Spaces

The fundamental question of topology is as follows: given two topological space (metric spaces, if you like) X and Y, are X and Y homeomorphic? If you suspect that the answer is "yes", then you need only to produce such a homeomorphism. However, if you think that the answer is "no", then you need to prove that no such homeomorphism can possibly exist! In this section we examine some simple examples which will convince us that some sophisticated tools might be necessary to answer this question.

Example 4.1.1. Are the spaces $X = (0,1)$ and $Y = \mathbb{R}$ homeomorphic?

While the spaces X and Y are quite different from a metric space perspective (one has diameter 1, the other is unbounded), they are the same topologically. To see this we construct a homeomorphism. First note that $(0,1)$ and $(-\pi/2, \pi/2)$ are homeomorphic by a simple map $f : (0,1) \to (-\pi/2, \pi/2)$ defined by

$$f(x) = \pi \cdot x - \pi/2.$$

It therefore suffices to find a homeomorphism $g : (-\pi/2, \pi/2) \to \mathbb{R}$, and this is given by

$$g(x) = \tan(x).$$

Indeed, g is continuous, bijective, and its inverse $g^{-1}(x) = \arctan(x)$ is also continuous.

Example 4.1.2. Are the spaces $X = [0,1]$ and $Y = \mathbb{R}$ homeomorphic?

This is similar to the last example, but X does feel topologically distinct from $\mathbb{R}$ in that X contains some boundary points. Thus we claim that X and Y are not homeomorphic, and our goal is to show that no homeomorphism $f : X \to Y$ can possibly exist. A standard trick is to look for a specific topological property that one space has and the other doesn't. In this case, we know that X is compact and that $\mathbb{R}$ is not. Then the image of any continuous map $f : X \to Y$ must also be compact and it follows that any such continuous map cannot be surjective! Therefore X and Y are not homeomorphic.

Example 4.1.3. Are the spaces $X = [0,1)$ and $Y = (0,1)$ homeomorphic?

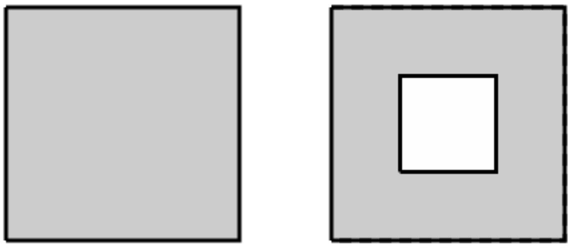


Figure 4.1: The spaces $X = [0, 1) \times (0, 1)$ and $Y = (0, 1) \times (0, 1)$ are shown in the top row. The second row shows each space with a point removed. For X, the point is chosen so that the resulting space has no "holes". For Y, any choice of point to remove results in hole in the space's interior.

We once again suspect that the answer is "no", but in this case neither space is compact, so we will need a different strategy. The following lemma (and its obvious generalizations) will be useful.

Lemma 4.1.1. *Let $f : X \to Y$ be a homeomorphism. Then for any $x \in X$, the restriction of f to $X\backslash\{x\}$ is a homeomorphism onto $Y\backslash\{f(x)\}$.*

Proof. The restricted map is clearly still a bijection. To see that it is continuous, let $U \subset Y\backslash\{f(x)\}$ be an open set. Then $U = U'\backslash\{f(x)\}$ for some open set $U' \subset Y$, and it follows that

$$f^{-1}(U) = f^{-1}(U'\backslash\{f(x)\}) = f^{-1}(U)\backslash\{x\}$$

is open in $X\backslash\{x\}$. Therefore f is continuous. Continuity of f^{-1} follows similarly. □

Now we note that, for our particular example, $X\backslash\{0\} = (0, 1)$ is a connected set, but $Y\backslash\{f(0)\}$ is not connected (removing any point from Y results in a set which is not connected). Since connectedness is preserved by continuous maps, it follows that there is no homeomorphism $f : X \to Y$, by contrapositive to the lemma.

Example 4.1.4. Are the spaces $X = [0, 1) \times [0, 1]$ and $Y = ([0, 1] \times [0, 1])\backslash((1/4, 3/4) \times (1/4, 3/4))$ homeomorphic (see Figure 4.1)?

Your intuition should be that the answer is "no". However, none of our previous tricks will work here: both X and Y are connected and compact, and removing a finite number of points from X or Y will not result in a disconnected space. However, Y is "obviously" different from X because it has a "hole". How do we detect the presence of this hole using topology?

The goal of this chapter is to develop a tool called *homology* which is an algorithm for counting "holes" of various dimensions in a topological space. To do so for a general topological or metric space is quite technical (this is discussed briefly in Section 4.5), so we will restrict to a special class of spaces called *simplicial complexes*. Roughly, these are spaces which are pieced together in a controlled way from a collection of triangles and higher-dimensional analogues of triangles. Since these objects have an intuitively "linear" structure, one might hope that the process of counting holes in the spaces can be reduced a linear algebra operation!

Once we have the tools of homology in hand, you will be able to distinguish the spaces from Example 4.1.4 with ease. You will do so in the exercises.

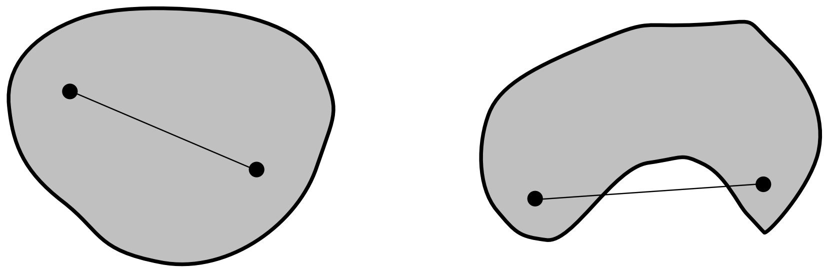
4.2 Simplicial Complexes

4.2.1 Geometric Simplicial Complexes

Convex Sets

A subset S of $\mathbb{R}^k$ is said to be *convex* if for any points $x, y \in S$, each point $(1-t)x + ty$, $t \in [0,1]$, along the interpolation between x and y is also contained in S. Otherwise S is said to be *nonconvex*.

Remark 4.2.1. *In the above, we are using $x, y \in S$ to denote* points *in $\mathbb{R}^k$, but the expression $(1-t)x + ty$ treats x and y as* vectors. *We will frequently conflate the notion of a point $x \in \mathbb{R}^k$ with the vector with basepoint at $\vec{0}$ and endpoint at x.*

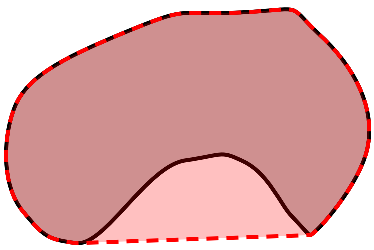


Convex Nonconvex

The *convex hull of S* is smallest convex subset of $\mathbb{R}^k$ which contains S and is denoted $\mathrm{cvx}(S)$. More precisely,

$$\mathrm{cvx}(S) = \bigcap \{C \mid S \subset C \subset \mathbb{R}^k \text{ and } C \text{ is convex}\}.$$

The figure below shows a set overlaid with its convex hull.



Simplices

Let $S = \{x_0, x_1, \ldots, x_n\}$ be a finite subset of $\mathbb{R}^k$. The set S is said to be in *general position* if its points are not contained in any affine subspace of $\mathbb{R}^k$ of dimension less than n (thus $n \leqslant k$). Recall that an *affine subspace* of $\mathbb{R}^k$ is a set of the form

$$x + V = \{x + v \mid v \in V\},$$

where $V \subset \mathbb{R}^k$ is a vector subspace (see the figure below).

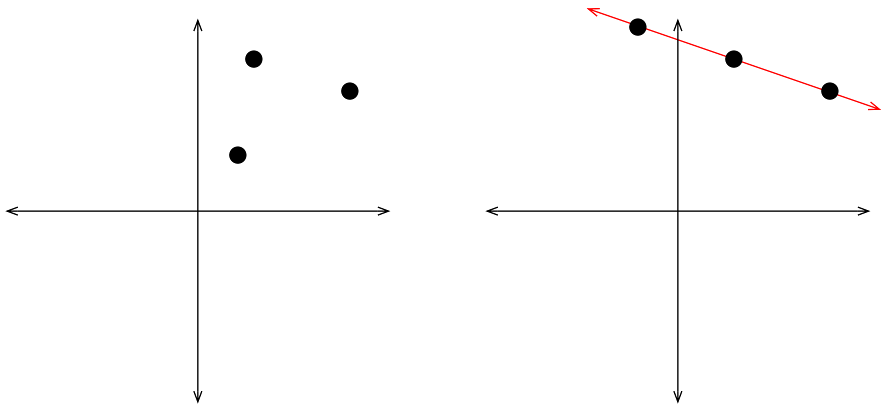


Figure 4.2: The figure on the left shows a set $\{x_0, x_1, x_2\}$ of 3 points in $\mathbb{R}^2$ which are in general position—any line can only contain 2 of the points. The figure on the right shows a set of points which are *not* in general position. The 1-dimensional affine subspace containing all the points is indicated in red.

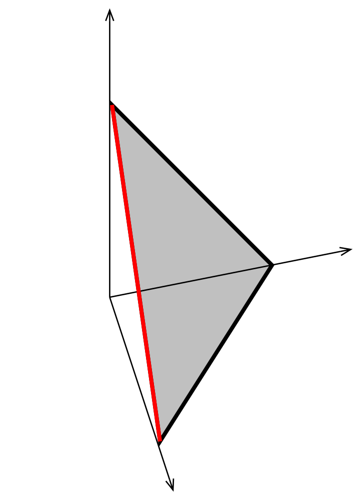
For a set S in general position, the *simplex associated to* S is the set $\sigma(S) = \mathrm{cvx}(S)$. The points x_i are called the *vertices* (the singular form is *vertex*) of $\sigma(S)$. Any pair of distinct points $x_i, x_j \in S$ determing their own simplex, called an *edge* of $\sigma(S)$. In general, for subset $T \subset S$, $\sigma(T)$ is called a *face* of $\sigma(S)$. The number n is called the *dimension* of $\sigma(S)$.

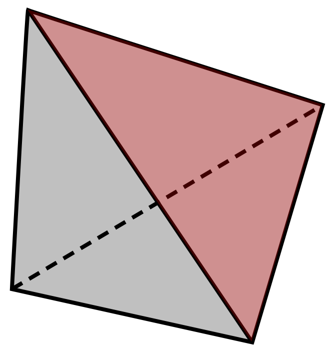
Frequently, we will only be interested in the simplex $\sigma(S)$ and not in the particular set S which defines it. We will refer to a set $\sigma \subset \mathbb{R}^n$ as a *simplex* if $\sigma = \sigma(S)$ for some set S in general position.

Example 4.2.1. It will be convenient to have a standard picture to refer to. The *standard n-dimensional simplex* is the simplex associated to the set

$$S = \{(1, 0, 0, \ldots, 0), (0, 1, 0, \ldots, 0), \ldots, (0, 0, 0, \ldots, 1)\} \subset \mathbb{R}^{n+1}$$

consisting of all points on the coordinate axes at Euclidean distance 1 from the origin. The figure on the left shows the standard 2-dimensional simplex with the 1-dimensional face (i.e., an edge) $\sigma((1, 0, 0), (0, 0, 1))$ highlighted. The figure on the right shows a (nonstandard) 3-dimensional simplex embedded in $\mathbb{R}^3$ with one of its 2-dimensional faces highlighted.





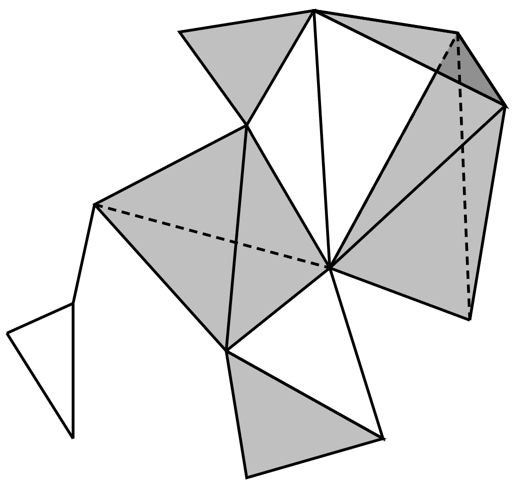
Simplicial Complexes

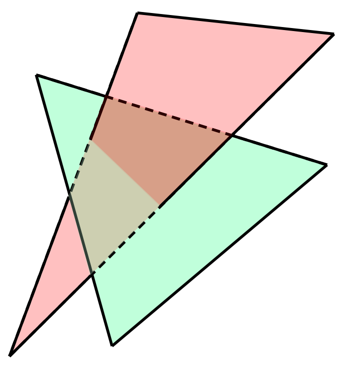
A *(geometric) simplicial complex* is a collection of simplices $\mathcal{X}$ in some $\mathbb{R}^n$ satisfying:

1. for any simplex $\sigma \in \mathcal{X}$, all faces of σ are also contained in $\mathcal{X}$,
2. for any two simplices $\sigma, \tau \in \mathcal{X}$, the intersection $\sigma \cap \tau$ is also a simplex and which is a face of both σ and τ.

