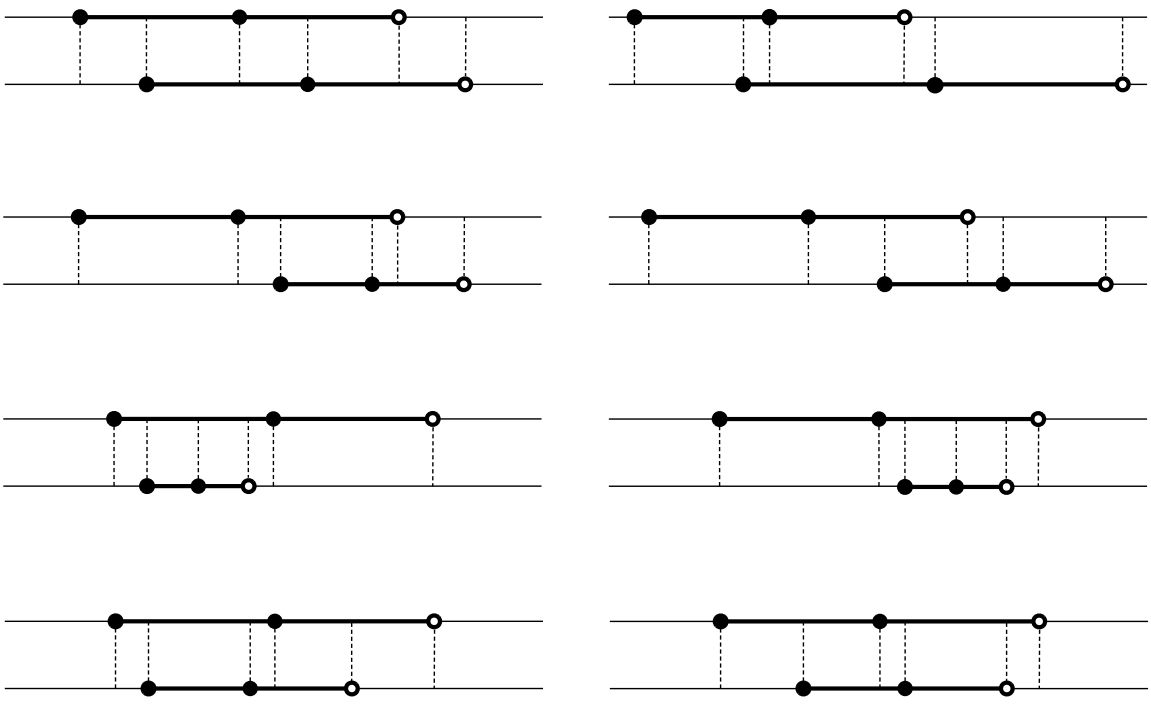
Introduction to Applied Algebraic Topology

Tom Needham



Last Updated: April 10, 2019

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.

Some Acknowledgements:

I would like to acknowledge contributions from several people to the creation of this text. The first time this course was run, I taught it jointly with Crichton Ogle and he had a large influence on the content and style of these notes. He discovered the novel proof of the Fundamental Theorem of Persistence Vector Spaces which is presented here. I would also like to thank Facundo Mémoli, who pushed for the creation of the course these notes are based on and with whom I have had countless discussions about Topological Data Analysis. Lastly, many thanks are due to students who have taken this course and graciously pointed out errors in the text and suggested improvements to exposition.

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

Contents

1 Review of Linear Algebra

A strong grasp of abstract linear algebra will be essential for the latter material in this text. Our focus will be somewhat different than that of a standard undergraduate linear algebra course. In particular, we will require a firm understanding of quotient constructions and some basic ideas from homological algebra. For a more in-depth treatment of some of the topics covered here, see, for example, [5, 6].

1.1 Abstract Vector Spaces

1.1.1 Definitions

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 main examples of fields.

Example 1.1.1. 1. The real numbers $\mathbb{R}$ form a field with the obvious multiplication and addition operations.

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

3. 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

We leave it as an exercise to show that F_2 satisfies the field axioms.

For the applications that we are interested in, it will be sufficient to keep the examples $\mathbb{R}$ and F_2 in mind.

Vector Spaces Over A Field

Let $\mathbb{F}$ be a field. A *vector space over* $\mathbb{F}$ is a set V together with an operation

$$\begin{aligned} + : V \times V &\to V \\ (\mathbf{v}_1, \mathbf{v}_2) &\mapsto \mathbf{v}_1 + \mathbf{v}_2 \end{aligned}$$

called *vector addition* and an operation

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

called *scalar multiplication*. In this context, elements of V are called *vectors* and elements of $\mathbb{F}$ are called *scalars*. We require that the following axioms are satisfied:

1. (Additive Associativity) For any elements $\mathbf{v}_1$, $\mathbf{v}_2$ and $\mathbf{v}_3$ in V,

$$(\mathbf{v}_1 + \mathbf{v}_2) + \mathbf{v}_3 = \mathbf{v}_1 + (\mathbf{v}_2 + \mathbf{v}_3).$$

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

$$\lambda_1 \cdot (\lambda_2 \cdot \mathbf{v}) = (\lambda_1 \lambda_2) \cdot \mathbf{v}.$$

3. (Additive Commutativity) For any $\mathbf{v}_1$ and $\mathbf{v}_2$ in V,

$$\mathbf{v}_1 + \mathbf{v}_2 = \mathbf{v}_2 + \mathbf{v}_1.$$

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

$$\mathbf{v} + \mathbf{0} = \mathbf{v}.$$

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

6. (Distributive Law I) For any $\lambda \in \mathbb{F}$ and $\mathbf{v}_1$ and $\mathbf{v}_2$ in V,
$$\lambda \cdot (\mathbf{v}_1 + \mathbf{v}_2) = \lambda \cdot \mathbf{v}_1 + \lambda \cdot \mathbf{v}_2.$$

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

8. (Scalar Multiple Identity) For any $\mathbf{v}$ in V,
$$1_{\mathbb{F}} \cdot \mathbf{v} = \mathbf{v}.$$

Note that we have adopted the notational conventions:

- vector spaces are denoted by capital letters U, V, W;
- vectors are denoted by bold letters $\mathbf{v}$, $\mathbf{w} \in V$;
- the zero vector is denoted $\mathbf{0}$, but we may also use $0_V \in V$ when the vector space to which it belongs needs to be emphasized;
- scalars are denoted by Greek letters μ, $\lambda \in \mathbb{F}$.

This notation scheme is generally followed for abstract vector spaces, but notation will frequently be specialized when referring to specific spaces.

1.1.2 Examples of Vector Spaces

In order to understand the utility of this abstract definition of a vector space, it is important to see the wide variety of examples.

Example 1.1.2. The example of a vector space over $\mathbb{R}$ that you are probably most familiar with is $\mathbb{R}^n$ (for some positive integer n). This is the set of n-tuples of real numbers, with vector addition given by the following formula: for vectors $\mathbf{v} = (v_1, v_2, \ldots, v_n)$ and $\mathbf{w} = (w_1, \ldots, w_n)$, we define
$$\mathbf{v} + \mathbf{w} = (v_1, v_2, \ldots, v_n) + (w_1, w_2, \ldots, w_n) = (v_1 + w_1, v_2 + w_2, \ldots, v_n + w_n).$$
Scalar multiplication of $\lambda \in \mathbb{R}$ and $\mathbf{v}$ is given by
$$\lambda \cdot \mathbf{v} = \lambda \cdot (v_1, v_2, \ldots, v_n) = (\lambda v_1, \lambda v_2, \ldots, \lambda v_n).$$
It is an easy but instructive exercise to check that these operations satisfy the axioms of a vector space over $\mathbb{R}$.

Example 1.1.3. For any field $\mathbb{F}$, the set of n-tuples of field elements $\mathbb{F}^n$ has a vector space structure with operations defined exactly as in the real case. In particular, $\mathbb{C}^n$ and F_2^n are vector spaces over $\mathbb{C}$ and F_2, respectively.

Example 1.1.4. Scalar multiplication by a real number still makes sense in $\mathbb{C}^n$ (since $\mathbb{R} \subset \mathbb{C}$), so $\mathbb{C}^n$ is also has the structure of a vector space over $\mathbb{R}$.

Example 1.1.5. Let

$$P_n(\mathbb{R}) = \{\text{functions } p : \mathbb{R} \to \mathbb{R} \mid p(x) = \lambda_0 + \lambda_1 x + \cdots + \lambda_n x^n \text{ for some scalars } \lambda_j \in \mathbb{R}\}.$$

That is, $P_n(\mathbb{R})$ denotes the set of polynomial functions $p : \mathbb{R} \to \mathbb{R}$ of degree at most n. Note that if a function $p : \mathbb{R} \to \mathbb{R}$ can be expressed in the form $p(x) = \lambda_0 + \lambda_1 x + \cdots + \lambda_n x^n$ with $\lambda_n \neq 0$, then the degree n and the coefficients λ_j are unique (take a moment to convince yourself of this). 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 $\mathbb{R}^{n+1}$. 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 1.3.1).

Example 1.1.6. More generally, for any field $\mathbb{F}$, let

$$P_n(\mathbb{F}) = \{\text{functions } p : \mathbb{F} \to \mathbb{F} \mid p(x) = \lambda_0 + \lambda_1 x + \lambda_2 x^2 + \cdots + \lambda_n x^n \text{ for some scalars } \lambda_j \in \mathbb{F}\}.$$

These polynomial spaces define vector fields over $\mathbb{F}$, with the vector space structure defined analogously to the real case. How does the set $P_n(F_2)$ compare to the full set of functions $\{p : F_2 \to F_2\}$ (with no "polynomial" restriction)?

Example 1.1.7. Consider the set $C^\infty([0,1], \mathbb{R})$ of smooth (infinitely differentiable) functions $f : [0,1] \to \mathbb{R}$ (at the endpoints, we take f to be infinitely left and right differentiable, respectively). We claim that this set forms a vector space with *pointwise addition* and *scalar multiplication*. That is, for vectors (functions) f and g in $C^\infty([0,1], \mathbb{R})$ and a scalar λ, we define the vectors (functions) $f + g$ and λf by

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

and

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

respectively. While $\mathbb{R}^n$ and $P_{n-1}(\mathbb{R})$ (from Example 1.1.5) appear to be equivalent in some way, $C^\infty([0,1], \mathbb{R})$ should feel "different". In fact, R^n and $P_{n-1}(\mathbb{R})$ are n-dimensional and

$C^\infty([0,1],\mathbb{R})$ is infinite-dimensional (dimension will be defined precisely in Section 1.2.2), 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*!).

Example 1.1.8. The space of smooth functions $C^\infty([0,1],\mathbb{C})$ is a vector space over $\mathbb{C}$. There is no corresponding notion of a smooth mapping space over F_2.

1.2 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 $\mathbf{e}_j \in \mathbb{R}^n$ denote the n-tuple with a 1 in the jth entry and zeros elsewhere; e.g., $\mathbf{e}_1 = (1,0,0,\ldots,0)$. Then any element $(v_1,\ldots,v_n) \in \mathbb{R}^n$ can be decomposed as

$$(v_1,\ldots,v_n) = v_1 \cdot \mathbf{e}_1 + v_2 \cdot \mathbf{e}_2 + \cdots + v_n \cdot \mathbf{e}_n.$$

The set $\{\mathbf{e}_1,\ldots,\mathbf{e}_n\}$ 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 $\mathbb{F}$.