Intuitively a simplicial complex is a shape obtained by gluing together a collection of simplices, where gluing is only allowed to take place along faces.

Example 4.2.2. The figure on the left shows a complicated simplicial complex in $\mathbb{R}^3$. The figure on the right is a collection of simplices in $\mathbb{R}^3$ which is *not* a simplicial complex since the simplices do not interesect along faces.





4.2.2 Abstract Simplicial Complexes

As a subspace of $\mathbb{R}^n$ (for some n), any simplicial complex inherits the structure of a metric space. We can therefore study simplicial complexes up to the equivalence relation of isometry. This notion of equivalence is too rigid in many applications, and we are actually primarily interested in *topological* properties of simplicial complexes. This means that we really wish to study simplicial complexes up to the equivalence relation of homeomorphism.

Fortunately, simplicial complexes are a class of geometric objects whose topological structure can be encoded very efficiently—this is exactly the reason that we wish to use simplicial complexes as a way to encode the topology of data! The topological information from a simplicial complex that we are after can be deduced from the combinatorial structure of the complex, which is encoded in the number of its simplices of various dimensions and in the way that the various simplices intersect. The topological information does not depend on the particular geometric embedding of the complex in a Euclidean space. With this motivation in mind, we will give a more abstract definition of a simplex in terms of the combinatorial (topological) information.

An *abstract simplicial complex* is a pair $X = (V(X), \Sigma(X))$ (we will also use the notation $X = (V, \Sigma)$), where $V(X)$ is a finite set and $\Sigma(X)$ is a collection of subsets of $V(X)$ such that for any $\sigma \in \Sigma(X)$ and any nonempty $\tau \subset \sigma$, $\tau \in \Sigma(X)$. The elements of $V(X)$ are called the *vertices* of X and the elements of $\Sigma(X)$ are called the *simplices* or *faces* of X. Faces containing exactly two vertices are called *edges*. Faces containing exactly $(k+1)$-vertices are called *k-dimensional faces*, or just *k-faces*. If σ is a k-face, then a *$(k-1)$-face of σ* is a $(k-1)$-face τ with $\tau \subset \sigma$.

Example 4.2.3. Let $\mathcal{X}$ be a simplicial complex. We can construct an abstract simplicial complex X associated to $\mathcal{X}$ by first taking $V(X)$ to be the union of all vertices of all simplices contained in $\mathcal{X}$. We include a subset of $V(X)$ in $\Sigma(X)$ if and only if the subset consists of the vertices of some simplex in $\mathcal{X}$. We leave it as an exercise to show that X is really an abstract simplicial complex.

Example 4.2.4. The *standard n-dimensional abstract simplex* is the abstract simplicial complex Δ^n with vertex set $\{0, 1, 2, \ldots, n\}$ and edge set consisting of every non-empty subset of the vertex set. How does this compare to the standard simplex defined in Example 4.2.1 (see the exercises)?

A *map of abstract simplicial complexes* X and Y is a map $f : V(X) \to V(Y)$ such that for all $\sigma \in \Sigma(X)$, $f(\sigma) \in \Sigma(Y)$. This means that for any collection $\{v_1, \ldots, v_{k+1}\}$ of vertices of X which define a k-simplex, the set $\{f(v_1), \ldots, f(v_{k+1})\}$ defines some simplex in Y. Note that we do not require f to be injective, so it is possible that the image set contains redundant entries, whence it defines a lower-dimensional simplex. A map of abstract simplicial complexes f from X to Y is called an *simplicial isomorphism* if it is a bijection and if for all $\tau \in \Sigma(Y)$, $f^{-1}(\tau) \in \Sigma(X)$. Two abstract simplicial complexes are *simplicially isomorphic* if there is a simplicial isomorphism between them.

Example 4.2.3 shows that for any geometric simplicial complex $\mathcal{X}$, we can construct an associated abstract simplicial complex. In the other direction, for an abstract simplicial complex X, a (geometric) simplicial complex $\mathcal{X}$ is called a *geometric realization of* X if the abstract simpicial complex associated to $\mathcal{X}$ is simplicially isomorphic to X. We use the notation $|X|$ for a geometric realization of X. Note that $|X|$ is highly nonunique! Also note that we have yet to show that any such $|X|$ exists.

Proposition 4.2.2. *A geometric realization exists for any abstract simplicial complex* $X = (V(X), \Sigma(X))$.

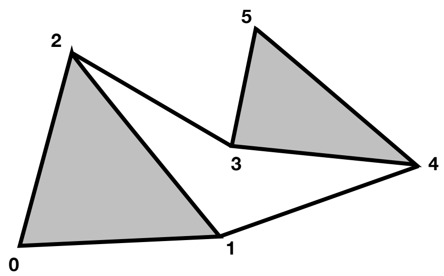
Proof. Assume $V(X)$ contains $n+1$ points. We embed the vertex set in $\mathbb{R}^{n+1}$ by mapping the elements of $V(X)$ to the vertices of the standard n-dimensional simplex. For each subset $\Sigma(X)$, we include the simplex in the geometric realization which is formed by taking the convex hull of the corresponding vertices. □

The construction given in the proof always works, but it is somewhat inefficient in the sense that an abstract simplicial complex with $n+1$ vertices can be geometrically realized in $\mathbb{R}^k$ for k far smaller than $n+1$.

Example 4.2.5. Consider the abstract simplicial complex $X = (V(X), \Sigma(X))$ with $V(X) = \{0, 1, 2, 3, 4, 5\}$ and

$$\begin{aligned}\Sigma(X) = \{&\{0,1,2\}, \{0,1\}, \{0,2\}, \{1,2\}\\ &\{3,4,5\}, \{3,4\}, \{3,5\}, \{4,5\}\\ &\{2,3\}, \{1,4\}\}\end{aligned}$$

A geometric realization of X is shown in the figure below.



The realization is a subset of $\mathbb{R}^2$, rather than the $\mathbb{R}^6$ required by the construction in the proof of Proposition 4.2.2.

Remark 4.2.3. *Given a geometric simplicial complex $\mathcal{X}$, we can form its abstract simplicial complex X as set with finitely many elements. The abstract simplicial complex is a very compact representation of the topology and combinatorics of $\mathcal{X}$, but it completely loses the metric space structure of of $\mathcal{X}$.*

4.3 Topological Invariants of Simplicial Complexes

4.3.1 Connected Components

Let $\mathcal{X}$ be a simplicial complex and let X denote its associated abstract simplicial complex. Let $V(X)$ denote the vertex set of X. We can define an equivalence relation $\sim$ on $V(X)$

by

$$v \sim v' \Leftrightarrow \{v, v'\} \text{ is an edge of } X.$$

We then have the following proposition which relates the connected components of the metric space $\mathcal{X}$ to the set $V(X)/\sim$.

Proposition 4.3.1. *The connected components of $\mathcal{X}$ are in bijective correspondence with $V(X)/\sim$.*

Proof. Let $\mathcal{C} = \{\mathcal{X}_0, \ldots, \mathcal{X}_N\}$ denote the set of connected components of $\mathcal{X}$ (remember, we are only considering *finite* simplicial complexes, so it is okay to assume that $\mathcal{X}$ has finitely many components). We define a map $\mathcal{C} \to V(X)/\sim$ by taking $\mathcal{X}_k$ to the equivalence class in $V(X)/\sim$ containing any vertex of $\mathcal{X}_k$. We need to check that this map is well-defined and that it is a bijection.

To see that the map is well-defined, let v, v' be vertices of $\mathcal{X}_k$. Since $\mathcal{X}_k$ is path-connected, there exists a sequence of edges $v_0, v_1, \ldots, v_m$ such that $v = v_0$, $v' = v_m$ and $\{v_i, v_{i+1}\}$ is an edge of $\mathcal{X}_k$ for all i. Then the transitivity of the equivalence relation $\sim$ implies that v and v' lie in the same equivalence class in $V(X)/\sim$.

To see that the map is a bijection, we can explicitly define its inverse. We define a map $V(X)/\sim \to \mathcal{C}$ by taking the equivalence class of v to the component of $\mathcal{X}$ which contains v. By a similar argument to the last paragraph, this map is well-defined. Moreover, it is easy to see that this is the inverse to the map defined in the first paragraph. $\square$

4.3.2 Back to Linear Algebra: Free Vector Spaces

Before moving on to studying more interesting topological invariants of simplicial complexes, we pause to introduce the very important notion of a free vector space on a set. Let $\mathbb{F}$ be a field and let S be a finite set. The *free vector space over $\mathbb{F}$ on the set S* is the vector space $V_{\mathbb{F}}(S)$ with underlying set consisting of functions $\phi : S \to \mathbb{F}$. We will sometimes shorten notation to $V(S) = V_{\mathbb{F}}(S)$ when the field is understood to be fixed. The vector space operations are defined pointwise: for $\phi, \phi' \in V_{\mathbb{F}}(S)$, $\lambda \in \mathbb{F}$ and $s \in S$,

$$(\phi + \phi')(s) = \phi(s) + \phi'(s),$$
$$(\lambda \cdot \phi)(s) = \lambda \cdot \phi(s).$$

The zero vector in $V_{\mathbb{F}}(S)$ is the *zero function*; i.e., the function which takes every element of S to zero.

For each $s \in S$, let ϕ_s denote the *characteristic function for s* defined by

$$\phi_s(s') = \begin{cases} 1_{\mathbb{K}} & s' = s \\ 0_{\mathbb{K}} & s' \neq s. \end{cases}$$

Proposition 4.3.2. *The set of characteristic functions of the elements of S forms a basis for $V_{\mathbb{F}}(S)$. It follows that $\dim(V_{\mathbb{F}}(S)) = |S|$.*

Proof. We need to show that $\{\phi_s\}_{s \in S}$ is spanning and linearly independent. Let ϕ be an arbitrary element of $V_{\mathbb{F}}(S)$. For each $s \in S$, let $\lambda_s = \phi(s)$. Then we can write ϕ as the linear combination

$$\phi = \sum_{s \in S} \lambda_s \phi_s,$$

and this shows that $\{\phi_s\}_{s \in S}$ is a spanning set.

To show that it is linearly independent, consider an arbitrary linear combination $\phi = \sum_{s \in S} \alpha_s \phi_s$. If the linear combination is equal to the zero function, then for each $s' \in S$, we have

$$0_{\mathbb{F}} = \phi(s') = \sum_{s \in S} \alpha_s \phi_s(s') = \alpha_{s'}.$$

Since s' was arbitrary, it must be that each coefficient in the linear combination is zero. $\square$

We will refer to the basis consisting of characteristic functions as the *standard basis for* $V_{\mathbb{F}}(S)$.

Any map $f : S \to T$ of sets induces a linear map $V_{\mathbb{F}}(f) : V_{\mathbb{F}}(S) \to V_{\mathbb{F}}(T)$ by extending linearly the function defined on basis functions by

$$V_{\mathbb{F}}(f)(\phi_s) = \phi_{f(s)}.$$

Proposition 4.3.3. *Let $f : S \to T$ be a map of sets. The induced linear map is given by the following general formula for $\phi \in V_{\mathbb{F}}(S)$:*

$$\left(V_{\mathbb{F}}(f)(\phi)\right)(t) = \sum_{s \in S \mid f(s) = t} \phi(s).$$

The proof of the proposition is left as an exercise.

Let S be a finite set and let $R \subset S \times S$ be a binary relation. We define a subspace $V_{\mathbb{F}}(R) \subset V_{\mathbb{F}}(S)$ by

$$V_{\mathbb{F}}(R) = \operatorname{span}\{\phi_s - \phi_{s'} \mid (s, s') \in R\}.$$

Proposition 4.3.4. *There is an isomorphism of vector spaces*

$$V_{\mathbb{F}}(S)/V_{\mathbb{F}}(R) \approx V_{\mathbb{F}}(S/R).$$

Proof. We define a map

$$L : V_{\mathbb{F}}(S/R) \to V_{\mathbb{F}}(S)/V_{\mathbb{F}}(R)$$

by linearly extending the map defined on basis vectors by

$$L(\phi_{[s]}) = [\phi_s].$$

First we need to check that this map is actually well-defined. This means that we need to show that $s \sim s'$ implies $[\phi_s] = [\phi_{s'}]$. The latter equality holds if and only if $\phi_s - \phi_{s'} \in V_{\mathbb{F}}(R)$, which holds by definition.

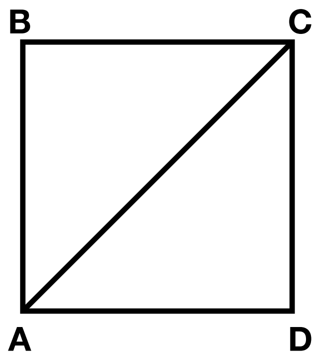
Next we need to show that the map is injective. Let $\sum_{[s]} \lambda_{[s]}\phi_{[s]}$ denote an arbitrary element of $V_{\mathbb{F}}(S/R)$ and assume that it maps by L to the zero vector. Then

$$0 = L\left(\sum_{[s]} \lambda_{[s]}\phi_{[s]}\right) = \sum_{[s]} \lambda_{[s]}[\phi_s]$$

implies that all $\lambda_{[s]} = 0$ by linear independence of the vectors $[\phi_s]$. The kernel of L is the zero vector, and L is therefore injective. Finally, the fact that the vector spaces are of the same dimension shows that L is surjective as well. $\square$

4.3.3 First Example

We first consider the example shown below, which we have already identified as a metric space with the metric inherited from $\mathbb{R}^2$. Denote this metric space by $\mathcal{X}$.