1.2.1 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

$$\sum_{\mathbf{s}\in S} \lambda_{\mathbf{s}}\mathbf{s},$$

where each $\lambda_{\mathbf{s}} \in \mathbb{F}$ 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 $\lambda_{\mathbf{s}} \in \mathbb{F}$, not all zero, such that

$$\sum_{\mathbf{s}\in S} \lambda_{\mathbf{s}}\mathbf{s} = \mathbf{0}.$$

If no such collection of scalars exist, the set is called *linearly independent* (note that the empty set is linearly independent, vacuously). A *basis* for V is a linearly independent spanning set.

The *zero vector space* or *trivial vector space* is the space that contains only the zero vector $\{\mathbf{0}\}$; up to labeling of the zero element this space is unique (there is only one trivial vector space). A vector space that contains at least one non-zero vector is referred to as *non-zero*. We have the following fundamental theorem.

Theorem 1.2.1. *Every non-zero 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 is equivalent to the Axiom of Choice. We will skip it for now, but the interested reader is invited to read a proof sketch in Section 7.2.

1.2.2 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. We now show that our definition of definition actually makes sense; that is, if we choose two different bases for V then they will always have the same number of elements.

Proposition 1.2.2. *The dimension of V is independent of choice of basis.*

Proof. Let A and B be bases for V. Our goal is to show that $|A| = |B|$. If $|A| = |B| = \infty$ then we are done, so let's assume by way of obtaining a contradiction (and without loss of generality) that $|A| = m < \infty$ and $|A| < |B|$. Write $A = \{\mathbf{a}_1, \ldots, \mathbf{a}_m\}$ and choose a subset of $m+1$ vectors from B, $\{\mathbf{b}_1, \ldots, \mathbf{b}_m, \mathbf{b}_{m+1}\} \subset B$. We will show that the vectors chosen from B are linearly dependent, hence the set B itself is linearly dependent, thus obtaining our contradiction.

For each $\mathbf{b}_j$, choose coefficients λ_k^j such that

$$\mathbf{b}_j = \lambda_1^j \mathbf{a}_1 + \lambda_2^j \mathbf{a}_2 + \cdots + \lambda_m^j \mathbf{a}_m = \sum_{k=1}^{m} \lambda_k^j \mathbf{a}_k.$$

To show that the subset of B is linearly dependent, we seek coefficients μ_j (not all zero) such that

$$\mu_1 \mathbf{b}_1 + \mu_2 \mathbf{b}_2 + \cdots + \mu_{m+1} \mathbf{b}_{m+1} = 0_V.$$

Writing the $\mathbf{b}_j$ in terms of the vectors in A, we seek μ_j such that

$$\mu_1 \sum_{k=1}^{m} \lambda_k^1 \mathbf{a}_k + \mu_2 \sum_{k=1}^{m} \lambda_k^2 \mathbf{a}_k + \cdots + \mu_{m+1} \sum_{k=1}^{m} \lambda_k^{m+1} \mathbf{a}_k = 0_V,$$

which can be rearranged as

$$\left(\sum_{j=1}^{m+1} \mu_j \lambda_1^j \right) \mathbf{a}_1 + \left(\sum_{j=1}^{m+1} \mu_j \lambda_2^j \right) \mathbf{a}_2 + \cdots + \left(\sum_{j=1}^{m+1} \mu_j \lambda_m^j \right) \mathbf{a}_m = 0_V$$

Since $\mathbf{a}_1, \dots \mathbf{a}_m$ are linearly independent, this occurs if and only if each coefficient is zero, so we obtain a system of linear equations

$$\begin{aligned}
\mu_1\lambda_1^1 + \mu_2\lambda_1^2 + \cdots + \mu_{m+1}\lambda_1^{m+1} &= 0 \\
\mu_1\lambda_2^1 + \mu_2\lambda_2^2 + \cdots + \mu_{m+1}\lambda_2^{m+1} &= 0 \\
&\vdots \\
\mu_1\lambda_m^1 + \mu_2\lambda_m^2 + \cdots + \mu_{m+1}\lambda_m^{m+1} &= 0.
\end{aligned}$$

There are $m+1$ unknowns (the μ_j) and m equations, so we can find a nontrivial solution (i.e., not all μ_j are zero). □

Example 1.2.1. 1. The set $B = \{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n\} \subset \mathbb{R}^n$ defined at the beginning of this section is a basis for $\mathbb{R}^n$, and is frequently referred to as the *canonical basis* for $\mathbb{R}^n$. Indeed, B is spanning, since any $(x_1, \dots, x_n)$ can be written as the linear combination

$$(x_1, x_2, \dots, x_n) = x_1\mathbf{e}_1 + x_2\mathbf{e}_2 + \cdots + x_n\mathbf{e}_n.$$

Moreover, B is linearly independent: if λ_j are scalars such that

$$\lambda_1\mathbf{e}_1 + \lambda_2\mathbf{e}_2 + \cdots + \lambda_n\mathbf{e}_n = \mathbf{0} = (0, 0, \dots, 0),$$

then $\lambda_j = 0$ for all j. 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 of polynomials $P_n(\mathbb{R})$ defined in Example 1.1.5 has a basis given by monomials $\{1 = x^0, x = x^1, x^2, x^3, \dots, x^n\}$. It follows that the dimension of $P_n(\mathbb{R})$ is $n+1$.

4. Let $P_\infty(\mathbb{R}) = \bigcup_{n=0}^{\infty} P_n(\mathbb{R})$ denote the set of all polynomials with real coefficients. Using the addition and scalar multiplication operations from the definition of $P_n(\mathbb{R})$, $P_\infty(\mathbb{R})$ is a vector space over $\mathbb{R}$. The infinite list of monomials $\{1, x, x^2, x^3, \dots\}$ defines a basis for $P_\infty(\mathbb{R})$, so $P_\infty(\mathbb{R})$ is infinite-dimensional.

5. The vector space $C^\infty([0,1], \mathbb{R})$ is infinite-dimensional (see the exercises).

1.3 Linear Transformations

1.3.1 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. $L(\mathbf{v}_1 + \mathbf{v}_2) = L(\mathbf{v}_1) + L(\mathbf{v}_2)$ for all $\mathbf{v}_1, \mathbf{v}_2 \in V$,

2. $L(\lambda\mathbf{v}) = \lambda L(\mathbf{v})$ for all $\mathbf{v} \in V$ 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 1.3.1. Let $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$ denote the standard basis for $\mathbb{R}^3$. Consider the linear map $L : \mathbb{R}^3 \to \mathbb{R}^3$ defined by

$$L(\mathbf{e}_1) = 2\mathbf{e}_2 + \mathbf{e}_3, \;\; L(\mathbf{e}_2) = \mathbf{e}_1, \;\; L(\mathbf{e}_3) = \mathbf{0}.$$

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 $\mathbf{e}_j$ 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

$$\mathbf{v} = \lambda_1\mathbf{e}_1 + \lambda_2\mathbf{e}_2 + \lambda_3\mathbf{e}_3$$

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

$$L(\mathbf{v}) = \lambda_1 L(\mathbf{e}_1) + \lambda_2 L(\mathbf{e}_2) + \lambda_3 L(\mathbf{e}_3) = 2\lambda_1\mathbf{e}_2 + \lambda_1\mathbf{e}_3 + \lambda_2\mathbf{e}_1.$$

As a concrete example, the vector $\mathbf{v} = (1, 2, 3) = \mathbf{e}_1 + 2\mathbf{e}_2 + 3\mathbf{e}_3$ takes the value

$$L(\mathbf{v}) = 2 \cdot 1 \cdot \mathbf{e}_2 + 1 \cdot \mathbf{e}_3 + 2 \cdot \mathbf{e}_1 = (2, 2, 1).$$

Linear Extensions

Let us expand on the observation from Example 1.3.1 that linear maps can be defined by defining their values on basis elements.

Proposition 1.3.1. *Let V and W be vector spaces, let $B = \{\boldsymbol{b}_1, \ldots, \boldsymbol{b}_n\}$ be a fixed basis for V and let $L : B \to W$ be some function. There exists a unique linear map $\overline{L} : V \to W$ such that $\overline{L}|_B = L$.*

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

$$\mathbf{v} = \lambda_1\mathbf{b}_1 + \lambda_2\mathbf{b}_2 + \cdots + \lambda_n\mathbf{b}_n$$

for some scalars λ_j. A map $\overline{L} : V \to W$ which is linear and which restricts to $\overline{L}|_B = L$ must take $\mathbf{v}$ to the vector

$$\overline{L}(\mathbf{v}) = \lambda_1 L(\mathbf{e}_1) + \lambda_2 L(\mathbf{e}_2) + \cdots + \lambda_n L(\mathbf{e}_n).$$

We therefore take this as our definition of $\overline{L}$. The desired properties of $\overline{L}$ follow by construction. □

The process of defining a linear map from its values on basis vectors is called *extending linearly*. It will be quite common to define a linear map by specifiying its values on a basis.

Linear Isomorphisms

A linear transformation which is a bijection is called a *linear isomorphism*. Vector spaces V and W are called *isomorphic* if there is a linear isomorphism between them $L : V \to W$. In this case we write $V \approx W$.

Proposition 1.3.2. *Let $L : V \to W$ be a linear isomorphism. The inverse function $L^{-1} : W \to V$ is a linear map.*

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

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

and

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

□

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

Lemma 1.3.3. *A linear map $L : V \to W$ is injective if and only if $L(\boldsymbol{v}) = 0_W$ implies $\boldsymbol{v} = 0_V$.*

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 $\mathbf{v} = 0_V$. On the other hand, assume that the only element of V which maps to 0_W is 0_V. Then for any $\mathbf{v}, \mathbf{v}' \in V$ with $\mathbf{v} \neq \mathbf{v}'$,

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

and it follows that L is injective. □

Lemma 1.3.4. *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 $L(\mathbf{s})$ for some $\mathbf{s} \in S$. A linear combination of elements of the image of S then satisfies

$$0_W = \sum_{\mathbf{s}\in S} \lambda_{\mathbf{s}} L(\mathbf{s}) = \sum_{\mathbf{s}\in S} L(\lambda_{\mathbf{s}}\mathbf{s}) = L\left(\sum_{\mathbf{s}\in S} \lambda_{\mathbf{s}}\mathbf{s}\right)$$

only if $\sum \lambda_{\mathbf{s}}\mathbf{s} = 0_V$ by Lemma 1.3.3. The independence of S then implies that $\lambda_{\mathbf{s}} = 0$ for all $\mathbf{s}$ 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 $\mathbf{w} \in W$, the surjectivity of L implies that there exists $\mathbf{v} \in V$ with $L(\mathbf{v}) = \mathbf{w}$ (although such a $\mathbf{v}$ is not necessarily unique). Since S is spanning, there exist coefficients $\lambda_{\mathbf{s}}$ such that $\mathbf{v} = \sum \lambda_{\mathbf{s}} \mathbf{s}$ and it follows that $\mathbf{w} = \sum \lambda_{\mathbf{s}} L(\mathbf{s})$. □

We have the following immediate corollary.

Corollary 1.3.5. *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 1.3.6. *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 $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ and $\{\mathbf{w}_1, \ldots, \mathbf{w}_m\}$ be bases for V and W, respectively. If $V \approx W$, there exists a linear isomorphism $L : V \to W$. By Lemma 1.3.4, 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 $L(\mathbf{v}_j) = \mathbf{w}_j$ and extending. This is clearly an isomorphism. □

Example 1.3.2. It follows from Example 1.2.1 and Proposition 1.3.6 that the spaces $P_n(\mathbb{R})$ and $\mathbb{R}^{n+1}$ are isomorphic.

1.3.2 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 1.3.3. Consider the linear map $L : \mathbb{R}^3 \to \mathbb{R}^3$ from Example 1.3.1. Using the standard column vector notation

$$\lambda_1 \mathbf{e}_1 + \lambda_2 \mathbf{e}_2 + \lambda_3 \mathbf{e}_3 \leftrightarrow \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \end{pmatrix},$$

the map can be written as matrix multiplication:

$$L(\lambda_1 \mathbf{e}_1 + \lambda_2 \mathbf{e}_2 + \lambda_3 \mathbf{e}_3) = \begin{pmatrix} 0 & 1 & 0 \\ 2 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \end{pmatrix} = \begin{pmatrix} \lambda_2 \\ 2\lambda_1 \\ \lambda_1 \end{pmatrix} \leftrightarrow \lambda_2 \mathbf{e}_1 + 2\lambda_1 \mathbf{e}_2 + \lambda_1 \mathbf{e}_3.$$

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 1.3.4. 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 $B = \{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ for V and $C = \{\mathbf{w}_1, \ldots, \mathbf{w}_m\}$ for W, we can represent L by a size $m \times n$ matrix as follows.

We will use a slightly nonstandard subscript notation to indicate that a vector or a matrix is being represented in a particular choice of basis. That is, we can express an arbitrary $\mathbf{v} \in V$ uniquely as a linear combination of basis elements

$$\mathbf{v} = \lambda_1 \mathbf{v}_1 + \lambda_2 \mathbf{v}_2 + \cdots + \lambda_n \mathbf{v}_n$$

and we let ${}_B\mathbf{v}$ denote the column vector of these scalar coefficients:

$${}_B\mathbf{v} = \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_n \end{pmatrix}.$$

We then have the following theorem.

Theorem 1.3.7. *Any linear transformation $L : V \to W$ can be represented as an $m \times n$ matrix ${}_CL_B$ (with respect to bases $B = \{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n\}$ and $C = \{\boldsymbol{w}_1, \ldots, \boldsymbol{w}_m\}$) so that for every $\boldsymbol{v} \in V$,*

$${}_CL(\boldsymbol{v}) = {}_CL_B \cdot {}_B\boldsymbol{v},$$

where the product on the right is matrix multiplication of the $m \times n$ matrix ${}_CL_B$ with the $n \times 1$ column vector ${}_B\boldsymbol{v}$. In particular, the matrix representation is given by the formula

$${}_CL_B = \left({}_CL(\boldsymbol{v}_1)\ {}_CL(\boldsymbol{v}_2)\ \cdots\ {}_CL(\boldsymbol{v}_n)\right).$$

Proof. First note that the operation which takes a vector $\mathbf{v}$ to its column vector representation ${}_B\mathbf{v}$ is linear. Indeed, for a pair of vectors $\mathbf{v}$ and $\mathbf{v}'$ with basis representations

$$\begin{aligned}\mathbf{v} &= \lambda_1\mathbf{v}_1 + \cdots + \lambda_n\mathbf{v}_n,\\ \mathbf{v}' &= \lambda_1'\mathbf{v}_1' + \cdots + \lambda_n'\mathbf{v}_n',\end{aligned}$$

we have

$$\begin{aligned}{}_B(\mathbf{v}+\mathbf{v}') &= {}_B((\lambda_1+\lambda_1')\mathbf{v}_1 + \cdots + (\lambda_n+\lambda_n')\mathbf{v}_n)\\ &= \begin{pmatrix}\lambda_1+\lambda_1'\\ \vdots\\ \lambda_n+\lambda_n'\end{pmatrix} = \begin{pmatrix}\lambda_1\\ \vdots\\ \lambda_n\end{pmatrix} + \begin{pmatrix}\lambda_1'\\ \vdots\\ \lambda_n'\end{pmatrix} = {}_B\mathbf{v} + {}_B\mathbf{v}'.\end{aligned}$$

A similar argument shows linearity over scalar multiples.

We now proceed with the proof. For any $\mathbf{v}$, we have (by the linearity demonstrated above)

$$\begin{aligned}{}_CL(\mathbf{v}) &=_C L(\lambda_1\mathbf{v}_1 + \cdots + \lambda_n\mathbf{v}_n)\\ &= \lambda_1{}_CL(\mathbf{v}_1) + \cdots + \lambda_n{}_CL(\mathbf{v}_n)\\ &= ({}_CL(\mathbf{v}_1)\ {}_CL(\mathbf{v}_2)\ \cdots\ {}_CL(\mathbf{v}_n))\cdot {}_B\mathbf{v}\\ &= {}_CL_B\cdot {}_B\mathbf{v}.\end{aligned}$$

□

Another way to express this theorem using more standard notation is as follows. For each $\mathbf{v}_j \in B$, we can write

$$L(\mathbf{v}_j) = \lambda_{1j}\mathbf{w}_1 + \lambda_{2j}\mathbf{w}_2 + \cdots + \lambda_{mj}\mathbf{w}_m$$

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.

1.4 Vector Space Constructions

1.4.1 Subspaces

Let V be a vector space over $\mathbb{F}$. A nonempty 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 $\mathbf{u}, \mathbf{v} \in U$, $\mathbf{u} + \mathbf{v} \in U$
2. (Closure Under Scalar Multiplication) for all $\mathbf{u} \in U$ and $\lambda \in \mathbb{F}$, $\lambda\mathbf{u} \in U$.

Note that our assumption that $U \neq \varnothing$, together with closure under scalar multiplication, implies that $0_V \in U$. The *dimension* of a vector subspace is just its dimension as a vector space, using the usual definition.

Example 1.4.1. 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 $\mathbf{v}_1, \ldots, \mathbf{v}_m \in V$ linearly independent, the set
$$\text{span}(\{\mathbf{v}_1, \ldots, \mathbf{v}_m\})$$
is an m-dimensional vector subspace
3. the full space V.

Example 1.4.2. 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 $C^\infty([0,1], \mathbb{R})$.

1.4.2 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

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

Proposition 1.4.1. *The kernel of a linear transformation $L : V \to W$ is a vector subspace of V.*

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

$$L(\mathbf{u} + \mathbf{v}) = L(\mathbf{u}) + L(\mathbf{v}) = 0_W + 0_W = 0_W$$

and

$$L(\lambda\mathbf{u}) = \lambda L(\mathbf{u}) = \lambda 0_W = 0_W.$$

□

Image

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

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

The image of L is also commonly referred to as the *range* of L and denoted range(L).

Proposition 1.4.2. *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.

1.4.3 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 1.4.3 (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 $\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$ be a basis for $\ker(L) \subset V$, so that $\text{null}(L) = k$. By Theorem 1.2.1, the set $\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$ can be extended to a basis $\{\mathbf{v}_1, \ldots, \mathbf{v}_k, \mathbf{w}_1, \ldots, \mathbf{w}_{n-k}\}$ for V. We claim that the set $B = \{L(\mathbf{w}_1), \ldots, L(\mathbf{w}_{n-k})\}$ forms a basis for image(L), and this will complete the proof of the theorem.

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

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

and this shows that B is spanning.

To see that B is linearly independent, suppose that

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

Since L is injective on $\text{span}\{\mathbf{w}_1, \ldots, \mathbf{w}_\ell\}$, it follows from Lemma 1.3.3 that $\sum \nu_\ell \mathbf{w}_\ell = 0_V$. Since the $\mathbf{w}_\ell$ are linearly independent, this implies that all $\nu_\ell = 0$. □

1.4.4 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

$$(\mathbf{v}_1, \mathbf{w}_1) + (\mathbf{v}_2, \mathbf{w}_2) = (\mathbf{v}_1 + \mathbf{v}_2, \mathbf{w}_1 + \mathbf{w}_2)$$

and scalar multiplication defined by

$$\lambda \cdot (\mathbf{v}, \mathbf{w}) = (\lambda \cdot \mathbf{v}, \lambda \cdot \mathbf{w}).$$

Proposition 1.4.4. *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 $\mathbf{w} \in W$, the set $\{(\mathbf{s}, \mathbf{w}) \mid \mathbf{s} \in S\}$ is an infinite linearly independent subset of $V \oplus W$.

On the other hand, if V and W are both finite-dimensional, let $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ be a basis for V and $\{\mathbf{w}_1, \ldots, \mathbf{w}_m\}$ a basis for W. Then it is easy to check that

$$\{(\mathbf{v}_j, 0)\} \cup \{(0, \mathbf{w}_k)\}$$

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

1.4.5 Quotient Spaces

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

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

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

Proposition 1.4.5. *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

$$[\mathbf{u}] + [\mathbf{v}] = [\mathbf{u} + \mathbf{v}]$$

and we define scalar multiplication by the formula

$$\lambda[\mathbf{u}] = [\lambda \mathbf{u}].$$

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 1.4.6. *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 1.2.1). We claim that

$$\{[\mathbf{b}] \mid \mathbf{b} \in B \backslash B'\} \tag{1.1}$$

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

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

The existence of the coefficients $\lambda_{\mathbf{b}}$ 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 $[\mathbf{b}] = [0]$ for any $\mathbf{b} \in B'$. Moreover, the set (1.1) is linearly independent, as

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

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

Finally, we claim that the set (1.1) contains $|B|-|B'|$ distinct elements. Its cardinality is certainly bounded above by this number, so we need to check that if $\mathbf{b}_1, \mathbf{b}_2 \in B\backslash B'$ satisfy $\mathbf{b}_1 \neq \mathbf{b}_2$, then $[\mathbf{b}_1] \neq [\mathbf{b}_2]$. This holds because $[\mathbf{b}_1] = [\mathbf{b}_2]$ if and only if $\mathbf{b}_1 - \mathbf{b}_2 \in B'$, which is impossible by linear independence. □

Quotient Maps

Let V be a vector space and U a subspace. There is a natural linear map $Q : V \to V/U$ called the *quotient map* defined by

$$Q(\mathbf{v}) = [\mathbf{v}].$$

Lemma 1.4.7. *The quotient map $Q : V \to V/U$ is a linear surjection.*

Proof. For any $\mathbf{v}, \mathbf{w} \in V$,

$$Q(\mathbf{v} + \mathbf{w}) = [\mathbf{v} + \mathbf{w}] = [\mathbf{v}] + [\mathbf{w}],$$

by definition of vector addition in V/U. Linearity over scalar multiplication follows similarly. Moreover, any $[\mathbf{v}] \in V/U$ is the image of the vector $\mathbf{v}$ under the quotient map, so Q is surjective. □

Proposition 1.4.8. *Any linear map $L : V \to W$, factors as the composition $L = L' \circ Q$, where*

$$Q : V \to V/\ker(L)$$

is the quotient map and

$$L' : V/\ker(L) \to W$$

is an injective map. If L is surjective, then L' is an isomorphism.

Proof. We define $L' : V/\ker(L) \to W$ by

$$L'([\mathbf{v}]) = L(\mathbf{v}).$$

We need to check that this is well defined. Let $\mathbf{v}' \in V$ such that $[\mathbf{v}] = [\mathbf{v}']$. Then

$$L'([\mathbf{v}']) - L'([\mathbf{v}]) = L(\mathbf{v}) - L(\mathbf{v}') = L(\mathbf{v} - \mathbf{v}') = 0_W,$$

since $[\mathbf{v}] = [\mathbf{v}']$ implies $\mathbf{v} - \mathbf{v}' \in \ker(L)$.