Now that we have the proper definitions, we easily see that this shape can be thought of as a (geometric) simplicial complex in $\mathbb{R}^2$. Let X denote the associated abstract simplicial complex for $\mathcal{X}$. Its vertex set is

$$\mathrm{Vert}(X) = \{A, B, C, D\}$$

and its edge set is

$$\mathrm{Edge}(X) = \{\{A, B\}, \{A, C\}, \{A, D\}, \{B, C\}, \{C, D\}\}.$$

To simpify notation, we will denote 1-dimensional simplices (edges) by, e.g., AB rather than $\{A, B\}$.

We can visually see that this metric space consists of one path component (it is connected), and that it contains a pair of “loops” which apparently cannot be shrunk to a point. Our goal is to develop a computational approach which will allow us to discern the apparent *topological* features of the shape from the *combinatorial* information given by its simplicial decomposition. This will be accomplished by using tools from linear algebra.

Consider the vector space $C_0(X) := V_{F_2}(\mathrm{Vert}(X))$, the free vector space over F_2 (the field with 2 elements) generated by $\mathrm{Vert}(X)$. (This is a standard notation that will be explained in the following chapter.) Similarly, let $C_1(X) := V_{F_2}(\text{Edge}(X))$ denote the free vector space over F_2 generated by $\text{Edge}(X)$. There is a natural, geometrically-motivated linear map

$$\partial_1 : C_1(X) \to C_0(X)$$

called the *boundary map*. Each basis element of $C_1(X)$ corresponds to an edge of X, and the boundary map takes a basis element to the linear combination of its boundary vertices. For example,

$$\partial_1(\phi_{AB}) = \phi_A + \phi_B,$$

where we are using the notation of Section 4.3.2 and denoting the basis element of $C_1(X)$ assocated to the edge AB by ϕ_{AB}. With respect to this basis, the matrix form of this linear map is given by

$$\partial_1 = \begin{array}{c} \\ {\scriptstyle A} \\ {\scriptstyle B} \\ {\scriptstyle C} \\ {\scriptstyle D} \end{array} \begin{array}{c} \begin{array}{ccccc} {\scriptstyle AB} & {\scriptstyle AC} & {\scriptstyle AD} & {\scriptstyle BC} & {\scriptstyle CD} \end{array} \\ \begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 \end{bmatrix} \end{array}.$$

By construction, the column space of ∂_1 is exactly the span of vectors of the form $\phi_v + \phi'_v$, where $v, v' \in \text{Vert}(X)$ and $\{v, v'\} \in \text{Edge}(X)$. Said differently, $\text{image}(\partial_1)$ is the free vector space associated to the set $\{(v, v') \mid \{v, v'\} \in \text{Edge}(X)\}$. Applying Proposition 4.3.4, we see that the quotient space

$$C_0(X)/\text{image}(\partial_1)$$

is isomorphic to the free vector space associated to the set $\text{Vert}(X)/\sim$, where $v \sim v' \Leftrightarrow \{v, v'\} \in E(X)$. Proposition 4.3.1 says that $V(X)/\sim$ is in bijective correspondence with the set of connected componenets of $\mathcal{X}$. We have just proved the following Proposition, which works for *any* simplicial complex $\mathcal{X}$ (defining $C_j(X, F_2)$ and ∂_1 in the appropriately generalized ways—see Section 4.4).

Proposition 4.3.5. *The number of connected components of $\mathcal{X}$ is given by the dimension of the vector space $C_0(X, F_2)/\text{image}(\partial_1)$.*

For our specific example, performing row-reduction on the matrix (keeping in mind that we are working over F_2, where $1 + 1 = 0$), we obtain

$$\begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

By inspection, we see that the first three columns of the reduced matrix form a basis for its column space. This means that the image of ∂_1 is 3-dimensional. We conclude that $C_0(X, F_2)/\text{image}(\partial_1)$ is 1-dimensional, corresponding to the fact that the space $\mathcal{X}$ is connected!

We have seen that the image of ∂_1 gives us important geometric about $\mathcal{X}$. The next natural step would be to examine the kernel of ∂_1. By the rank-nullity theorem (Theorem 2.4.3) the kernel of ∂_1 must be 2-dimensional. We leave it to the reader to check that the vectors

$$\begin{pmatrix} 1 \\ 1 \\ 0 \\ 1 \\ 0 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} 0 \\ 1 \\ 1 \\ 0 \\ 1 \end{pmatrix}$$

form a basis for $\ker(\partial_1)$. In terms of our basis, this means that the kernel is spanned by the vectors

$$\phi_{AB} + \phi_{AC} + \phi_{BC} \quad \text{and} \quad \phi_{AC} + \phi_{AD} + \phi_{CD}.$$

Looking back at the picture of $\mathcal{X}$, we see that these vectors exactly describe the apparent loops in the shape! Indeed, one can intuitively think of the sums in the vector space $C_1(X, F_2)$ as unions, so that the vectors listed above correspond to the unions of edges

$$\{A,B\} \cup \{A,C\} \cup \{B,C\} \quad \text{and} \quad \{A,C\} \cup \{A,D\} \cup \{C,D\}, \tag{4.1}$$

respectively. It is visually obvious that these are loops in $\mathcal{X}$.

Apparently (at least for this simple example) the dimension of the kernel of ∂_1 counts the number of loops in the simplicial complex. At this point, one might object: there are other loops in $\mathcal{X}$ which are not contained in the list (4.1). The most obvious is the loop

$$\{A,B\} \cup \{B,C\} \cup \{C,D\} \cup \{A,D\}.$$

This is where we see that the extra vector space structure is important in our construction. This loop corresponds (under our informal association of unions with sums) to the vector

$$\phi_{AB} + \phi_{BC} + \phi_{CD} + \phi_{AD}.$$

In matrix notation, we have

$$\phi_{AB} + \phi_{BC} + \phi_{CD} + \phi_{AD} = (1,0,1,1,1),$$

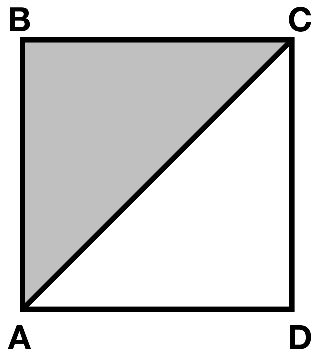
which can be expressed as the linear combination

$$(1,0,1,1,1) = (1,1,0,1,0) + (0,1,1,0,1) = (\phi_{AB} + \phi_{AC} + \phi_{BC}) + (\phi_{AC} + \phi_{AD} + \phi_{CD}).$$

(Remember that we are working over the field F_2!) Thus the loop $\{A,B\} \cup \{B,C\} \cup \{C,D\} \cup \{A,D\}$ can be viewed as a linear combination of the loops in the list (4.1). We finally conclude that the dimension of the kernel of ∂_1 counts the loops in $\mathcal{X}$ which are *independent* in this sense.

4.3.4 Second Example

Now consider the metric space shown below. We denote this simplicial complex by $\mathcal{Y}$ and the associated abstract simplicial complex by Y.



The space $\mathcal{Y}$ is clearly a slight modification of the space $\mathcal{X}$ from the previous section. In particular, the vertex and edge sets of Y the same as those of X:

$$V(Y) = \{A, B, C, D\} \quad \text{and} \quad E(Y) = \{\{A, B\}, \{A, C\}, \{A, D\}, \{B, C\}, \{C, D\}\}.$$

To obtain $\mathcal{Y}$ from $\mathcal{X}$, we add in a single 2-simplex. The *face set* for Y is thus

$$F(Y) = \{\{A, B, C\}\}.$$

Similar to the previous section, we use $C_0(Y, F_2)$ and $C_1(Y, F_2)$ to denote the free vector spaces over F_2 generated by $V(Y)$ and $E(Y)$, respectively. Due to the presence of a 2-dimensional simplex, we also introduce the notation $C_2(Y, F_2)$ for the free vector space over F_2 generated by $F(Y)$. As before, we are interested in the boundary map $\partial_1 : C_1(Y, F_2) \to C_0(Y, F_2)$, which has exactly the same matrix representation as the map ∂_1 in the previous section.

From the previous section, we see that ∂_1 still has 3-dimensional image. Applying Proposition 4.3.5, we easily deduce the (visually obvious) fact that $\mathcal{Y}$ has a single connected component. Likewise, the kernel of ∂_1 is 2-dimensional, and is spanned by the vectors

$$\phi_{AB} + \phi_{AC} + \phi_{BC} \quad \text{and} \quad \phi_{AC} + \phi_{AD} + \phi_{CD}.$$

Now we have run into a problem: the loop corresponding to $\{A, B\}$, $\{A, C\}$ and $\{B, C\}$ has been "filled in" by a 2-simplex in order to construct $\mathcal{Y}$. Due to the presence of a higher-dimensional simplex in $\mathcal{Y}$, there is another natural map of interest. This is the map $\partial_2 : C_2(Y, F_2) \to C_1(Y, F_2)$ defined on the basis for (the 1-dimensional vector space) $C_2(Y, F_2)$ by

$$\phi_{ABC} \mapsto \phi_{AB} + \phi_{AC} + \phi_{BC}.$$

That is, ∂_2 takes the vector corresponding to the 2-simplex $\{A, B, C\}$ to the linear combination of vectors corresponding to edges along its boundary. For this reason, ∂_2 is also called a *boundary map*. In matrix form, we have

$$\partial_2 = \begin{array}{c} \\ {\scriptstyle AB} \\ {\scriptstyle AC} \\ {\scriptstyle AD} \\ {\scriptstyle BC} \\ {\scriptstyle CD} \end{array} \begin{array}{c} {\scriptstyle ABC} \\ \begin{bmatrix} 1 \\ 1 \\ 0 \\ 1 \\ 0 \end{bmatrix} \end{array}.$$

The image of ∂_2 is clearly the span of the vector $\phi_{AB} + \phi_{AC} + \phi_{BC}$, which corresponds to the boundary of the single 2-simplex in $\mathcal{Y}$. We therefore see that the number of linearly independent loops in $\mathcal{Y}$ (which are not filled in by a face) is given by the dimension of the vector space

$$\begin{aligned} \ker(\partial_1)/\text{image}(\partial_2) &= \text{span}_{F_2}\{\phi_{AB} + \phi_{AC} + \phi_{BC}, \phi_{AC} + \phi_{AD} + \phi_{CD}\}/\text{span}_{F_2}\{\phi_{AB} + \phi_{AC} + \phi_{BC}\} \\ &\approx \text{span}_{F_2}\{\phi_{AC} + \phi_{AD} + \phi_{CD}\}. \end{aligned}$$

This construction works for general simplicial complexes, so we state our conclusion in the following (slightly ill-defined) proposition.

Proposition 4.3.6. *The number of linearly independent loops in a simplicial complex $\mathcal{X}$ is given by the dimension of the vector space* $\ker(\partial_1)/\text{image}(\partial_2)$.

There is an ambiguity in the proposition: how do we know that $\text{image}(\partial_2) \subset \ker(\partial_1)$? After all, if this is not the case then the proposition doesn't even make sense. We will delay the proof until the next section, where it will be done in far greater generality. For now, we claim that this proposition makes intuitive sense by the construction of the boundary maps. The kernel of ∂_1 contains linear combinations of edges which "have no boundary"—that is, the union of the edges gives a closed loop. The image of ∂_2 contains linear combinations of edges which are the boundary of a 2-simplex—that is, unions of edges which bound a face. Thus the vector space $\ker(\partial_1)/\text{image}(\partial_2)$ contains all loops, modulo those loops which are filled in by a face.

4.4 Homology of Simplicial Complexes over F_2

4.4.1 Chain Complexes

Chain Groups

Let $\mathcal{X}$ be a (finite) simplicial complex and let X denote its associated abstract simplicial complex. We define the *k-th chain group* of $\mathcal{X}$ over F_2 to be the free vector space over F_2 generated by the set of k-dimensional simplices of X. The k-th chain group is denoted $C_k(X)$, or sometimes simply by C_k when the simplicial complex $\mathcal{X}$ is understood to be fixed.