Clearly L' is linear, by the definition of vector addition in the quotient space. We also have injectivity, since $L'([\mathbf{v}]) = 0_W$ implies that $L(\mathbf{v}) = 0_W$, hence that $\mathbf{v} \in \ker(L)$, so $[\mathbf{v}] = [0_V]$. Finally, it follows easily from the assumption of surjectivity of L that the map L' is surjective. The last thing to check is that $L = L' \circ Q$. For any $\mathbf{v} \in V$, we have

$$L' \circ Q(\mathbf{v}) = L'([\mathbf{v}]) = L(\mathbf{v}).$$

□

Split Maps

A linear map $L : V \to W$ is called *split* if there is another linear map $L' : W \to V$ such that $L \circ L'$ is the identity map on W. It turns out that every surjective linear map L is split.

Theorem 1.4.9. *Every surjective linear map $L : V \to W$ is split.*

Proof. By Proposition 1.4.8, it suffices to show that for any vector space V and any subspace U, the quotient map $Q : V \to V/U$ splits. Let B be a basis for U and let B' denote a completion of this basis to a basis for V (once again, using Theorem 1.2.1). We saw in the proof of Proposition 1.4.6 that $\{[\mathbf{b}] \mid \mathbf{b} \in B' \backslash B\}$ gives a basis for V/U and we define the map $Q' : V/U \to V$ by setting $Q'([\mathbf{b}]) = \mathbf{b}$ for each element of this basis and extending linearly. For each basis element $[\mathbf{b}]$, we have

$$Q \circ Q'([\mathbf{b}]) = Q(\mathbf{b}) = [\mathbf{b}].$$

Therefore $Q \circ Q'$ is the identity map on basis vectors and it follows easily that it is equal to the identity map in general. □

Cokernels

Let $L : V \to W$ be a linear map of finite-dimensional vector spaces. There is a vector space associated to L, called the *cokernel* of L and denoted $\operatorname{coker}(L)$. The cokernel is

defined to be the vector space

$$\text{coker}(L) = W/\text{image}(L).$$

Since we have already shown that $\text{image}(L)$ is a subspace of W, it follows immediately that $\text{coker}(L)$ has a vector space structure.

1.5 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.

1.5.1 Inner Products

An *inner product* on V is a map

$$\begin{aligned}\langle \cdot, \cdot \rangle : V \times V &\to \mathbb{R} \\ (\mathbf{v}_1, \mathbf{v}_2) &\mapsto \langle \mathbf{v}_1, \mathbf{v}_2 \rangle\end{aligned}$$

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

1. (Positive-Definiteness) $\langle \mathbf{v}, \mathbf{v} \rangle \geqslant 0$ and equality holds if and only if $\mathbf{v}$ is equal to 0_V,
2. (Symmetry) $\langle \mathbf{v}, \mathbf{w} \rangle = \langle \mathbf{w}, \mathbf{v} \rangle$,
3. (Bilinearity) $\langle \mathbf{u} + \mathbf{v}, \mathbf{w} \rangle = \langle \mathbf{u}, \mathbf{w} \rangle + \langle \mathbf{v}, \mathbf{w} \rangle$ and $\langle \lambda \mathbf{v}, \mathbf{w} \rangle = \lambda \langle \mathbf{v}, \mathbf{w} \rangle$. It follows from the symmetry property that $\langle \mathbf{v}, \mathbf{u} + \mathbf{w} \rangle = \langle \mathbf{v}, \mathbf{u} \rangle + \langle \mathbf{v}, \mathbf{w} \rangle$ and $\langle \mathbf{v}, \lambda \mathbf{w} \rangle = \lambda \langle \mathbf{v}, \mathbf{w} \rangle$.

Example 1.5.1. 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 1.5.2. As usual, we think of elements of $\mathbb{R}^n$ as n-tuples of real numbers, or as $1 \times n$-dimensional matrices with real entries. Then the column vector representation of $\mathbf{v} \in \mathbb{R}^n$ with respect to the standard basis B can be expressed as ${}_B\mathbf{v} = \mathbf{v}^T$, where the superscript T denotes matrix transpose. Notice that the dot product can be expressed in matrix form as $\mathbf{v}\mathbf{w}^T$ for any $\mathbf{v}, \mathbf{w} \in \mathbb{R}^n$. More generally, let M be an $n \times n$ matrix with real entries. Then the map

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

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. $\mathbf{v} \cdot M \cdot \mathbf{v}^T > 0$ for all $\mathbf{v} \neq \vec{0}$.

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 1.5.3. 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 1.5.1. *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?

1.5.2 Norms

A *norm* on V is a map

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

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

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

There is one immediate source of norms on V.

Proposition 1.5.2. *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 1.5.3 (Cauchy-Schwarz Inequality)**.** *For any inner product $\langle\cdot,\cdot\rangle$ and any $\boldsymbol{u}, \boldsymbol{v} \in V$,*

$$|\langle \boldsymbol{u}, \boldsymbol{v}\rangle| \leqslant \langle \boldsymbol{u}, \boldsymbol{u}\rangle \langle \boldsymbol{v}, \boldsymbol{v}\rangle .$$

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

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

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

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

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

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

and taking square roots proves the result. □

Example 1.5.4. 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 $\mathbf{v} = (a_1, \ldots, a_n) \in \mathbb{R}^n$ by the formula

$$\|\mathbf{v}\|_p = \left(|\mathbf{v}_1|^p + \cdots + |\mathbf{v}_n|^p \right)^{1/p}.$$

For $p = \infty$, define

$$\|\mathbf{v}\|_\infty = \max_i |\mathbf{v}_i|.$$

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

$$\|\mathbf{v}\|_2 = \langle \mathbf{v}, \mathbf{v} \rangle.$$

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.*

1.6 Exercises

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

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

$$-\mathbf{v} = -1 \cdot \mathbf{v}.$$

3. Define a "vector addition" operation $\widehat{+}$ on the set $V = \mathbb{R}$ by $v \widehat{+} w = \max\{v, w\}$. We also define a scalar multiplication operation by $\alpha \cdot v = \alpha v$ (i.e., the usual multiplication of real numbers). Determine which of the 8 vector space axioms the triple $(V, \widehat{+}, \cdot)$ obeys.

4. Let $\mathbb{R}^{m\times n}$ denote the set of all $m \times n$ matrices with real coefficients. Show that $\mathbb{R}^{m\times n}$ is a vector space with the following operations:

$$\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix} + \begin{pmatrix} b_{11} & b_{12} & \cdots & b_{1n} \\ b_{21} & b_{22} & \cdots & b_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ b_{m1} & b_{m2} & \cdots & b_{mn} \end{pmatrix}$$
$$= \begin{pmatrix} a_{11}+b_{11} & a_{12}+b_{12} & \cdots & a_{1n}+b_{1n} \\ a_{21}+b_{21} & a_{22}+b_{22} & \cdots & a_{2n}+b_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{m1}+b_{m1} & a_{m2}+b_{m2} & \cdots & a_{mn}+b_{mn} \end{pmatrix},$$

$$\lambda \cdot \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix} = \begin{pmatrix} \lambda a_{11} & \lambda a_{12} & \cdots & \lambda a_{1n} \\ \lambda a_{21} & \lambda a_{22} & \cdots & \lambda a_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ \lambda a_{m1} & \lambda a_{m2} & \cdots & \lambda a_{mn} \end{pmatrix}$$

5. Let $B_1 = \{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$ denote the standard basis for $\mathbb{R}^3$. Let

$$B_2 = \{\mathbf{v}_1 = \mathbf{e}_1 + \mathbf{e}_2, \mathbf{v}_2 = \mathbf{e}_2, \mathbf{v}_3 = \mathbf{e}_3\}$$

and

$$B_3 = \{\mathbf{w}_1 = \mathbf{e}_1, \mathbf{w}_2 = \mathbf{e}_2 - \mathbf{e}_3, \mathbf{w}_3 = \mathbf{e}_3\}.$$

It is a fact (which you do not need to verify!) that B_2 and B_3 also give bases for $\mathbb{R}^3$. Consider the linear map $L : \mathbb{R}^3 \to \mathbb{R}^3$ defined on the standard basis vectors by

$$L(\mathbf{e}_1) = \mathbf{e}_1 + \mathbf{e}_2, \quad L(\mathbf{e}_2) = \mathbf{e}_2, \quad L(\mathbf{e}_3) = \mathbf{e}_3.$$

Compute the following matrix representations of the map L:

a) ${}_{B_1}L_{B_1}$

b) ${}_{B_2}L_{B_1}$

c) ${}_{B_3}L_{B_1}$

6. Let B_1, B_2 and B_3 denote the bases for $\mathbb{R}^3$ defined in the previous problem. Let $I : \mathbb{R}^3 \to \mathbb{R}^3$ denote the *identity linear map* defined by $I(\mathbf{v}) = \mathbf{v}$ for every $\mathbf{v} \in \mathbb{R}^3$.
   a) Explain why ${}_{B_i}I_{B_i}$ is the 3×3 *identity matrix* for each $i = 1, 2, 3$. (The identity matrix is the diagonal matrix of all 1's).
   b) Compute the matrices ${}_{B_2}I_{B_1}$ and ${}_{B_1}I_{B_3}$.
   c) Compute the matrix product ${}_{B_1}I_{B_3} \cdot {}_{B_3}L_{B_1}$ and show that the result is equal to the matrix ${}_{B_1}L_{B_1}$ from the previous problem. We conclude that matrix multiplication by ${}_{B_1}I_{B_3}$ has the effect of changing the matrix representation of L from ${}_{B_3}L_{B_1}$ to ${}_{B_1}L_{B_1}$. In general, matrices of the form ${}_{B}I_{C}$ are called *change of basis matrices*.

7. Let V and W be vector spaces over $\mathbb{R}$ with $\dim(V) = n$ and $\dim(W) = m$, and let $\mathrm{Lin}(V, W)$ denote the set of all linear transformations $L : V \to W$.
   a) For linear maps $L, M \in \mathrm{Lin}(V, W)$, we define the linear map $L + M$ by
   $$(L + M)(\mathbf{v}) = L(\mathbf{v}) + M(\mathbf{v}).$$
   For a scalar $\lambda \in \mathbb{F}$, we define the linear map λL by
   $$(\lambda L)(\mathbf{v}) = \lambda L(\mathbf{v}) = L(\lambda \mathbf{v}).$$
   Show that these operations turn $\mathrm{Lin}(V, W)$ into a vector space.
   b) Let $B = \{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ be a basis for V and let $C = \{\mathbf{w}_1, \ldots, \mathbf{w}_m\}$ be a basis for W. Explain how these bases allow us to choose a basis for $\mathrm{Lin}(V, W)$. Use this to compute the dimension of $\mathrm{Lin}(V, W)$.
   Hint: Consider the set of maps δ_{ij} defined by
   $$\delta_{ij}(\mathbf{v}_i) = \mathbf{w}_j, \quad \text{and} \quad \delta_{ij}(\mathbf{v}_k) = 0 \ \forall \ k \neq i.$$
   c) Show that the map $F : \mathrm{Lin}(V, W) \to \mathbb{R}^{m \times n}$ defined by
   $$F(L) = {}_{C}L_{B}$$
   is a linear isomorphism. For the basis elements δ_{ij} defined above, what is $F(\delta_{ij})$?