Remark 4.4.1. *As defined, these C_k are really just vector spaces. We call these vector spaces* chain groups *in order to match with the terminology used in most literature. A* *group is a more general algebraic structure than a vector space—in particular, the additive structure of any vector space turns it into a group. By using this more general structure, we can define chain groups with other coefficients; e.g. one frequently considers chain groups over the integers $C_k(X;\mathbb{Z})$. For our purposes, it will be sufficient to consider the chain groups as vector spaces. See Section 4.5 for a brief discussion of these more general chain groups.*

Boundary Maps

For each k, we define the *boundary map*

$$\partial_k : C_k \to C_{k-1}$$

by defining it on the basis

$$\{\phi_\sigma \mid \sigma \text{ is a } k\text{-simplex of } X\} \tag{4.2}$$

via the formula

$$\partial_k \phi_\sigma = \sum \{\phi_\tau \mid \tau \text{ is a } (k-1)\text{-dimensional face of } \sigma\}.$$

The reason for calling ∂_k the "boundary map" should be clear. Indeed, ∂_k takes a k-simplex to the sum of $(k-1)$-simplices which lie along its boundary!

Alternate Forms of ∂_k

We will give two alternative forms of the map ∂_k which will be convenient in different contexts. First let $\sigma = \{v_0, \ldots, v_k\}$ be a k-simplex of X with vertices v_i. Then ∂_k applied to this simplex takes the explicit form

$$\partial_k \phi_\sigma = \sum_{j=0}^{k} \phi_{\{v_0,\ldots,\widehat{v_j},\ldots,v_k\}}.$$

Here we use the hat to denote ommision of the vertex; that is,

$$\{v_0, \ldots, \widehat{v_j}, \ldots, v_k\} = \{v_0, v_1, \ldots, v_{j-1}, v_{j+1}, \ldots, v_{k-1}, v_k\}.$$

The linear map ∂_k also has a convenient matrix form. Let us pick an *ordering* of the standard basis (4.2) for C_k; that is, we write the basis as

$$(\phi_{\sigma_1}, \phi_{\sigma_2}, \ldots, \phi_{\sigma_{n_k}}),$$

where the σ_j form a complete list of the k-dimensional simplices of X. Likewise, we pick an ordering for the standard basis for C_{k-1}:

$$(\phi_{\tau_1}, \phi_{\tau_2}, \ldots, \phi_{\tau_{n_{k-1}}}),$$

where the τ_i form a complete list of the $(k-1)$-simplices of X. Then the matrix form of ∂_k (with respect to these bases) has a 1 for its (i,j) entry if the ith $(k-1)$-simplex is a face of the jth k-simplex. Otherwise its (i,j) entry is 0.

Main Property of ∂_k

The boundary maps ∂_k have a very important property:

Theorem 4.4.2. *For every k,*

$$\partial_k \circ \partial_{k+1} : C_{k+1} \to C_{k-1}$$

is the zero map.

For brevity, this theorem is frequently stated as simply $\partial^2 = 0$. Moreover, the theorem can be stated more geometrically as "the boundary of a boundary is empty". This is intuitively clear, but requires a formal proof to check our intuition.

Proof. Let $\sigma = \{v_0, v_1, \ldots, v_{k+1}\}$ be a $(k+1)$-simplex of X. Then

$$\partial_{k+1}\phi_\sigma = \sum_{j=0}^{k+1} \phi_{\{v_0,\ldots,\widehat{v_j},\ldots,v_{k+1}\}}.$$

By linearity,

$$\begin{aligned}\partial_k \circ \partial_{k+1}\phi_\sigma &= \sum_{j=0}^{k+1} \partial_k \phi_{\{v_0,\ldots,\widehat{v_j},\ldots,v_{k+1}\}} \\ &= \sum_{j=0}^{k+1} \sum_{i \neq j} \phi_{\{v_0,\ldots,\widehat{v_i},\ldots,\widehat{v_j},\ldots,v_{k+1}\}}.\end{aligned}$$

Expanding this sum, we see that each vector $\phi_{\{v_0,\ldots,\widehat{v_i},\ldots,\widehat{v_j},\ldots,v_{k+1}\}}$ appears exactly twice. Since we are working over the field F_2, this means that $\partial_k \circ \partial_{k+1}\phi_\sigma$ is zero. This shows that the claim holds on arbitrary basis vectors and it follows that it holds in general. $\square$

Cycles and Boundaries

There are a pair of interesting subspaces associated to each linear map ∂_k. Let $Z_k(X)$ denote the kernel of ∂_k. As in the case of chain groups, we will shorten the notation to Z_k when the simplicial complex $\mathcal{X}$ is understood. We refer to elements of Z_k as *k-cycles*. Let $B_k(X) = B_k$ denote the image of ∂_{k+1} (note the shift in index!). Elements of B_k are called *k-boundaries*. We have the following immediate corollary of Theorem 4.4.2.

Corollary 4.4.3. *For all k, $B_k \subset Z_k$.*

Proof. Let $\phi \in B_k$. Then, by definition, $\phi = \partial_{k+1}\psi$ for some $\psi \in C_{k+1}$. Theorem 4.4.2 implies that $\partial_k \circ \partial_{k+1}\psi$ is zero; i.e. $\partial_k\phi = 0$ and $\phi \in Z_k$. $\square$

Homology Groups

We finally define the main objects of interest for this section. The *k-th homology group of $\mathcal{X}$* is the quotient space

$$H_k(X) = Z_k(X)/B_k(X).$$

Note that this is well-defined by Corollary 4.4.3.

Remark 4.4.4. *As we remarked about the chain groups $C_k(X)$, each $H_k(X)$ is actually a vector space over F_2. We use the terminology "homology group" in order to agree with the literature. In more general homology theories, the homology groups have a richer algebraic structure—see Section 4.5.*

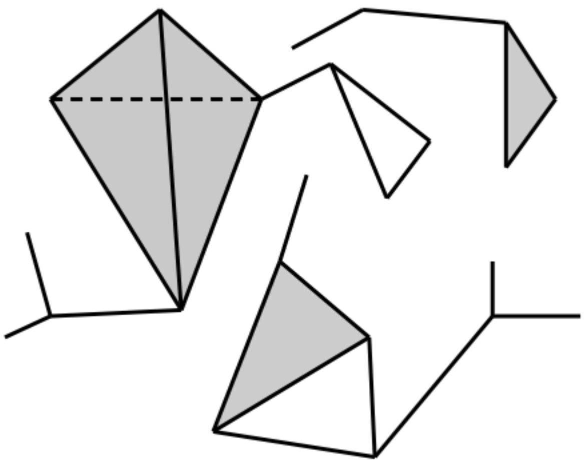
Betti Numbers

The *k-th Betti number of $\mathcal{X}$* is the integer

$$\beta_k(\mathcal{X}) := \dim H_k(X).$$

Intuitively, the k-th Betti number of $\mathcal{X}$ counts the number of k-dimensional holes in the topological space $\mathcal{X}$.

Example 4.4.1. Consider the simplicial complex shown below, where the tetrahedron is not filled in by a 3-simplex. The Betti numbers for the figure below are: $\beta_0 = 3$, indicating that it has 3 connected components; $\beta_1 = 2$, represented by the two empty triangles; $\beta_2 = 1$, represented by the empty tetrahedron. These Betti numbers were computed "by inspection", but we could also compute the homology groups explicitly to arrive at the same conclusion.



4.5 More Advanced Topics in Homology

Homology theory is a fundamental part of the subfield of topology called *algebraic topology.* As such, it can be developed in many incredibly sophisticated ways. In this section we mention some variants of the homology theory discussed in these notes and some more advanced properties of homology. A rigorous treatment of these topics is beyond the scope of these notes, but an awareness of them will be useful for the reader who wishes to study the topic further. We will therefore informally discuss some directions one can take in further research into homology theory. A standard reference for learning more advanced algebraic topology is [2].

4.5.1 Variants of Homology

Homology with Other Coefficients

A very natural way to generalize our construction of homology groups is to start by taking our chain groups to be free vector spaces over different fields. An argument could be made that it would be more natural to start by defining homology using chain groups over $\mathbb{R}$, since the majority of readers are likely to be most comfortable with real vector spaces. Indeed, given an abstract simplicial complex X, we can define chain groups

$$C_k(X;\mathbb{R}) = V_{\mathbb{R}}(\{k\text{-dimensional simplices of } X\}).$$

We can then adjust our definitions to get *homology with real coefficients*. The main technical drawback is that we would then need to introduce a more complicated boundary map which takes into account an “orientation” on each simplex of X. Our motivation for working over F_2 was precisely to avoid these technicalities!

More generally, we could define chain groups to be free vector spaces over more general fields, such as the field with three elements F_3 (for any prime number p, there is a corresponding field with p elements F_p). Homology over other finite fields can be useful for distinguishing spaces, and is actually built into several of the available persistence homology programs. Even more generally, one can define homology by starting with chain groups defined as modules over some other group or ring, which are algebraic structures that generalize vector spaces and fields.

Homology of Topological Spaces

In this chapter we defined homology of simplicial complexes with F_2-coefficients. The previous subsection indicates that we could define homology theories of simplicial complexes with more general coefficients by some mild adjustments to our definitions. There is another natural question: can we extend homology to treat more general topological spaces? The answer to the question is “yes”, but it turns out to be not so straightforward to do so.

Let X be a topological space (or a metric space, if you prefer). A *triangulation* of X is a geometric simplicial complex $\mathcal{X}$ such that X and $\mathcal{X}$ are homeomorphic ($\mathcal{X} \subset \mathbb{R}^n$ endowed with the subspace topology). One could then calculate the homology of the simplicial complex $\mathcal{X}$ using our definition. Unfortunately, some obvious issue arise immediately:

- Triangulations are highly non-unique. Is it possible that two triangulations $\mathcal{X}$ and $\mathcal{X}'$ of the space X have different homology groups? (See Figure 4.3.)

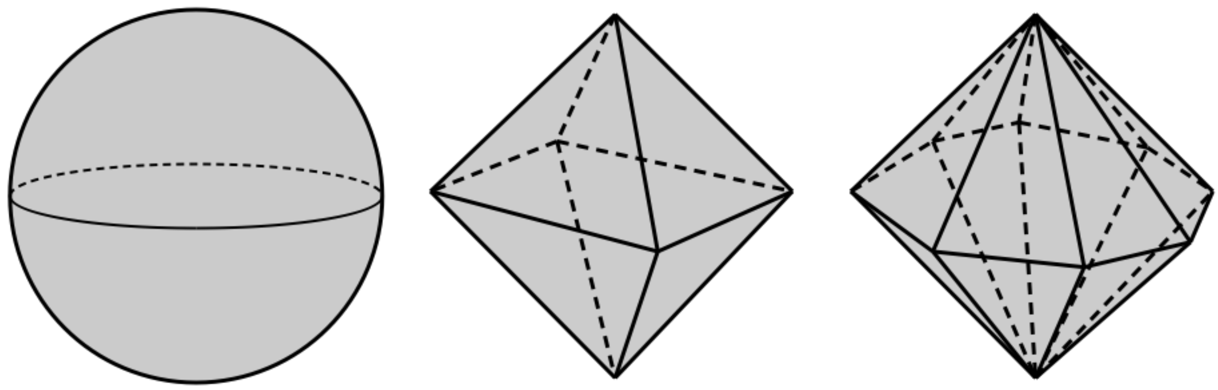


Figure 4.3: A 2-sphere $S^2 \subset \mathbb{R}^3$ and a pair of distinct triangulations.

- Does every space X admit a triangulation? If not, then this strategy will not allow us to calculate the homology of a general space.

The second question has been a subject of intense study in pure mathematics for decades. One obvious obstruction is that our definition of a simplicial complex requires that every simplicial complex is compact, hence any noncompact space cannot be triangulated. It was recently proved [5] that there even exist relatively simple *compact* topological spaces (compact 5-dimensional manifolds, which are 5-dimensional analogues of 2-dimensional surfaces such as spheres and donuts) that have no triangulation. To treat general topological spaces, we therefore need a more flexible version of homology. One such theory is called *singular homology* . The idea is to form the k-th chain group of a topological space X as

$$C^{sing}(X;\mathbb{F}) = V_{\mathbb{F}}(\{\text{continuous maps of the standard } k\text{-simplex into } X\}).$$

Boundary maps and a homology theory can be defined from there. Singular homology has the important property that if two spaces are homeomorphic, then their singular homology groups must agree (this is called *functoriality*, and is treated in the next section). It is known that for a simplicial complex, the singular homology groups and simplicial homology groups (as we have defined in this chapter) are the same, and it follows that if $\mathcal{X}$ and $\mathcal{X}'$ are different triangulations of the same space X, then all spaces will have the same homology groups.

The problem with using this singular homology approach for applications is that the singular chain groups are infinite-dimensional and therefore impossible to work with directly. There are many sophisticated tools used to treat them abstractly, but in practice one must always convert a space into a finite simplicial complex in order to do direct calculations. The discussion in this section shows that, if any such triangulation exists, then it doesn't matter which triangulation we choose for a space—the resulting homology will always be the same!

4.5.2 Functoriality

Consider simplicial complexes $\mathcal{X}$ and $\mathcal{Y}$ with abstract simplicial complexes X and Y. Let $f : V(X) \to V(Y)$ be a map of abstract simplicial complexes. This map induces a well-defined map $C_k(f) : C_k(X) \to C_k(Y)$ of chain groups (over F_2) by defining

$$C_k(f)(\phi_\sigma) = \begin{cases} \phi_{f(\sigma)} & \text{if } \phi_{f(\sigma)} \in C_k(Y) \\ 0 & \text{otherwise} \end{cases}$$

for each k-simplex $\sigma \in C_k(X)$ and extending linearly. Recall that a map of simplicial complexes is not required to be injective, so it is possible that $\phi_{f(\sigma)}$ is a lower-dimensional simplex—this is the reason for the conditional definition of $C_k(f)$.

Theorem 4.5.1. *The maps $C_k(f) : C_k(X) \to C_k(Y)$ induce well-defined linear maps $H_k(f) : H_k(X) \to H_k(Y)$ on homology vector spaces.*

Proof. We first claim that the following diagram commutes.

$$\begin{array}{ccc} C_k(X) & \xrightarrow{\partial_k} & C_{k-1}(X) \\ {\scriptstyle C_k(f)}\downarrow & & \downarrow{\scriptstyle C_{k-1}(f)} \\ C_k(Y) & \xrightarrow{\partial_k} & C_{k-1}(Y) \end{array}$$

This means that if we start in the upper left corner and proceed to the lower right through either of the two possible paths, the resulting map is the same. To check this, let $\phi_\sigma \in C_k(X)$ be a basis element. We will assume that f takes the k-simplex σ to a k-simplex $f(\sigma)$—the case in which $f(\sigma)$ is a lower-dimensional simplex follows similarly. Then

$$\partial_k\left(C_k(f)(\phi_\sigma)\right) = \partial_k \phi_{f(\sigma)} = \sum \{\phi_\xi \mid \xi \text{ is a } (k-1)\text{-face of } f(\sigma)\}\,. \tag{4.3}$$

On the other hand,

$$\begin{aligned} C_{k-1}(f)(\partial_k(\sigma)) &= C_{k-1}(f)\left(\sum\{\phi_\tau \mid \tau \text{ is a } (k-1)\text{-face of } \sigma\}\right) \\ &= \sum\{\phi_{f(\tau)} \mid \tau \text{ is a } (k-1)\text{-face of } \sigma\}. \end{aligned} \tag{4.4}$$

Next note that the assumption that $f(\sigma)$ is a k-simplex implies that f is injective on the vertices of σ. It follows immediately that the expressions (4.3) and (4.4) are equal.

The fact that the diagram commutes implies that $C_k(f)$ takes $Z_k(X)$ into $Z_k(Y)$ and $B_k(X)$ into $B_k(Y)$. Therefore $C_k(f)$ induces a well-defined linear map $H_k(f)$ from the quotient space $H_k(X) = Z_k(X)/B_k(X)$ into the quotient space $H_k(Y) = Z_k(Y)/B_k(Y)$. □

The property described by the theorem—that maps between spaces induce maps between homology vector spaces—is an example of a general mathematical principle called *functoriality*. A much more general property is enjoyed by singular homology groups. We state it here without proof because the result is very useful, although a bit too technical to prove for our purposes.

Theorem 4.5.2. *If topological spaces X and Y are homotopy equivalent, then their singular homology vector spaces $H_k^{sing}(X; F_2)$ and $H_k^{sing}(Y; F_2)$ are isomorphic.*

Although we haven't proved the theorem, we will use the following corollary.

Corollary 4.5.3. *Let X and Y be topological spaces with triangulations $\mathcal{X}$ and $\mathcal{Y}$. If the simplicial homology groups of $\mathcal{X}$ and $\mathcal{Y}$ are not all isomorphic, then X and Y are not homotopy equivalent.*

4.6 Exercises

1. Let $S \subset \mathbb{R}^k$ be a set. Show that $\mathrm{cvx}(S)$ is equal to the set
$$\bigcup \{(1-t)x + ty \mid t \in [0,1],\ x, y \in S\}.$$
2. Write down the abstract simplicial complex associated to the geometric simplicial complex shown below.
3. Prove Proposition 4.3.3.
4. For each of the simplicial complexes shown in Figure 4.4, write down matrix expressions for ∂_1 and ∂_2 with respect to natural choices of bases. Compute the images and kernels of each map (feel free to use a computer algebra system to do so). Calculate the dimension of $\ker(\partial_1)/\mathrm{image}(\partial_2)$ in each case. Does your answer make sense intuitively?
5. Let $\mathcal{X}$ be a simplicial complex and let $X = (V(X), \Sigma(X))$ be as defined in Example 4.2.3. Show that X defines an abstract simplicial complex.
6. Let $\mathcal{X}$ be the standard n-simplex defined in Example 4.2.1 and let X be the abstract simplex associated to $\mathcal{X}$ (see Example 4.2.3). Show that X is simplicially isomorphic to the standard abstract n-simplex defined in Example 4.2.4.
7. Prove that the spaces from Example 4.1.4 are not homeomorphic. Here is a suggested strategy: find a triangulation of each space. Compute the homology vector spaces for each triangulation and show that they are not all the same. Using functoriality (in particular, Corollary 4.5.3), conclude that the spaces are not homotopy equivalent, hence not homeomorphic.

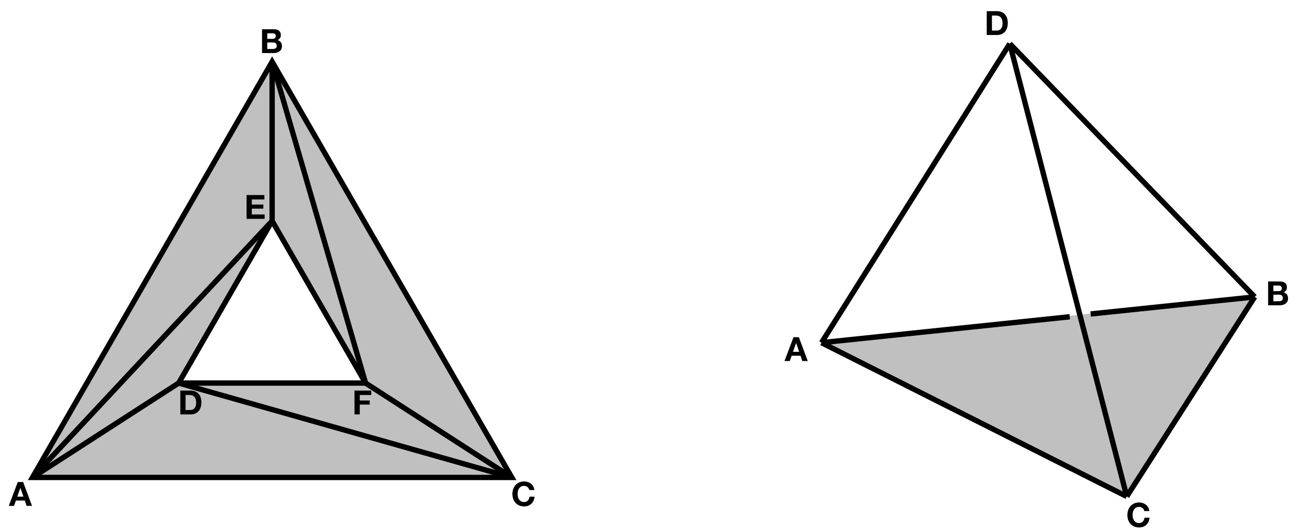


Figure 4.4: The space on the left is a simplicial complex in $\mathbb{R}^2$. The space on the right is a simplicial complex in $\mathbb{R}^3$. It has the 1-skeleton of a 3-simplex, with a single face $\{A, B, C\}$ filled in.

5 Persistent Homology

As we stated in Chapter 3, a typical real-world dataset comes in the form of a *point cloud*—that is, a finite subset of some ambient metric space. Every point cloud determines a *finite metric space* (X, d) by taking d to be the restriction of the metric from the ambient space. Our goal is therefore to study the topology of finite metric spaces. But here we see a problem: finite metric spaces are classified up to homeomorphism by the number of points in the space, and the number of datapoints in a large dataset is not a very interesting invariant. On the other hand, we can intuitively distinguish between topological types of finite metric spaces, as in the example shown in Figure 5.1. The point clouds contain the same number of points, so they are homeomorphic. But the point cloud on the left appears intuitively to be unstructured, while the point cloud on the right appears to have a topological feature (a "hole", or a 1-dimensional homology cycle!)

So the question becomes, how do we algorithmically encode the apparent topological differences between these finite metric spaces? The approach that we will take is through persistent homology of their associated Vietoris-Rips complexes. The rough idea is as follows. Let (X, d) be a finite metric space. For each distance parameter r, we associate to X a simplicial complex $\mathrm{VR}(X, r)$ defined in terms of d. We can then calculate the simplicial homology of this complex. The homology vector spaces change with the parameter r, and those homology classes which survive for a long interval of r values (i.e., those which "persist") are deemed topologically relevant, while homology classes that appear and quickly disappear are treated as noise. The goal of this chapter is to fill in the details of this process.

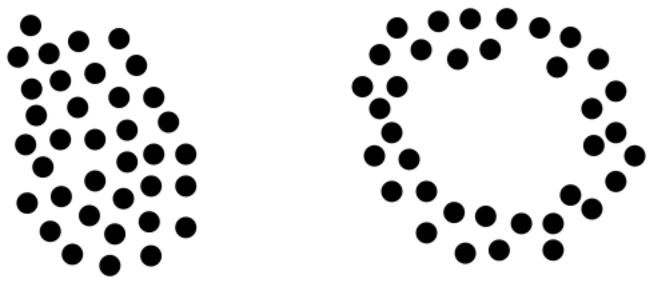


Figure 5.1: A pair of point clouds with the same number of points.

5.1 Vietoris-Rips Complex

5.1.1 Definition

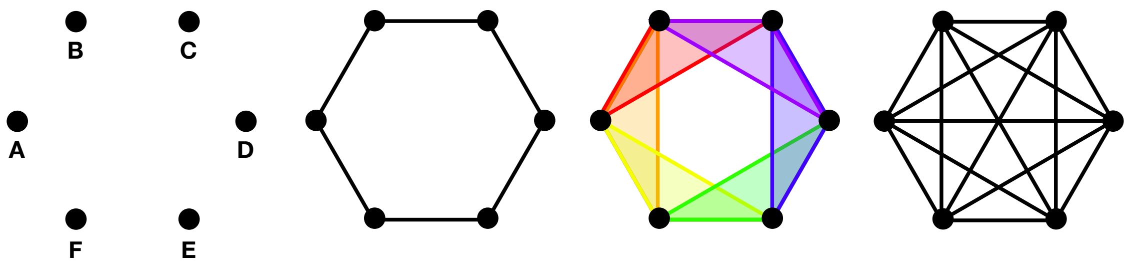
Let (X, d) be a finite metric space. For each real number $r \geqslant 0$, we can associate a simplicial complex to X called the *Vietoris-Rips complex* and denoted $\mathrm{VR}(X, r)$ as follows. The vertex set of $\mathrm{VR}(X, r)$ is simply the set X. A subset $\{x_0, x_1, \ldots, x_n\}$ of X is declared to be a simplex of $\mathrm{VR}(X, r)$ if and only if

$$d(x_i, x_j) \leqslant r \quad \forall\, i, j \in \{0, 1, \ldots, n\}.$$

5.1.2 A Simple Example

Clearly, the definition of $\mathrm{VR}(X, r)$ depends heavily on the choice of r. Let's look at a simple example for various choices of r.

Example 5.1.1. Consider the metric space X consisting of 6 points forming the vertices of a regular hexagon of side length 1 in the Euclidean plane. We label the points of X as $A - F$. These points form the vertex set of $\mathrm{VR}(X, r)$ for any choice of $r \geqslant 0$. The Vietoris-Rips complexes of X can be grouped into four categories.



1. For $0 \leqslant r < 1$, $\mathrm{VR}(X, r)$ is just the set of discrete vertices.

2. For $1 \leqslant r < \sqrt{3}$, the Vietoris-Rips complex can be visualized as the simplicial complex pictured in the figure second from the left. For each pair of consective vertices, there is an edge in $\mathrm{VR}(X, r)$, because consective vertices are at distance 1 from each other. Non-consecutive vertices are at distance at least $\sqrt{3}$ from one-another, so there are no other simplices in $\mathrm{VR}(X, r)$.

3. For $\sqrt{3} \leqslant r < 2$, each triple of consecutive vertices forms a 2-simplex in $\mathrm{VR}(X, r)$. For example, the elements of the set $\{A, B, C\}$ satisfy

   $$d(A, B) = 1 \leqslant \sqrt{3}, \quad d(B, C) = 1 \leqslant \sqrt{3}, \quad d(A, C) = \sqrt{3}.$$

   Any triple of vertices which are not consecutive will contain a pair of vertices which are 2 units apart, so there are no other 2-simplices in $\mathrm{VR}(X, r)$. Moreover, any set containing 4 points will contain a pair of vertices which are 2 units apart, so there are no higher-dimensional simplices in $\mathrm{VR}(X, r)$. The figure second from the right shows the 2-dimensional simplices in $\mathrm{VR}(X, r)$. Note that the figure is not actually a simplicial complex, since the simplices aren't attached to each other in the correct way! It will typically be best to think of $\mathrm{VR}(X, r)$ as an abstract simplicial complex. Of course, $\mathrm{VR}(X, r)$ has a geometric realization, but it will typically be difficult or impossible to visualize.