8. Show that $P_n(\mathbb{R})$ is a vector space over $\mathbb{R}$. (See Example 1.1.5.)

9. Show that $C^\infty([0, 1], \mathbb{R})$ is a vector space over $\mathbb{R}$ (See Example 1.1.7.) You can assume basic facts from calculus.

10. Show that F_2 is a field. (See Example **??**.)

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

12. Prove that the set $\{\sin(k\pi t) \mid k \in \mathbb{Z}_{>0}\} \subset C^\infty([0,1],\mathbb{R})$ is linearly independent. Use this to conclude that $C^\infty([0,1],\mathbb{R})$ is infinite-dimensional (you will need to apply a theorem to do so!).

13. Show that the derivative map defined in Exercise 1.3.4 is a linear transformation between vector spaces.

14. 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 $C^0([0,1],\mathbb{R})$.

15. Let $V = \mathbb{R}^3$. Show that the set $W = \{(x,y,z) \mid z = 0\}$ is a subspace of V.

16. Prove Proposition 1.4.2.

17. Complete the proof of Proposition 1.4.5. This can be broken down into several steps: Let V be a vector space over $\mathbb{F}$ and $W \subset V$ a subspace of V.

    a) Define a relation $\sim$ on V by $\mathbf{v}_1 \sim \mathbf{v}_2$ if and only if $\mathbf{v}_1 = \mathbf{v}_2 + \mathbf{w}$ for some $\mathbf{w} \in W$. Show that this defines an *equivalence relation* on V.

    b) We define the *quotient space* V/W to be the set of equivalence classes

    $$V/W = \{[\mathbf{v}] \mid \mathbf{v} \in V\}.$$

    We define addition on V/W by

    $$[\mathbf{v}_1] + [\mathbf{v}_2] = [\mathbf{v}_1 + \mathbf{v}_2].$$

    First show that this operation is *well-defined*. This means that you must show that if $\mathbf{v}_1'$ and $\mathbf{v}_2'$ are vectors such that $[\mathbf{v}_1'] = [\mathbf{v}_1]$ and $[\mathbf{v}_2'] = [\mathbf{v}_2]$, then $[\mathbf{v}_1' + \mathbf{v}_2'] = [\mathbf{v}_1 + \mathbf{v}_2]$.

    c) Show that this addition operation is associative and commutative.

    d) We define the zero vector $0_{V/W} \in V/W$ to be the element

    $$0_{V/W} = [0_V].$$

    Show that $0_{V/W}$ is an additive identity in V/W. Also show that every element of V/W has an additive inverse.

    e) We define scalar multiplication in V/W as follows. For $\lambda \in \mathbb{F}$ and $[\mathbf{v}] \in V/W$, let

    $$\lambda \cdot [\mathbf{v}] = [\lambda \cdot \mathbf{v}].$$

    Show that this operation is well-defined.

    f) Show that scalar multiplication is associative and that there is an identity element for scalar multiplication.

    g) Prove that the two distributive laws in the vector space axioms hold for vector addition and scalar multiplication in V/W.

We conclude that V/W is a vector space!

18. Let $V = \mathbb{R}^3$ and let $W = \{(x, y, z) \mid z = 0\}$ (which we have already shown to be a subspace of V!). Prove that V/W is linearly isomorphic to $\mathbb{R}$.

19. Show that the $\langle \cdot, \cdot \rangle_{L^2}$ defines an inner product on $C^\infty([0, 1], \mathbb{R})$ (see Example 1.5.3).

20. Let $P_n(\mathbb{R})$ denote the vector space of degree (at most) n polynomials with real coefficients, considered as real-valued functions. Let $X = \{x_0, x_1, \ldots, x_m\}$ denote an arbitrary collection of $(m + 1) \geqslant 1$ real numbers. We define a product $\langle \cdot, \cdot \rangle_X$ on $P_n(\mathbb{R})$ by the formula

$$\langle p, q \rangle_X = \sum_{i=0}^{m} p(x_i) \cdot q(x_i),$$

for any $p, q \in P_n(\mathbb{R})$.

a) Show that this product is symmetric and bilinear.

b) Show that if $m \geqslant n$ and all x_i are distinct, then this product is also positive-definite. Conclude that $\langle \cdot, \cdot \rangle_X$ defines an inner product on $P_n(\mathbb{R})$ in this case.

Hint: You may assume the following algebraic fact: a real, degree (at most) n polynomial which is not the constant zero function has at most n distinct roots.

c) Show that the product is not necessarily positive-definite when $m < n$ by finding a counterexample.

21. Let $V = C^\infty([0, 1], \mathbb{R}^n)$ denote the set of smooth (infinitely differentiable) functions $f : [0, 1] \to \mathbb{R}^n$. The set V is a vector space with vector addition and scalar multiplication defined pointwise (i.e., $(f+g)(t) = f(t)+g(t)$ and $(\lambda \cdot f)(t) = \lambda f(t)$). Show that the set

$$U = \{f \in V \mid f(0) = \mathbf{0}\}$$

is a linear subspace of V and prove that the quotient space V/U is isomorphic to $\mathbb{R}^n$.

22. For this problem, we fix the standard basis $B = \{\mathbf{e}_1, \ldots, \mathbf{e}_n\}$ for $\mathbb{R}^n$ and consider all vectors $\mathbf{v} \in \mathbb{R}^n$ to be written as column vectors with respect to B. That is, we shorten notation and write $\mathbf{v} = {}_B\mathbf{v}$. We will use superscript T to denote matrix transpose. Recall that the standard inner product $\langle \cdot, \cdot \rangle_{\mathbb{R}^n}$ (i.e., the dot product) in $\mathbb{R}^n$ is given for vectors $\mathbf{v} = (v_1, v_2, \ldots, v_n)^T$ and $\mathbf{w} = (w_1, w_2, \ldots, w_n)^T$ by

$$\langle \mathbf{v}, \mathbf{w} \rangle_{\mathbb{R}^n} = v_1 w_1 + v_2 w_2 + \cdots + v_n w_n = \mathbf{v}^T \cdot \mathbf{w}.$$

More generally, for any $n \times n$ matrix A, we define a pairing $\langle \cdot, \cdot \rangle_A$ by the formula

$$\langle \mathbf{v}, \mathbf{w} \rangle_A = \mathbf{v}^T \cdot A \cdot \mathbf{w}.$$

Note that we can represent the standard inner product as $\langle \cdot, \cdot \rangle_I$ in this notation, where I is the $n \times n$ identity matrix.

a) A pairing $\langle \cdot, \cdot \rangle : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ is called *bilinear* if for all vectors $\mathbf{u}, \mathbf{v}$ and $\mathbf{w} \in \mathbb{R}^n$ and for every scalar λ,

$$\langle \mathbf{u} + \mathbf{v}, \mathbf{w} \rangle = \langle \mathbf{u}, \mathbf{w} \rangle + \langle \mathbf{v}, \mathbf{w} \rangle , \quad \langle \mathbf{v}, \mathbf{u} + \mathbf{w} \rangle = \langle \mathbf{v}, \mathbf{u} \rangle + \langle \mathbf{v}, \mathbf{w} \rangle$$

and

$$\langle \lambda \mathbf{v}, \mathbf{w} \rangle = \lambda \langle \mathbf{v}, \mathbf{w} \rangle = \langle \mathbf{v}, \lambda \mathbf{w} \rangle .$$

Show that any bilinear pairing is completely determined by its values $\langle \mathbf{v}_i, \mathbf{v}_j \rangle$ on any basis $B = \{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$.

b) Show that for any matrix A, the pairing $\langle \cdot, \cdot \rangle_A$ is bilinear. Conversely, show that for any bilinear pairing $\langle \cdot, \cdot \rangle$, there is some matrix A such that $\langle \cdot, \cdot \rangle = \langle \cdot, \cdot \rangle_A$. We conclude that there is a bijection

$$\{\text{bilinear pairings in } \mathbb{R}^n\} \leftrightarrow \{n \times n \text{ matrices}\}.$$

Hint: For the converse, consider the values $\langle \mathbf{e}_i, \mathbf{e}_j \rangle$ and use part (a).

c) Show that the bilinear pairing $\langle \cdot, \cdot \rangle_A$ is symmetric if and only if the matrix A is symmetric (i.e., $A^T = A$). We conclude that there is a bijection

$$\{\text{symmetric, bilinear pairings in } \mathbb{R}^n\} \leftrightarrow \{\text{symmetric } n \times n \text{ matrices}\}.$$

d) Show that the bilinear, symmetric pairing $\langle \cdot, \cdot \rangle_A$ is positive-definite if and only if all eigenvalues of the symmetric matrix A are positive. We conclude that there is a bijection

$$\{\text{inner products on } \mathbb{R}^n\} \leftrightarrow \{\text{symmetric } n \times n \text{ matrices with positive eigenvalues}\}.$$

Hint: For the "forward" direction, you can assume Schur's theorem, which states that for any symmetric matrix A, there is a basis $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ for $\mathbb{R}^n$ consisting of eigenvectors of A. Then consider the values $\langle \mathbf{v}_i, \mathbf{v}_i \rangle_A$. To prove the converse, you can further assume that a symmetric matrix is diagonalizable by orthogonal matrices. That is, $A = O^T D O$, where D is the diagonal matrix of (positive!) eigenvalues of A, and O is a matrix with the property that $O^T \cdot O$ is the $n \times n$ identity matrix. In particular, this means that O is invertible and therefore corresponds to a linear isomorphism.

2 Metric Space Topology

In applications, we begin with a set of datapoints, which 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 $\mathbf{v}, \mathbf{w} \in \mathbb{R}^n$ given by $\|\mathbf{v} - \mathbf{w}\|$, 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 [8].

2.1 Metric Spaces

2.1.1 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*.

2.1.2 Examples of Metric Spaces

Basic Examples

Example 2.1.1. 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 2.1.2. 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. Indeed,

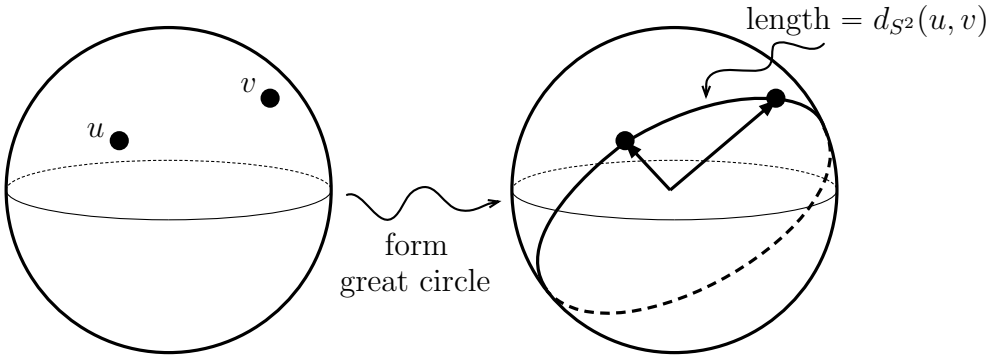
1. $d(v, w) = \|v - w\| \geqslant 0$, with equality if and only if $v - w = 0_V$, i.e., $v = w$,
2. $d(v, w) = \|v - w\| = \|w - v\| = d(w, v)$, and
3. for all $u, v, w \in V$, $d(u, w) = \|u - w\| = \|u - v + v - w\| \leqslant \|u - v\| + \|v - w\| = d(u, v) + d(v, w)$.