4. For $r \geqslant 2$, every set of vertices is included in $\mathrm{VR}(X, r)$—this is simply because the greatest distance between *any* two points in X is 2. This means that $\mathrm{VR}(X, r)$ contains 3, 4 and 5-dimensional simplices. The figure on the right shows the set of edges in $\mathrm{VR}(X, r)$ (i.e., there is an edge joining any two vertices).

5.1.3 Observations

From Example 5.1.1, we can immediately make some observations about $\mathrm{VR}(X, r)$ for any finite metric space (X, d). The following propositions are clear.

Proposition 5.1.1. *For $r \leqslant \min\{d(x, x') \mid x, x' \in X\}$, $\mathrm{VR}(X, r)$ is homeomorphic to X.*

Let S be a finite set. The *complete simplicial complex on S* is the simplicial complex containing a simplex for every subset of S.

Proposition 5.1.2. *For $r \geqslant \max\{d(x, x') \mid x, x' \in X\}$, $\mathrm{VR}(X, r)$ is the complete simplicial complex on X.*

Another important phenomenon illustrated by Example 5.1.1 is that the topology of $\mathrm{VR}(X, r)$ changes with r according to the geometry of X. In the example, we see that a loop is formed when $r = 1$, that the loop "persists" as 2-dimensional simplices are attached when $1 \geqslant r < 2$, and the loop finally dissappears when $r = 2$ as it is filled in by higher-dimensional simplices. This is the key idea of persistent homology, which we will define below.

5.1.4 Other Complexes

There are a variety of other ways to associate a simplicial complex to a finite metric space, for example, the α-complex is defined for a finite metric space (X, d) which is isometrically embedded in some larger metric space (Y, d)—for simplicity, assume that (Y, d) is $\mathbb{R}^N$ with Euclidean distance. For each $x \in X$, the *Voronoi cell* of x is the set

$$\mathrm{Vor}(x) = \{y \in Y \mid d(x, y) \leqslant d(x', y) \text{ for all } x' \in X\}.$$

For each $r > 0$, we define the *α-cell*

$$A(x, r) = \overline{B_d(x, r)} \cap \mathrm{Vor}(x).$$

The *α-complex* of X with scale parameter r is the simplicial complex with vertex set X and k-faces consisting of sets $\{x_0, \ldots, x_k\}$ such that

$$\bigcap_{j=1}^{k} A(x_j, r) \neq \varnothing.$$

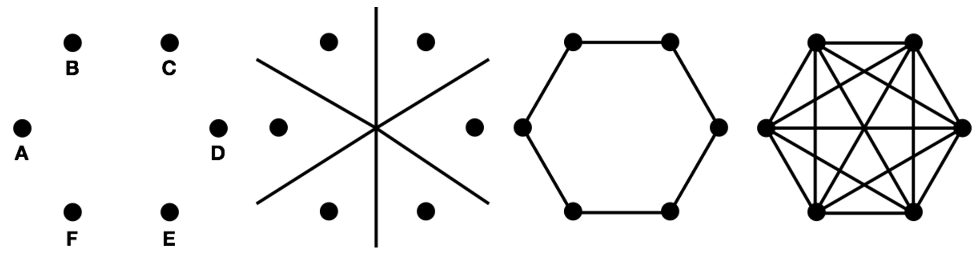


Figure 5.2: A point cloud, its Voronoi decomposition and its α-complexes at scale parameters $r = 3/4$ and $r = 5/4$.

An example is shown in Figure 5.2. The α-complex is typically smaller than the Vietoris-Rips complex in that it contains fewer simplices. However, the algorithms to compute it require one to compute the Voronoi cells of the ambient space, which are computationally expensive when the ambient space is high-dimensional.

Some other parameterized complexes that can be associated to a finite metric space are the Cĕch complex and the witness complex. See [1] for more details.

5.2 Linear Algebra with Persistence Vector Spaces

Inspired by the previous section, and the Vietoris-Rips complex construction, we wish to introduce a version of simplicial homology which is parameterized over r. All of the concepts used in linear algebra generalize to give parameterized versions.

5.2.1 Definitions

Persistence Vector Space

Let $\mathbb{F}$ be a field (we will mainly be using $\mathbb{F} = F_2$, so it is okay to just keep this choice in mind). A *persistence vector space over* $\mathbb{F}$ is a family $\{V_r\}_{r\in\mathbb{R}_+}$ of vector spaces V_r over $\mathbb{F}$ together with a family of linear maps $L_{r,r'} : V_r \to V_{r'}$ for $r \leqslant r'$. Moreover, we require the linear maps to satisfy the following compatibility condition: if $r \leqslant r' \leqslant r''$, then $L_{r,r''} = L_{r',r''} \circ L_{r,r'}$.

We will denote a persistence vector space by $\{V_r\}$, with the understanding that persistence vector spaces are always parameterized over $\mathbb{R}_+$. When talking about multiply persistence vector spaces, we will need to distinguish their families of linear maps. In this case, we will use the notation $L^V_{r,r'}$ for the maps associated to $\{V_r\}$.

Linear Transformation Between Persistence Vector Spaces

Let $\{V_r\}$ and $\{W_r\}$ be persistence vector spaces over $\mathbb{F}$. A *linear transformation of persistence vector spaces* is a family $\{f_r\}$ of linear maps $f_r : V_r \to W_r$ which preserves the structure of the maps $L_{v,v'}$. That is, for all $r \leqslant r'$ the following diagram commutes:

$$\begin{array}{ccc} V_r & \xrightarrow{L^V_{r,r'}} & V_{r'} \\ {\scriptstyle f_r}\downarrow & & \downarrow{\scriptstyle f_{r'}} \\ W_r & \xrightarrow{L^W_{r,r'}} & W_{r'} \end{array}$$

To say that the diagram commutes means that either one of the possible paths from V_r to $W_{r'}$ yields the same result. More precisely,

$$f_{r'} \circ L^V_{r,r'} = L^W_{r,r'} \circ f_r.$$

A linear transformation of persistence vector spaces is called an *isomorphism* if it admits a two-sided inverse.

Sub-Persistence Vector Space

A *sub-persistence vector space* is a collection $\{U_r\}$ of linear subspaces $U_r \subset V_r$ such that $L^V_{r,r'}(U_r) \subset U_{r'}$ holds for each $r \leqslant r'$.

Let $f = \{f_r : V_r \to W_r\}$ be a linear map of persistence vector spaces. The *kernel* of f is the sub-persistence vector space

$$\ker(f) = \{\ker(f_r)\}$$

and the *image* of f is the sub-persistence vector space

$$\mathrm{im}(f) = \{\mathrm{im}(f_r)\}.$$

Quotient Persistence Vector Space

Let $\{U_r\}$ be a sub-persistence vector space of $\{V_r\}$. The *quotient persistence vector space* is the persistence vector space $\{V_r/U_r\}$ where the linear maps $L^{V/U}_{r,r'}$ are given for $v \in V_r$ by the formula

$$L^{V/U}_{r,r'}([v]) = \left[L^V_{r,r'}(v)\right].$$

Direct Sum of Persistence Vector Spaces

For persistence vector spaces $\{V_r\}$ and $\{W_r\}$ over $\mathbb{F}$, we define the *direct sum* $\{V_r\} \oplus \{W_r\}$ to be the persistence vector space with

$$\left(\{V_r\} \oplus \{W_r\}\right)_r = V_r \oplus W_r$$

and linear maps

$$L^{V \oplus W}_{r,r'} : V_r \oplus W_r \to V_{r'} \oplus W_{r'}$$

given by

$$L^{V \oplus W}_{r,r'}(v, w) = \left(L^V_{r,r'}(v), L^W_{r,r'}(w)\right).$$

Free Persistence Vector Space

A *filtered set* is a set X together with a map $\rho : X \to \mathbb{R}_+$. For each $r \in \mathbb{R}_+$, we define the *sublevel set*

$$X[r] = \{x \in X \mid \rho(x) \leqslant r\}.$$

The *free vector space generated by* (X, ρ) is the persistence vector space denoted $\{V_{\mathbb{F}}(X, \rho)_r\}$ and defined by

$$V_{\mathbb{F}}(X, \rho)_r = V_{\mathbb{F}}(X[r]).$$

Note that for each pair $r \leqslant r'$, $X[r] \subset X[r']$ so that we can define the linear maps $L^{V_{\mathbb{F}}(X)}_{r,r'}$ can be defined by the formula

$$L^{V_{\mathbb{F}}(X,\rho)}_{r,r'}(\phi_x) = \phi_x$$

for each basis vector $\phi_x \in V_{\mathbb{F}}(X, \rho)_r$.

A general persistence vector space $\{V_r\}$ is called *free* if it can be expressed as $\{V_{\mathbb{F}}(X, \rho)_r\}$ for some filtered set (X, ρ). It is called *finitely generated* if it is free and the filtered set (X, ρ) can be chosen so that X is finite.

Example 5.2.1. Our main examples of free persistence vector spaces are *persistence chain complexes*. Given a finite metric space (X, d), we construct its Vietoris-Rips complex $\mathrm{VR}(X, r)$ for each scale parameter r. We then obtain a filtered set (X, ρ), where finish this

5.2.2 Matrix Representations of Linear Maps of Persistence Vector Spaces

Recall that choosing bases for finite-dimensional vector spaces V and W allows us to represent an abstract linear transformation $L : V \to W$ as matrix multiplication. Our goal is to define a similar representation for linear maps between persistence vector spaces. As one would expect, the representation is a bit more involved. For this section, we will restrict our attention to finitely generated persistence vector spaces.

(X, Y)-Matrices

Let (X, Y) be a pair of finite sets. An (X, Y)*-matrix over* $\mathbb{F}$ is a size $|X| \times |Y|$ matrix over $\mathbb{F}$ whose entries are indexed by the sets X and Y rather than integers. That is, it is a matrix (a_{xy}) of entries $a_{xy} \in \mathbb{F}$. For each $x \in X$, we use $r(x)$ to denote the row associated to x. Similarly, $c(y)$ denotes the column associated to $y \in Y$.

Note that if X and Y are ordered, then the ordering turns an (X, Y)-matrix into a matrix in the usual sense.

The Matrix Associated to a Linear Map

Let (X, ρ) and (Y, σ) be finite filtered sets and let

$$f : \{V_{\mathbb{F}}(Y, \sigma)_r\} \to \{V_{\mathbb{F}}(X, \rho)_r\}$$

be a linear transformation of the associated persistence vector spaces. Since X and Y are finite, it must be the case that there exists some large real number R such that $r \geqslant R$ implies $V_{\mathbb{F}}(X,\rho)_r = V_{\mathbb{F}}(X)$ and $V_{\mathbb{F}}(Y,\sigma)_r = V_{\mathbb{F}}(Y)$. Moreover, if both $r, r' \geqslant R$ then the linear maps f_r and $f_{r'}$ are the *same map* on their common domain $V_{\mathbb{F}}(Y,\sigma)_r = V_{\mathbb{F}}(Y,\sigma)_{r'} = V_{\mathbb{F}}(Y)$. When $r \geqslant R$, we denote this common map by

$$f_\infty : V_{\mathbb{F}}(Y) \to V_{\mathbb{F}}(X).$$

Using the standard bases $\{\phi_x\}_{x\in X}$ and $\{\phi_y\}_{y\in Y}$ for $V_{\mathbb{F}}(X)$ and $V_{\mathbb{F}}(Y)$, respectively, there is an (X,Y)-matrix associated to f_∞, which we denote $A(f) = (a_{xy})$. We once again remark that, in order to actually think of (a_{xy}) as a matrix in the usual sense, we need to pick an ordering of the bases. This will not be necessary in what follows.

We have a representation result which says that $A(f)$ contains all of the information carried by f. To prove it, we will need the following simple lemma.

Lemma 5.2.1. *A linear combination $\sum_{x\in X} a_x\phi_x \in V_{\mathbb{F}}(X)$ lies in $V_{\mathbb{F}}(X,\rho)_r$ if and only if $a_x = 0$ whenever $\rho(x) > r$.*

Proof. By definition, the vector space $V_{\mathbb{F}}(X,\rho)_r$ consists exactly of linear combinations of ϕ_x with $x \in X[r]$; i.e. $\phi(x) \leqslant r$. A linear combination $\sum a_x\phi_x$ is of this form if and only if $a_x = 0$ when $x \notin X[r]$; i.e. when $\phi(x) > r$. $\square$

An (X,Y)-matrix (a_{xy}) is called (ρ,σ)-*adapted* if it has the property that $a_{xy} = 0$ whenever $\rho(x) > \sigma(y)$.

Proposition 5.2.2. *The (X,Y)-matrix $A(f)$ is (ρ,σ)-adapted and any (ρ,σ)-adapted matrix A uniquely determines a linear transformation of perisistence vector spaces*

$$f_A : \{V_{\mathbb{F}}(Y,\sigma)_r\} \to \{V_{\mathbb{F}}(X,\rho)_r\}$$

such that the correspondences $f \mapsto A(f)$ and $A \mapsto f_A$ are inverses of each other.

Proof. The image of the basis vector $\phi_y \in V_{\mathbb{F}}(Y)$ is the linear combination $\sum_{x\in X} a_{xy}\phi_x$, by definition of $A(f)$. On the other hand, the structure of f implies that the image of ϕ_y lies in $V_{\mathbb{F}}(X,\rho)_{\sigma(y)}$, so $A(f)$ is (ρ,σ)-adapted by Lemma 5.2.1.

Now let $A = (a_{xy})$ be an arbitrary (ρ,σ)-adapted matrix. We define a linear map of persistence vector spaces by restriction. That is, a basis element $\phi_y \in V_{\mathbb{F}}(Y,\sigma)_r$ is mapped to $\sum_{x\in X} a_{xy}\phi_x$. The assumption that A is (ρ,σ)-adapted guarantees that the image of this map lies in $V_{\mathbb{F}}(X,\rho)$, by Lemma 5.2.1. $\square$

5.3 Persistence Homology

A *filtered simplicial complex* is a family $\{\mathcal{X}_r\}$ of simplicial complexes indexed by $\mathbb{R}$, such that for each $r \leqslant r'$, $\mathcal{X}_r$ is a subcomplex of $\mathcal{X}_{r'}$. Let $\{X_r\}$ denote the associated family of abstract simplicial complexes.

Example 5.3.1. Our main examples are the simplicial complexes constructed for point clouds in Section **??**. For example, the collection of Vietoris-Rips complexes $\{\mathrm{VR}(X, r)\}$ defines a filtered simplicial complex.

The *k-th peristence homology vector space* of a filtered simplicial complex is the persistence vector space $\{PH_k(X)_r\}$ with $PH_k(X)_r$ the k-th simplicial homology of the simplicial complex $\{\mathcal{X}_r\}$. This does determine a persistence vector space: for $r \leqslant r'$, the map $L_{r,r'}^{PH_k(X)}$ is the map induced on homology by the inclusion map $X_r \hookrightarrow X_{r'}$.

5.4 Examples

5.4.1 Example 1

5.4.2 Example 2

5.5 Exercises

1. Consider the point cloud formed by the 14 points are at the vertices of the cube of side length $\sqrt{2}$ (this value is chosen to make calculations a bit simpler),

   $$(0,0,0),(0,0,\sqrt{2}),(\sqrt{2},0,0),(\sqrt{2},0,\sqrt{2}),(0,\sqrt{2},0),(0,\sqrt{2},\sqrt{2}),(\sqrt{2},\sqrt{2},0),(\sqrt{2},\sqrt{2},\sqrt{2})$$

   and the midpoints of the faces of the cube

   $$(\sqrt{2}/2,\sqrt{2}/2,0),(\sqrt{2}/2,0,\sqrt{2}/2),(0,\sqrt{2}/2,\sqrt{2}/2),$$
   $$(\sqrt{2}/2,\sqrt{2}/2,\sqrt{2}),(\sqrt{2}/2,\sqrt{2},\sqrt{2}/2),(\sqrt{2},\sqrt{2}/2,\sqrt{2}/2).$$

   Work out the persistent homology of this point cloud.

6 Representions of Persistent Homology

6.1 Structure Theorem for Persistence Vector Spaces

We have so far developed a way to assign a topological invariant to a point cloud: first turn the point cloud into a filtered simplicial complex (via, say, the Vietoris-Rips construction), then compute its persistent homology. The resulting topological invariant initially seems quite complex; it is, after all, a collection of infinitely many vector spaces. The goal of this section is to show that the persistence vector spaces arising from the persistent homology construction can actually be described in a simple way.

6.1.1 Preliminaries

Before stating the main result, we need to introduce some notation.

Birth-Death Persistence Vector Spaces

Let $b \in [0, \infty)$ and $d \in [0, \infty]$ be real numbers with $b < d$. Let $P_{\mathbb{F}}(b, d)$ denote the persistence vector space with

$$P_{\mathbb{F}}(b,d)_r = \begin{cases} \mathbb{F} & r \in [b,d) \\ 0 & r \notin [b,d), \end{cases}$$

where 0 denotes the vector space containing only 0. The linear maps are defined by

$$L_{r,r'}^{P_{\mathbb{F}}(b,d)} = \begin{cases} 0 & r < b \\ 1 & b \leqslant r \leqslant r' < d \\ 0 & r' \geqslant d, \end{cases}$$

where 0 is shorthand for the zero map and 1 is shorthand for the identity map $\mathbb{F} \to \mathbb{F}$.

We refer to these special vector spaces as *birth-death persistence vector spaces* . The numbers (b, d) are called a *birth-death pair* .

The following example suggests that birth-death persistence vector spaces play a special role when dealing with free persistence vector spaces on finite filtered sets.

Example 6.1.1. Let (X, ρ) be a finite filtered set with $X = \{x_1, \ldots, x_n\}$ and consider the persistence vector space $\{V_{\mathbb{F}}(X, \rho)_r\}$. We define a map

$$f : \{V_{\mathbb{F}}(X,\rho)_r\} \to \bigoplus_j P_{\mathbb{F}}(\rho(x_j), +\infty)$$

as follows. For fixed r, note that $V_{\mathbb{F}}(X,\rho)_r = V_{\mathbb{F}}(X[r])$ and

$$\left(\bigoplus_j P_{\mathbb{F}}(\rho(x_j), +\infty)\right)_r \approx \bigoplus_{x\in X[r]} \mathbb{F}.$$

We can think of elements of the target space as $(\{1\}, X[r])$-matrices (that is, as row vectors indexed by $X[r]$). Then the map f_r is given on the canonical basis for $V_{\mathbb{F}}(X[r])$ by taking ϕ_x $(x \in X[r])$ to the $(\{1\}, X[r])$-matrix with a 1 in the position $c(x)$ and zeros elsewhere.

It is easy to see that for any $r \leqslant r'$, $f_{r'} \circ L_{r,r'}^{V_{\mathbb{F}}(X,rho)} = L_{r,r'}^{P_{\mathbb{F}}(b,d)} \circ f_r$ and that each f_r is invertible. It follows that the map f is an isomorphism of persistence vector spaces.

Birth-death persistence vector spaces are useful as "building blocks" to construct other persistence vector spaces. They have the following important uniqueness property.

Proposition 6.1.1. *Let*

$$\{V_r\} = \bigoplus_{i=1}^{n} P(b_i, d_i) \quad \textit{and} \quad \{\overline{V}_r\} = \bigoplus_{j=1}^{m} P(\overline{b}_j, \overline{d}_j).$$

If $\{V_r\} \approx \{\overline{V}_r\}$, then $n = m$ and the set of pairs (b_i, d_i) with multiplicities is equal to the set of $(\overline{b}_j, \overline{d}_j)$ with multiplicities.

We state the proposition more succinctly as: direct sums of decompositions into birth-death pairs are unique up to reordering.

Proof. Let $b_{min} = \min\{b_i\}$ and $\overline{b}_{min} = \min\{\overline{b}_j\}$; i.e., the "first birth times". These values are given by

$$b_{min} = \min\{r \mid V_r \neq 0\} \quad \text{and} \quad \overline{b}_{min} = \min\{r \mid \overline{V}_r \neq 0\},$$

and the isomorphism $\{V_r\} \approx \{\overline{V}_r\}$ therefore implies $b_{min} = \overline{b}_{min}$ (since $V_r = 0$ if and only if $\overline{V}_r = 0$). We denote this common value by b.

Next we consider the "first death times"

$$d_{min} = \min\{d_i \mid b_i = b\} \quad \text{and} \; \overline{d}_{min} = \min\{\overline{d}_j \mid \overline{b}_j = b\}.$$

There is also an intrinsic characterization of these values. Indeed,

$$d_{min} = \min\{r' \mid \ker(L_{r,r'}) \neq 0\} \quad \text{and} \quad \overline{d}_{min} = \min\{r' \mid \ker(\overline{L}_{r,r'}) \neq 0\}.$$

Our isomorphism implies that $L_{r,r'}$ is injective if and only if $\overline{L}_{r,r'}$ is as well, so we have $d_{min} = \overline{d}_{min}$ and we denote this common value by d.

The work above shows that $P(b,d)$ appears in both decompositions at least once, say k-times in the first decomposition and $\overline{k}$ in the second. Our goal is to show that it appears the same number of times in each decomposition; that is $k = \overline{k}$. Without loss of generality, we can assume that $P(b,d)$ appears as the first k (respectively $\overline{k}$) summands of $\{V_r\}$ (respectively $\{\overline{V}_r\}$). Let

$$\{W_r\} = \bigoplus_{i=1}^{k} P(b,d) \subset \{V_r\} \quad \text{and} \quad \{\overline{W}_r\} = \bigoplus_{j=1}^{\overline{k}} P(b,d) \subset \{\overline{V}_r\},$$

where each space is a sub-persistence vector space. We once again have intrinsic characterizations; we claim that W_r and $\overline{W}_r$ are each isomorphic to the kernel of the linear transformation

$$\text{image}\,(L_{b,r}) \xrightarrow{L_{r,d}|_{\text{image}\left(L_{b,r}\right)}} V_d.$$

This shows that $\{W_r\} \approx \{\overline{W}_r\}$.

We therefore have

$$\{V_r\}/\{W_r\} \approx \{\overline{V}_r\}/\{\overline{W}_r\}.$$

Moreover,

$$\{V_r\}/\{W_r\} \approx \bigoplus_{i=k+1}^{n} P(b_i, d_i) \qquad \text{and} \qquad \{\overline{V}_r\}/\{\overline{W}_r\} \approx \bigoplus_{i=k+1}^{n} P(\overline{b}_i, \overline{d}_i),$$

and we the proof follows by induction on the number of summands in the direct sum decompositions. $\square$

Finitely Presented Persistence Vector Spaces

We will also need the following definition. A persistence vector space is called *finitely presented* if it is isomorphic to a persistence vector space of the form $\{W_r\}/\text{image}(f)$ for some linear transformation of persistence vector spaces

$$f : \{V_r\} \to \{W_r\},$$

where $\{V_r\}$ and $\{W_r\}$ are both finitely generated free persistence vector spaces.

Example 6.1.2. The main example of a finitely presented persistence vector space is a persistence homology vector space.

Let $\{X_r\}$ denote the set of abstract simplicial complexes associated to filtered simplicial complex $\{\mathcal{X}_r\}$. For each k, the collection of chain groups $\{C_k(X_r)\}$ forms a persistence vector space, with maps $L_{r,r'}^{C_k(X_r)}$ induced by inclusions. Consider the map $\partial_k : \{C_k(X_r)\} \to \{C_{k-1}(X_r)\}$ with $(\partial_k)_r$ given by the usual boundary map on $C_k(X_r)$. It is straightforward to see that for all $r \leqslant r'$, we have $(\partial_k)_{r'} \circ L_{r,r'}^{C_k(X_r)} = L_{r,r'}^{C_{k-1}(X_r)} \circ (\partial_k)_r$, so that ∂_k defines a linear map of persistence vector spaces. It follows that the collection $\{Z_k(X_r)\}$ of k-dimensional cycles also forms a persistence vector space.

Therefore $\{PH_k(X_r)\} = \{C_k(X_r)\}/\text{image}(\partial_k)$ is a finitely presented persistence vector space.

Persistence Vector Spaces Associated to Adapted Matrices

Let (X, ρ) and (Y, σ) be finite filtered sets. There is a map from the set of (ρ, σ)-adapted matrices to the set persistence vector spaces called the *θ-correspondence* , defined for a (ρ, σ)-adapted matrix A by

$$A \mapsto \theta(A) = \{V_{\mathbb{F}}(X, \rho)_r\}/\text{image}(f_A).$$

The is entirely analogous to the θ-correspondence for vector spaces defined in Section 2.4.7.

We have the following facts, which are more-or-less obvious, but will be useful to refer to.

Proposition 6.1.2. *The persistence vector space $\theta(A)$ is finitely presented. Moreover, any finitely-presented persistence vector space is of the form $\theta(A)$ for some adapted matrix A.*

Proposition 6.1.3. *Let A be a (ρ, σ)-adapted matrix, B a (ρ, ρ)-adapted matrix and C a (σ, σ)-adapted matrix. Then BAC is (ρ, σ)-adapted and $\theta(A)$ is isomorphic to $\theta(BAC)$.*

We also have the following example, which will prove very useful in proving the main theorem of this section.

Example 6.1.3. Let (X, ρ) and (Y, σ) be finite-filtered sets and let A be a (ρ, σ)-adapted matrix with the special property that it has at most one nonzero entry in each row and in each column and that the nonzero entries are all equal to 1. Consider the persistence vector space $\theta(A)$.