Example 2.1.3. Let X be any nonempty set. Define a metric d^δ on X by setting

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

You will show in the exercises that this function really defines a metric, called the *discrete metric* on X.

Example 2.1.4. 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.

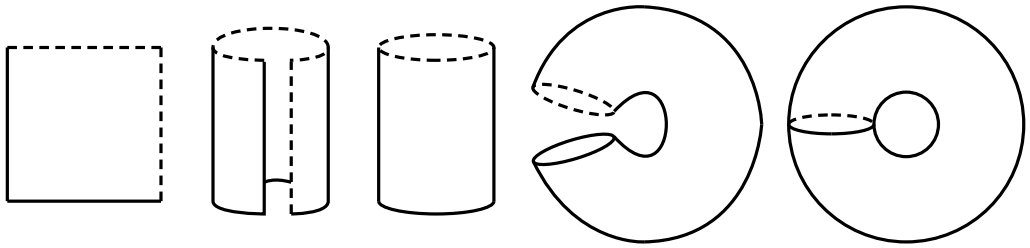


Example 2.1.5. Consider the square $T := [0,1) \times [0,1) \subset \mathbb{R}^2$. We define the distance d_T between points $p_1 = (x_1, y_1)$ and $p_2 = (x_2, y_2)$ in T by the formula

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

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.

Subspaces and Product Spaces

Example 2.1.6. A *metric subspace* of a metric space (X, d) is a metric space (Y, d_Y), where $Y \subset X$ and $d_Y = d|_{Y \times Y}$ is the *subspace metric* ; 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. The fact that d_Y is a metric follows immediately from the assumption that d is.

Example 2.1.7. Let (X, d_X) and (Y, d_Y) be metric spaces. The *product metric* $d_{X \times Y}$ on the set $X \times Y$ is defined by

$$d_{X \times Y}((x, y), (x', y')) = d_X(x, x') + d_Y(y, y').$$

It is straightforward to check that $d_{X \times Y}$ sastisfies the properties of a metric.

Example 2.1.8. As a specific example of the product metric construction, consider $\mathbb{R}$ with its standard metric. The product metric on $\mathbb{R} \times \mathbb{R}$ is given by

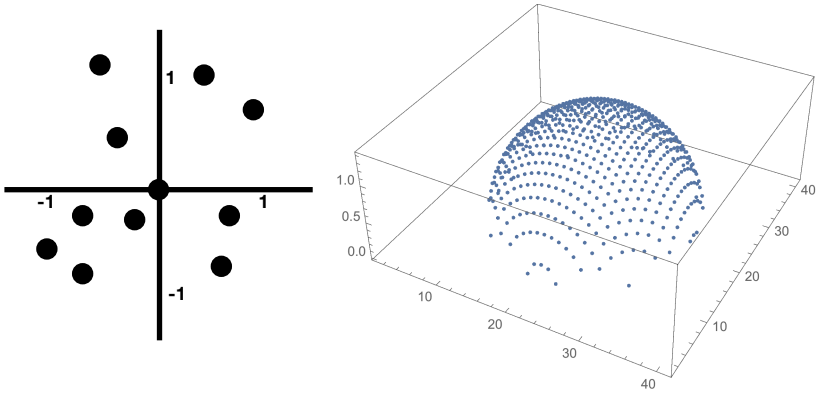
$$d_{\mathbb{R} \times \mathbb{R}}((x, y), (x', y')) = |x - x'| + |y - y'|,$$

and we see that the product metric is the same as the metric induced by the ℓ_1-norm.

Common Examples Arising in Data Analysis

Example 2.1.9. 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 2.1.10. 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 $v, w \in V$ and $v \neq w$. Graphs are realized geometrically by drawing the vertex set and joining vertices v and w by a line segment when $\{v, w\} \in E$. Figure 2.1 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".

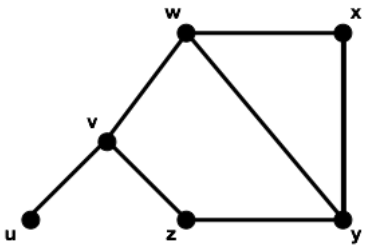


Figure 2.1: A graph.

A *path* in the graph $G = (V, E)$ between vertices $v, w \in V$ is a sequence of vertices $(v = v_1, v_2, \ldots, v_m = w)$ such that each $\{v_j, v_{j+1}\} \in E$. A graph is called *connected* if there is a path joining any pair of vertices and it is called a *tree* if there is a unique path. The *length* of a path is the number of edges in it; that is, the path $(v = v_1, v_2, \ldots, v_m = w)$ has length $m - 1$. For example, in Figure 2.1, (u, v, w, x) is a path of length 3 joining the vertices u and x, while the path $(u, v, z, y.w.x)$ has length 5. The graph in Figure 2.2 is a tree and the unique path joining v and w has length 3.

A connected 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 graph shown in Figure 2.1, $d_G(v, y) = 2$, because we could take either path (v, w, y) or (v, z, y) to join the vertices and there is no shorter path.

In a similar vein, an *oriented* graph is represented by a pair $G = (V, E)$ where V is as before and E is a subset of $V \times V$, with $(v_1, v_2) \in E$ indicating the existence of an oriented edge starting at v_1 and ending at v_2. A path in an oriented graph G is required to be consistent with the orientation of each edge; in other words, $(v_1, v_2, \ldots, v_m)$ is an oriented path iff $(v_i, v_{i+1}) \in E, 1 \leqslant i \leqslant (m-1)$. Associated to any oriented graph is an unoriented one derived by ignoring the orientation (or ordering) of each edge. One could similarly define a "graph distance" in the oriented graph setting. Such a "distance" would satisfy the reflexivity property and the triangle inequality from the definition of a metric, but we would run into the technical issues that not every pair of points can be joined by an oriented path (in which case we might declare the distance between the vertices to be ∞) and the symmetry property of a metric may not hold. The resulting "distance" is what is known as an *extended pseudo-metric* . An example is shown in figure 2.3

2.1.3 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

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

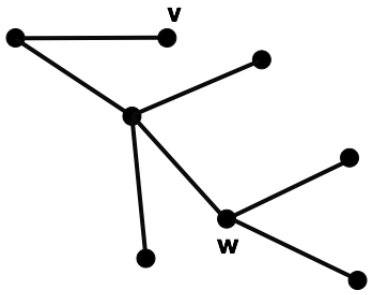


Figure 2.2: A tree.

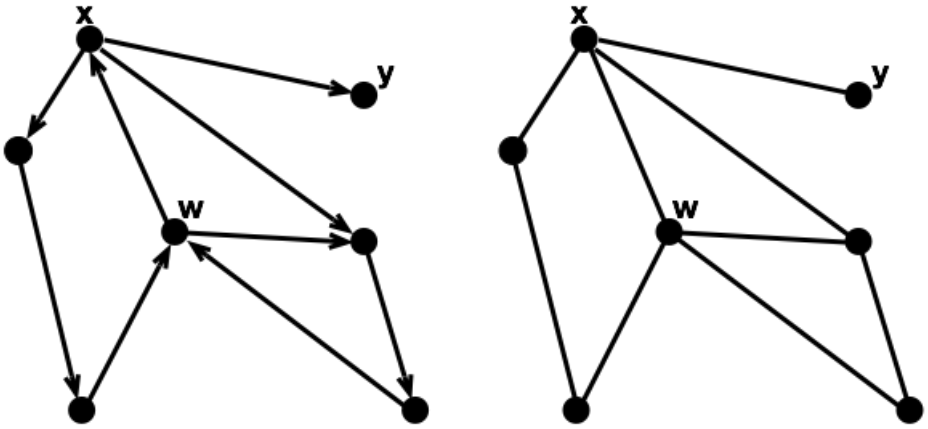


Figure 2.3: The figure on the left shows an oriented graph G, with edge orientation indicated by an arrow. The figure on the right shows its underlying (unoriented) graph G'. In the oriented graph G, we have $d_G(x, y) = 1$, $d_G(y, x) = \infty$, $d_G(x, w) = 3$ and $d_G(w, x) = 1$. In the underying graph G' we have $d_{G'}(x, y) = d_{G'}(y, x) = 1$ and $d_{G'}(x, w) = d_{G'}(w, x) = 1$.

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.

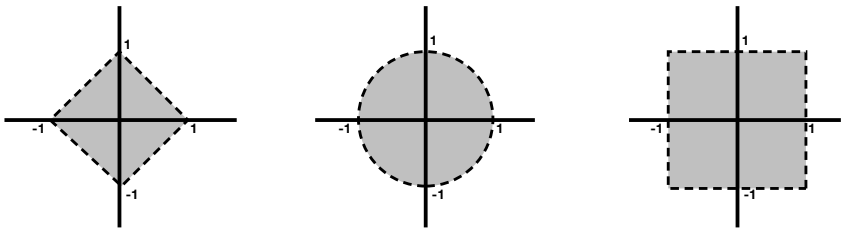
Example 2.1.11. For $\mathbb{R}$ with its standard metric, the open metric balls are open intervals.

Example 2.1.12. For a set X with the discrete metric, the open metric balls are of the form

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

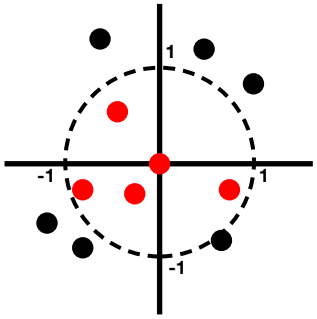
for all $x \in X$.

Example 2.1.13. 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.

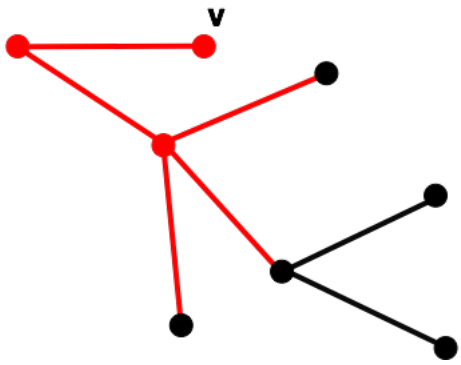


$B_{\ell_1}(\vec{0}, 1)$ $B_{\ell_2}(\vec{0}, 1)$ $B_{\ell_\infty}(\vec{0}, 1)$

Example 2.1.14. 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 2.1.15. The next example returns to the tree metric space T of Example 2.1.10. 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 2.1.1. *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 2.1.2. *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 $\mathcal{U} = \{U_1, \ldots, U_n\}$ 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 $i = 1, \ldots, n$ 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 2.1.16. 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 $\mathcal{U} = \{U_n\}_{n=1,2,3,\ldots}$ 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 (in fact, it is closed).

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

Proposition 2.1.3. *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 closed;

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 2.1.17. 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 $a < b$
- $\{1\}$
- $[0,1] \cup [2,3] \cup [4,5] \cup \cdots = \bigcup_{i=0}^{\infty}[2i, 2i+1]$. Here we have a closed set expressed as an infinite union of closed sets. Note that the union of infinitely many closed sets is *not* necessarily closed; this example is just a special case.
- Consider the collection of closed sets $[1/n, 1-1/n]$, with n a positive integer. The union of these sets is $(0,1)$, which is not closed!