Label the elements of Y as $\{y_1, \ldots, y_n\}$. Then for each $y_j \in Y$,

$$f_A(\phi_{y_j}) = \sum_{x \in X} a_{xy_j} \phi_x = \phi_{x_j}$$

for some $x_j \in X$. Indeed, the x_j is the unique element of X so that $a_{x_j y_j} = 1$. We claim that $\theta(A)$ can be decomposed as

$$\theta(A) \approx \bigoplus_j P(\rho(x_j), \sigma(y_j)) \oplus \bigoplus_{x \notin \{x_1, \ldots, x_n\}} P(\rho(x), +\infty).$$

To see this, note that

$$\left(\bigoplus_j P(\rho(x_j), \sigma(y_j)) \oplus \bigoplus_{x \notin \{x_1, \ldots, x_n\}} P(\rho(x), +\infty) \right)_r = \bigoplus_{\rho(x_j) \leqslant r < \sigma(y_j)} \mathbb{F} \oplus \bigoplus_{x \in X[r], x \neq x_j} \mathbb{F}. \quad (6.1)$$

On the other hand,

$$\theta(A)_r = \{V_{\mathbb{F}}(X, \rho)_r\}/\text{image}((f_A)_r) = \text{span}_{\mathbb{F}}\{[\phi_x] \mid x \in X[r]\}, \quad (6.2)$$

with $[\phi_x] = [0]$ if and only if $x = x_j$ with $\sigma(y_j) < r$. Decomposing (6.2) further, we have

$$\operatorname{span}_{\mathbb{F}}\{[\phi_x] \mid x \in X[r]\} \approx \bigoplus_{\rho(x_j) \leqslant r < \sigma(y_j)} \operatorname{span}\{\phi_{x_j}\} \oplus \bigoplus_{x \in X[r], x \neq x_j} \operatorname{span}_{\mathbb{F}}\{\phi_x\},$$

which is clearly isomorphic to (6.1). This idea can be used to create an isomorphism of persistence vector spaces, as in Example 6.1.1.

Row and Column Operations for Adapted Matrices

Let (X, ρ) and (Y, σ) be finite filtered sets and let A be a (ρ, σ)-adapted matrix. Just as in the case of unadapted matrices (see Section 2.4.6), there are certain useful row and column operations for (ρ, σ)-adapted matrices :

1. Multiply all entries in a row/column by the same nonzero element of F,
2. Add a nonzero multiple of $r(x)$ to $r(x')$ when $\rho(x) \geqslant \rho(x')$,
3. Add a nonzero multiple of $c(y)$ to $c(y')$ when $\sigma(y) \leqslant \sigma(y')$

Proposition 6.1.4. *The row and column operations for (ρ, σ)-adapted matrices listed above preserve the property of being (ρ, σ)-adapted.*

Proof. Let $A = (a_{xy})$ be a (ρ, σ)-adapted matrix. The first operation doesn't change locations of zeros in the matrix, so it obviously preserves the property. Consider the second operation of adding $\lambda r(x)$ to $r(x')$, where $\rho(x) \geqslant \rho(x')$. Assume that $\rho(x') > \sigma(y)$; we then need to show that the matrix $A' = (a'_{xy})$ obtained after the row operation has $a'_{x'y} = 0$. This entry is given by $a'_{x'y} = a_{x'y} + \lambda a_{xy} = 0 + \lambda \cdot 0$, since $\rho(x) \geqslant \rho(x') > \sigma(y)$ implies that $a_{x'y} = a_{xy} = 0$. The proof in the case of the third operation is similar. □

We have the following proposition in analogy with Proposition 2.4.7, whose proof is left as an exercise.

Proposition 6.1.5. *Each adapted row and column operation corresponds to an isomorphism of a persistence vector space.*

Proposition 6.1.6. *For any (ρ, σ)-adapted matrix A, there exists a (ρ, σ)-adapted matrix A' with at most one nonzero entry in each row and each column, the nonzero entry equal to 1, such that $\theta(A) \approx \theta(A')$.*

Proof. Let A be a (ρ, σ)-adapted matrix for some filtered sets (X, ρ) and (Y, σ). We will prove the proposition by constructing A' inductively. Let $\overline{y} \in Y$ be such that $\sigma(\overline{y}) \leqslant \sigma(y)$ for all y such that $c(y)$ does not contain all zeros (assuming that this is a nonempty set, because if this were not the case then we would already be done). Now let $\overline{x} \in X$ be such that $\rho(\overline{x}) \geqslant \rho(x)$ for all x satisfying $a_{x\overline{y}} \neq 0$ (once again, this is a nonempty set; this time by our assumption on $c(\overline{y})$). We wish to "zero out" entries in $c(\overline{y})$ except for the $(\overline{x}, \overline{y})$-entry using row operations. This is possible because we can add a nonzero multiple of $r(\overline{x})$ to any other row $r(x)$ with $\rho(\overline{x}) \geqslant \rho(x)$ and the rows satisfying this condition are exactly those which potentially contain non-zero entries in the column $c(\overline{y})$! We can similarly "zero out" entries in $r(\overline{x})$ besides $a_{\overline{xy}}$ and the result is a (ρ, σ)-adapted matrix such that $r(\overline{x})$ and $c(\overline{y})$ each contain exactly one nonzero entry. Multiplying this row or column by $1/a_{\overline{xy}}$ ensures that the nonzero entry is equal to 1.

Now we continue inductively; define $\tilde{y}$ to be the element of Y satisfying $\sigma(\tilde{y}) \leqslant \sigma(y)$ for all y such that $y > \overline{y}$ and $c(y)$ does not contain all zeros and $\tilde{x}$ to be the element of X satisfying $\rho(\tilde{x}) \geqslant \rho(x)$ for all $x > \overline{x}$ with $a_{x\tilde{y}} \neq 0$. Running the same procedure will "zero out" the row $r(\tilde{x})$ and column $c(\tilde{y})$ as desired without effecting $r(\overline{x})$ or $c(\overline{y})$. This process can be repeated until we obtain a matrix A' with the desired properties. □

6.2 Main Theorem

We are now ready to state our main result. All vector spaces are assumed to be over the same field $\mathbb{F}$, so the field is supressed from the notation.

Theorem 6.2.1 (Fundamental Theorem of Persistent Homology). *For any finitely presented persistence vector space $\{V_r\}$, there exists a finite collection of birth-death pairs $(b_1, d_1), (b_2, d_2), \ldots, (b_n, d_n)$ with $b_i \in [0, \infty)$ and $d_i \in [0, \infty]$ such that*

$$\{V_r\} \approx P(b_1, d_1) \oplus P(b_2, d_2) \oplus \cdots \oplus P(b_n, d_n).$$

Moreover, the decomposition is unique up to reordering the factors.

Proof. By Proposition 6.1.2, there exist finite filtered sets (X, ρ) and (Y, σ) such that we can represent $\{V_r\}$ as $\theta(A)$ for some (ρ, σ)-adapted matrix A. Proposition 6.1.6 implies that there is a matrix A' with at most one entry equal to 1 in each row and in each column, with all other entries equal to 0, such that $\theta(A') \approx \theta(A)$. We already showed in Example 6.1.3 that $\theta(A')$ has a decomposition of the desired form. This proves the existence part of the theorem. The uniqueness part of the theorem follows immediately from Proposition 6.1.1.

□

6.3 Barcodes

By applying Theorem 6.2.1, we can now associate to any filtered simplicial complex $\{\mathcal{X}_r\}$ a topological signature called a *barcode* . To do so, we calculate the persistence homology of $\{\mathcal{X}_r\}$ to obtain a collection of finitely presented persistence vector spaces $\{H_k(\mathcal{X}_r)\}$ for $k = 0, 1, 2 \ldots$. Theorem 6.2.1 tells us that for each k there is a (unique up to reordering) decomposition

$$\{H_k(\mathcal{X}_r)\} \approx P(b_1, d_1) \oplus P(b_2, d_2) \oplus \cdots \oplus P(b_n, d_n),$$

which gives a multiset of birth-death pairs. To form the barcode, we draw a line segment for each birth-death pair. The barcode is then a collection of line segments, typically represented in the form

To illustrate this idea, we provide examples below.

6.3.1 Examples of Barcodes

6.4 Persistence Diagrams

There is another convenient representation of the persistence homology of a filtered simplicial complex called a *persistence diagram* . ...

6.4.1 Examples of Persistence Diagrams

6.5 Computing Barcodes

6.6 Other Representations of Persistent Homology

6.7 Exercises

1. Complete the proof of Proposition 6.1.4 by showing that the remaining operation preserves the property of being (ρ, σ)-adapted.

2. Prove Proposition 6.1.5.

7 Metrics on the Space of Barcodes (Under Construction)

8 Applications (Under Construction)

9 Appendix

9.1 Every Vector Space Has a Basis

In this section we provide a proof of Theorem 2.2.1, that every vector space has a basis. The proof relies on the *Axiom of Choice* :
Let $\mathcal{A}$ be a collection of nonempty, disjoint sets. There exists a set A containing exactly one element from each set in $\mathcal{A}$.

The Axiom of Choice can't be proved from any of the other usual axioms of set theory, and can therefore only be accepted or not accepted. Most modern mathematicians choose to accept the Axiom of Choice as a basic axiom of set theory. In fact, a major motivation for accepting it is that it is *equivalent* to the statement that every vector space has a basis!

It is well known that the Axiom of Choice is equivalent to *Zorn's Lemma* , which is a statement about ordered sets. To state it, we need to introduce some definitions.

Let A be a set. A *partial order* on A is a relation $\prec$ (i.e. we denote that $a, b \in A$ are related by the notation $a \prec b$) such that $a \prec a$ holds for any $a \in A$ and for every $a, b, c \in A$, $a \prec b$ and $b \prec c$ implies $a \prec c$. An element $a \in A$ is a *maximum* of A if $a \prec b$ implies that $a = b$. Let $B \subset A$. An element $a \in A$ is an *upper bound* on B if $b \prec a$ for every $b \in B$. We say that B is *totally ordered* if for every $a, b \in B$, either $a \prec b$ or $b \prec a$.

Example 9.1.1. The set $\mathbb{R}$ has partial order $\prec = \leqslant$. In this case, every subset is totally ordered.

On the other hand, we could take our set A to be the power set $\mathcal{P}(\mathbb{R})$ of all subsets of $\mathbb{R}$. This set admits a partial order given by inclusion, $\prec = \subseteq$. Many subsets are not totally ordered. The whole set $\mathcal{P}(\mathbb{R})$ is not totally ordered: $[0, 1]$ and $[1, 2]$ that neither is a subset of the other. One example of a totally ordered subset of $\mathcal{P}(\mathbb{R})$ is the set

$$\{[0, n] \mid n \in Z_{>0}\}.$$

We can now state Zorn's Lemma.

Theorem 9.1.1 (Zorn's Lemma)**.** *Let A be a set with partial order $\prec$. If every totally ordered subset B admits an upper bound, then A has a maximum.*

We are now prepared to prove Theorem 2.2.1.

Proof. Let $\mathcal{X}$ denote the collection of all linearly independent subsets of V. This set is partially ordered by inclusion. We wish to apply Zorn's Lemma to show that $\mathcal{X}$ contains a maximum element B. If such a maximum exists, then it must be a basis. Indeed, it is linearly independent by definition. It must also be spanning, since for any $v \in V$,

$B \subset B \cup \{v\}$, there are two possibilities: either $v \in B$ in which case it is clear that $v \in \operatorname{span}(B)$ or $v \notin B$, so it must be that $B \cup \{v\}$ is not linearly independent and we once again conclude that $v \in \operatorname{span}(B)$.

It remains to show that we can apply Zorn's Lemma. Let $\mathcal{Y} \subset \mathcal{X}$ be a collection of linearly independent subsets which is totally ordered by inclusion. We wish to show that it is upper bounded by an element of $\mathcal{X}$. Let $Y_0 = \cup_{Y \in \mathcal{Y}} Y$. Then for any $Y \in \mathcal{Y}$, we certainly have $Y \subset Y_0$. It therefore remains to show that $Y_0 \in \mathcal{X}$; that is, Y_0 is linearly independent. Any finite linear combination of elements of Y_0 can be written as a linear combination of elements of some set $Y \in \mathcal{Y}$, by virtue of the total ordering of $\mathcal{Y}$. Therefore Y_0 is linearly independent.

To show that any linearly independent subset S of V can be extended to a basis, we can mimic the above argument by replacing $\mathcal{X}$ with the set of all linearly independent subsets of V which contain S. $\square$

Bibliography

[1] Carlsson, Gunnar. “Topological pattern recognition for point cloud data.” Acta Numerica 23 (2014): 289-368.

[2] Hatcher, Allen. “Algebraic topology. 2002.” Cambridge UP, Cambridge 606, no. 9 (2002).

[3] Hoffman, K. and Kunze, R., Linear algebra. 1971. Englewood Cliffs, New Jersey.

[4] Johnson, Lee W., Ronald Dean Riess, and Jimmy Thomas Arnold. Introduction to linear algebra. Addison-Wesley Longman, 1993.

[5] Manolescu, Ciprian. “Pin (2)-equivariant Seiberg-Witten Floer homology and the triangulation conjecture.” Journal of the American Mathematical Society 29, no. 1 (2016): 147-176.

[6] Munkres, James R. Topology. Prentice Hall, 2000.

Index