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 2.4.2 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 2.1.4. *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 $V = Y \cap U$. □

2.2 Topological Spaces

2.2.1 Definition of a Topological Space

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 2.2.1. 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 [8].

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 2.2.2. 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).

2.2.2 Equivalence of Topological Spaces

Another reason that we introduce the more general notion of a topological space (rather than strictly dealing with metric spaces) is that different metrics on the same set can produce the same topology. More precisely, let $\mathcal{T}$ and $\mathcal{T}'$ be topologies on the same set X. We say that $\mathcal{T}$ is *finer* than $\mathcal{T}'$ if $\mathcal{T}' \subset \mathcal{T}$. Similarly, $\mathcal{T}$ is *coarser* than $\mathcal{T}'$ if $\mathcal{T} \subset \mathcal{T}'$. Finally, $\mathcal{T}$ is *equivalent to* $\mathcal{T}'$ if $\mathcal{T} = \mathcal{T}'$.

The following lemma gives a useful way to determine when one topology is finer than another.

Lemma 2.2.1. *Let $\mathcal{T}$ and $\mathcal{T}'$ be topologies on a set X. Then $\mathcal{T}' \subset \mathcal{T}$ if and only if for each $U \in \mathcal{T}'$ and for each $x \in U$, there is an open set $V \in \mathcal{T}$ such that $x \in V \subset U$.*

Proof. Suppose that $\mathcal{T}' \subset \mathcal{T}$ and let $U \in \mathcal{T}'$ and $x \in U$. Then $U \in \mathcal{T}$ by assumption and we obviously have $x \in U \subset U$. Conversely, suppose that the property holds and let $U \in \mathcal{T}'$. Then for each $x \in U$ there is an open set $V_x \in \mathcal{T}$ such that $x \in V_x \subset U$. Taking the union of all such V_x, we have

$$U = \bigcup_{x \in U} V_x$$

and the latter must be an element of $\mathcal{T}$, since the topology is closed under arbitrary intersections. □

This leads to the following characterization in the special case that the topologies are generated by a metric.

Corollary 2.2.2. *Let $\mathcal{T}$ and $\mathcal{T}'$ be metric topologies on a set X which are generated by metrics d and d', respectively. Then $\mathcal{T}' \subset \mathcal{T}$ if and only if for all $x \in X$ and for each $\epsilon > 0$, there exists $\delta > 0$ such that*

$$B_d(x, \delta) \subset B_{d'}(x, \epsilon).$$

Proof. It suffices to prove that the metric ball condition is equivalent to the subset condition of Lemma 2.2.1. Assume that the condition from the lemma holds and let $x \in X$ and $\epsilon > 0$. Then $B_{d'}(x, \epsilon)$ is an open set in $\mathcal{T}'$ containing x, and it follows by our assumption that there is some open set V in $\mathcal{T}$ such that

$$x \in V \subset B_{d'}(x, \epsilon).$$

Since $\mathcal{T}$ is a metric topology, there exists $\delta > 0$ such that

$$B_d(x, \delta) \subset V,$$

and our claim follows by combining this with the previous inclusion.

To prove the converse, suppose that the metric ball inclusion condition holds and let $U \in \mathcal{T}'$ and $x \in U$. Then there is some $\epsilon > 0$ such that

$$B_{d'}(x, \epsilon) \subset U.$$

The claim follows by taking $V = B_d(x, \delta)$, where $\delta > 0$ satisfies

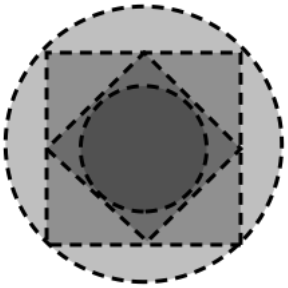
$$B_d(x, \delta) \subset B_{d'}(x, \epsilon).$$

□

Example 2.2.3. We have an infinite family $\{d_p\}$ of metrics $\mathbb{R}^n$ induced by the ℓ_p-norms. It turns out that, even though all of the metrics are distinct, they all induce the same topology on $\mathbb{R}^n$. You will prove this in the homework for the important cases of $p = 1, 2, \infty$. The idea of the proof is to show that there are constants c_1, c_2 and c_3 such that for any $x \in \mathbb{R}^n$ and for any $\epsilon > 0$, we have

$$B_{d_2}(x, \epsilon) \subset B_{d_1}(x, c_1\epsilon) \subset B_{d_\infty}(x, c_2\epsilon) \subset B_{d_2}(x, c_3\epsilon).$$

It then follows that the topologies are equivalent by Corollary 2.2.2. This idea is illustrated by the following figure and details are given in the exercises.



2.2.3 Basis for a Topology

A useful way of describing topology is via a basis, which is a subset of the topology satisfying certain conditions described in the following proposition.

Proposition 2.2.3. *Let $(X, \mathcal{T})$ be a topological space and let $\mathcal{B} \subset \mathcal{T}$. The following are equivalent:*

1. *Every open set $U \in \mathcal{T}$ can be expressed as a union $U = \bigcup_\alpha V_\alpha$ of some collection $\{V_\alpha\}$ of elements of $\mathcal{B}$;*
2. *For each open set $U \in \mathcal{T}$, for each $x \in U$, there exists $V \in \mathcal{B}$ such that $x \in V \subset U$.*

We leave the proof of the proposition as an exercise. If $\mathcal{B}$ satisfies the conditions of the proposition then it is called a *basis* for $\mathcal{T}$. A basis is called *minimal* if for every proper subset $\mathcal{B}' \subset \mathcal{B}$, $\mathcal{B}'$ is not a basis.

Example 2.2.4. For any metric space (X, d), a basis for the metric topology is given by

$$\mathcal{B} = \{B_d(x, r) \mid x \in X \text{ and } r \in \mathbb{R}_{>0}\}.$$

Indeed, this follows immediately from the definition the metric topology!

2.2.4 Subspaces and Products

Subspaces

Let X be a topological space and $Y \subset X$ a subset. The *subspace topology* on Y is the topology whose open sets are of the form $U \cap Y$, where U is open in X. This is indeed a topology on Y:

1. $\varnothing = \varnothing \cap Y$ and $Y = X \cap Y$;
2. for any collection of open sets $\{U_\alpha \cap Y\}$,
$$\bigcup_\alpha (U_\alpha \cap Y) = Y \cap \bigcup_\alpha U_\alpha;$$
3. for a finite collection $\{U_1 \cap Y, U_2 \cap Y, \ldots, U_n \cap Y\}$,
$$\bigcap_j (U_j \cap Y) = Y \cap \bigcap_j U_j.$$

We leave the proof of the following as an exercise.

Proposition 2.2.4. *Let (X, d) be a metric space and $Y \subset X$ a subset. Then the subspace topology with respect to the metric topology on X is equivalent to the metric topology on Y with respect to the subspace metric.*

Product Spaces

Let X and Y be topological spaces. The *product topology* on the set $X \times Y$ is the topology generated by the basis

$$\{U \times V \mid U \text{ is open in X and } V \text{ is open in } Y\}.$$

It is straightforward to check that this defines a topology and we leave the following proof as an exercise.

Proposition 2.2.5. *Let (X, d_X) and (Y, d_Y) be metric spaces. Then the product topology of the metric topologies on X and Y is equivalent to the metric topology for the product metric $d_{X\times Y}$.*

Remark 2.2.6. *For a finite collection $\{X_1, \ldots, X_n\}$ of topological spaces, we can define the product topology on $X_1 \times X_2 \times \cdots \times X_n$ to be the topology with basis consisting of sets of the form $U_1 \times U_2 \times \cdots \times U_n$ with U_j open in X_j. This definition continues to makes sense even for an infinite collection of topological spaces, but passing to infinite products results in many technical issues. We do not wish to treat the issue in depth here, but the reader should be aware that the term* product topology *generally refers to a coarser topology than this one when dealing with infinite products.*

2.2.5 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 $\text{int}(Y)$.

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 2.2.7. *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. □

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 2.2.8. *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

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

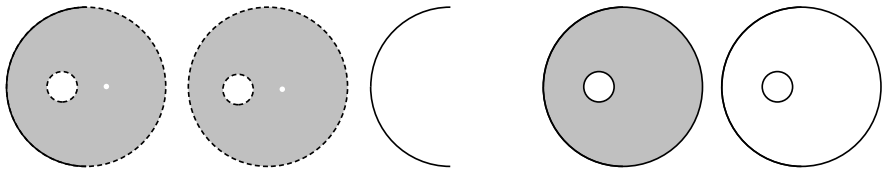
First note that $\overline{Y}$ is a closed set which contains Y, so it must be that $Z \subset \overline{Y}$. It remains to show that $\overline{Y} \subset Z$. Let $y \in \overline{Y}$. If $y \in Y$, then y is an element of each set in the intersection defining Z, so it is an element of Z and we are 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 2.2.7 that $y \in C$. Since C was arbitrary, it must be that $y \in Z$. □

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 2.2.5. 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 $\text{int}(Y)$, boundary ∂Y, closure $\overline{Y}$ and the boundary of its closure $\partial\overline{Y}$.



2.3 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 2.3.1. *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. □

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

Notice that the the first property in the proposition only depends on the open sets of the spaces X and Y. Inspired by this, we define a function $f : X \to Y$ between topological spaces (which are not necessarily metric spaces!) to be *continuous* if for every open set $U \subset Y$, the set $f^{-1}(U)$ is open in X.

Example 2.3.1. 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 sense of metric spaces or topological spaces.

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

Lemma 2.3.2. *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.*

2.4 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.

2.4.1 Compactness

An *open cover* of a metric space (X, d) is a collection $\mathcal{U}$ of open sets such that $\bigcup_{U \in \mathcal{U}} U = X$. 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 2.4.1. 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 $y_j \in U_j$ 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 2.4.2. 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 2.4.3. 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 2.4.1. *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

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

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 $c \in [0, 1]$ (as $c = 1$ is an upper bound on C), so it must be contained in some open set $A \in \mathcal{A}$. Then $(c - \epsilon, c] \subset A$ for sufficiently small $\epsilon > 0$. Moreover, for sufficiently small $\epsilon \in (0, c]$, $c - \epsilon \in C$; otherwise there is an interval $(c - \epsilon, c) \notin C$ and any point in this interval must upper bound C, violating our definition of c. We write $[0, c] = [0, c - \epsilon] \cup (c - \epsilon, c]$, and note that our assumptions imply that $[0, c - \epsilon]$ is contained in a finite subcover of $\mathcal{A}$ and it follows that we have that $[0, c]$ is contained in a finite 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 2.4.2. *Let $f : X \to Y$ be a continuous map of topological 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 2.4.3. *A finite product of compact topological spaces is also compact.*

Proof. By induction, it suffices to prove that the product of two compact spaces $X \times Y$ is compact. We sketch the proof here and leave the details as an exercise. Let $\mathcal{U}$ be an open cover of $X \times Y$. For each $x \in X$, the slice $\{x\} \times Y$ is homeomorphic to Y and hence compact. It can therefore be covered by finitely many elements of $\mathcal{U}$; let N denote the union of these finitely many elements. We make the following claim, whose proof is left as an exercise.

Claim: For any open neighborhood N in $X \times Y$ of any slice $\{x\} \times Y$, there is an open set $V \subset X$ such that $x \in V$ and $V \times Y \subset N$.

Assuming the claim, we choose an open neighborhood N_x as above for each $x \in X$, then choose an open neighborhood V_x with $x \in V_x$ and such that $V_x \times Y \subset N$. The collection $\{V_x\}$ covers X, so it admits a finite subcover $\{V_{x_1}, V_{x_2}, \ldots, V_{x_n}\}$. We then obtain a finite subcover of $\mathcal{U}$ by including the finitely many constituent open sets of each N_{x_j}. □

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 2.4.4 (Extreme Value Theorem)**.** *Let (X, d) be a compact metric space. Any continuous function $X \to \mathbb{R}$ achieves its maximum and minimum values; that is, there exist $c, d \in X$ such that for any $x \in X$, $f(c) \leqslant f(x) \leqslant f(d)$.*

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

Theorem 2.4.5 (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.*

2.4.2 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 2.4.6. *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. □

A fundamental theorem from calculus follows immediately and in much greater generality.

Theorem 2.4.7 (The Intermediate Value Theorem)**.** *Let X be a connected space and let $f : X \to \mathbb{R}$ be a continuous map. If $a, b \in \mathbb{R}$, say with $a < b$, are in the image of f then so is any point $c \in (a, b)$.*

Proof. Suppose that this is not the case and that there is some $c \in (a, b)$ which is not in the image of f. Let $U = f^{-1}(-\infty, c)$ and $V = f^{-1}(c, \infty)$. Since f is continuous, U and V are open subsets of X. Moreover, U, V are nonempty and not all of X, since the points $a \in (-\infty, c)$ and $b \in (c, \infty)$ lie in the image of f. The union $U \subset V$ covers X, since c is not in the image of f. Finally, $U \cap V \subset f^{-1}((-\infty, c) \cap (c, \infty)) = \varnothing$, so it follows that U, V give a separation of X and we have obtained a contradiction. □

Theorem 2.4.8. *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 2.4.3 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 2.4.4. 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. □

Corollary 2.4.9. *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. □

The following is a useful alternate characterization of connectedness.

Proposition 2.4.10. *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 $Y = X$. □

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 2.4.11. *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. □

It is generally much easier to prove that a space is connected by showing that it is path-connected than it is to show connectedness directly from the definition. Consider the following examples.

Example 2.4.4. For any n, $\mathbb{R}^n$ is connected. Indeed, for any $x, y \in \mathbb{R}^n$, the function $\gamma : [0, 1] \to \mathbb{R}^n$ defined by

$$\gamma(t) = (1 - t)x + ty$$

(using scalar multiplication and vector addition) gives a path between the points.

Example 2.4.5. For any n, $\mathbb{R}^n \backslash \{0\}$ is connected. For any $x, y \in \mathbb{R}^n$, if the path defined in the previous example does not pass through the origin, then we are done. If it does, the path can be modified by picking a third point $z \in \mathbb{R}^n$ which does not lie along the line through and x and y and concatenating the path from x to z with the path from z to y.

Example 2.4.6. The unit sphere $S^n \subset \mathbb{R}^{n+1}$ is connected. For any $x, y \in S^n$ there is a path γ. in $\mathbb{R}^n \backslash \{0\}$ joining x to y. We modify this path to to get a path η which lies in the sphere via the formula

$$\eta(t) = \frac{\gamma(t)}{\|\gamma(t)\|}.$$

2.5 Equivalence Relations

2.5.1 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 2.5.1. *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. □

Proposition 2.5.2. *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 2.5.1. 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

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

2.5.2 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 2.5.3. *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 2.3.2. 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 2.5.2. 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).

2.5.3 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 2.5.4. *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 2.5.3. 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 $f \circ g$ is equal to the identity map on Y, so there is nothing to prove here. On the other hand $g \circ f : X \to X$ 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.

2.6 Exercises

1. Let $\mathbb{R}$ denote the real numbers endowed with the *standard metric topology*. This is the topology which defines a nonempty set $S \subset \mathbb{R}$ to be open if and only if for each $p \in S$ there exists $\epsilon > 0$ such that the metric ball $B_\epsilon(p)$ is contained in S. Prove that S is open if and only if for each $p \in S$, there is an open interval $(b,c) \subset \mathbb{R}$ such that $p \in (b,c) \subset S$.

2. Show that for any set X, the discrete metric d^δ defined in Example 2.1.3 is really a metric.

3. Let X be a set and let d and d' be metrics on X. Show that the respective metric topologies on X are equivalent if and only if for each $x \in X$ and each $\epsilon > 0$, there

exist $\epsilon_1, \epsilon_2 > 0$ such that

$$B_d(x, \epsilon_1) \subset B_{d'}(x, \epsilon) \quad \text{and} \quad B_{d'}(x, \epsilon_2) \subset B_d(x, \epsilon).$$

4. Let (X, d) be a metric space and let $Y \subset X$ be a subset. Let d_Y denote the subspace metric on Y. Then there are two natural topologies on Y: the subspace topology inherited from the metric topology on X and the metric topology for the subspace metric d_Y. Show that these two topologies are equivalent.

5. Let (X, d_X) and (Y, d_Y) be metric spaces and let $d_{X\times Y}$ denote the product metric on the set $X \times Y$. There are two natural topologies on $X \times Y$: the product topology inherited from the metric topologies on X and Y and the metric topology for the product metric $d_{X\times Y}$. Show that these two topologies are equivalent.

6. Prove Proposition 2.1.3.

7. Show that the topologies generated by d_2, d_1 and d_∞ on $\mathbb{R}^n$ are equivalent.

   Hint: Find constants $c_1, c_2, c_3 > 0$ such that for all $\mathbf{v} \in \mathbb{R}^n$ and $\epsilon > 0$,

   $$\|\mathbf{v}\|_1 \leqslant c_1\|\mathbf{v}\|_2 \leqslant c_2\|\mathbf{v}\|_\infty \leqslant c_3\|\mathbf{v}\|_1,$$

   then use the previous problem. Note that it is possible to find constants which do not depend on $\mathbf{v}$ (although they may depend on the dimension n).

8. Work out a more explicit representation of the function d_{S^2} defined in Example 2.1.4. 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.

9. Prove that the function d_{S^2} defined in Example 2.1.4 is a metric.

10. Prove that the function d_T defined in Example 2.1.5 is a metric.

11. Prove Proposition 2.2.3.

12. Let d denote the standard metric on $\mathbb{R}$. Show that the set

    $$\mathcal{B} = \{B_d(x, r) \mid x \in \mathbb{Q} \text{ and } r \in \mathbb{Q}_{>0}\}$$

    gives a basis for the standard topology on $\mathbb{R}$ (where $\mathbb{Q}$ denotes the rational numbers and $\mathbb{Q}_{>0}$ is the set of positive rational numbers). Show that this basis is not minimal.

13. Find a minimal basis for $\mathbb{R}^\delta$.

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

15. 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) $([0,1] \times [0,1]) \backslash ((1/4, 3/4) \times (1/4, 3/4))$

16. Prove Lemma 2.3.2.

17. Show that if $\phi : T \to U$ is a homeomorphism of topological spaces and $T' \subset T$ is a subspace (with the subspace topology), then the restricted map $\phi|_{T'} : T' \to U$ takes T' homeomorphically onto the subspace $\phi(T') \subset U$.

18. Complete the proof of Proposition 2.4.3 by proving the claim which appears in the proof sketch. This is commonly referred to as the *Tube Lemma*.

19. Let X be a set. We use X^δ to denote X endowed with the discrete topology and X^c to denote X endowed with the *coarse topology* (this is the topology on X given by $\{\varnothing, X\}$). Let Y be an arbitrary topological space.

    a) What are the continuous maps $f : X^\delta \to Y$?

    b) What are the continuous maps $f : X^c \to X^\delta$?

20. Show that if X is a compact space and C is a closed subset of X, then C is also compact.

21. Show that if X is a Hausdorff space, then any compact subset of X must be closed.

22. Prove the Heine-Borel Theorem, Theorem 2.4.5.

    Hint: Apply Theorem 2.4.1 and Proposition 2.4.3 to prove that any box of the form $[a_1, b_1] \times [a_2, b_2] \times \cdots \times [a_n, b_n]$ is compact. Any bounded set can be contained in such a rectangle, and one then applies Exercise 20. Conversely, try applying Exercise 21.

23. Prove the Extreme Value Theorem, Theorem 2.4.4.

    Hint: Let $A = f(X)$, which is compact by Proposition 2.4.2. Show that A has maximum and minimum values, perhaps by contradiction.

24. Prove that a subset $S \subset \mathbb{R}$ is path-connected if and only if it is an interval (by "interval", we mean bounded or unbounded, open, closed, or half-open; e.g., $(0, 1]$, $[2, \infty)$, $\mathbb{R}$ are all examples of intervals).

25. Prove Proposition 2.5.4.

26. a) Show that $\mathbb{R}^n \simeq \{0\}$.

    b) Show that $\mathbb{R}^n \backslash \{0\} \simeq S^{n-1}$, where S^{n-1} is the unit sphere in $\mathbb{R}^n$. (Hint: first show that the open (standard) metric ball $B_d(0, 2) \subset \mathbb{R}^n$ is homeomorphic to $\mathbb{R}^n$.)

    c) Show that the intervals $(0, 1)$ and $[0, 1]$ are homotopy equivalent.

27. Let $Y = \{a, b\}$ be a two-point set with the discrete topology. Consider the maps $f_0, f_1 : Y \to Y$ with f_0 the constant map $f_0(a) = f_0(b) = a$ and with f_1 the identity map. Show that f_0 is not homotopic to f_1. Use this to show that Y is not homotopy equivalent to a one-point space $X = \{x\}$.

3 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*.

3.1 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 3.1.1. 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 3.1.2. 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 3.1.3. Are the spaces $X = [0, 1)$ and $Y = (0, 1)$ homeomorphic?

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.

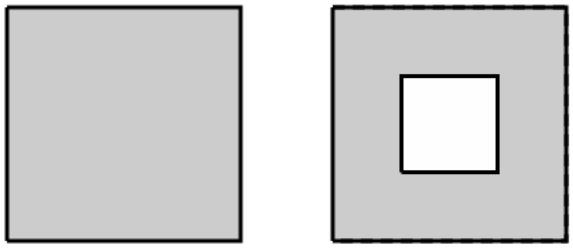


Figure 3.1: The spaces $X = [0,1) \times (0,1)$ and $Y = (0,1) \times (0,1)$ are shown in the top row. The second row shows each space with a point removed. For X, the point is chosen so that the resulting space has no "holes". For Y, any choice of point to remove results in hole in the space's interior.

Lemma 3.1.1. *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 3.1.4. 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 3.1)?

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 3.5.2), 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 3.1.4 with ease. You will do so in the exercises.

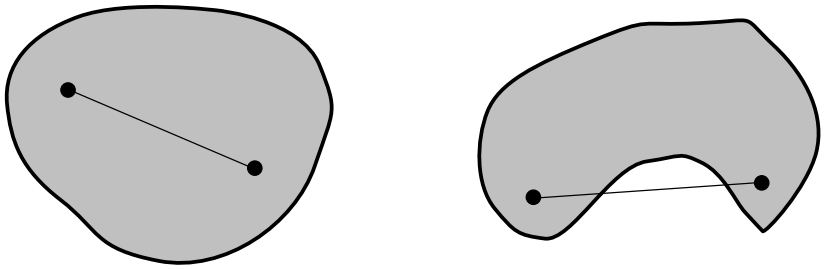
3.2 Simplicial Complexes

3.2.1 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 3.2.1. *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.*

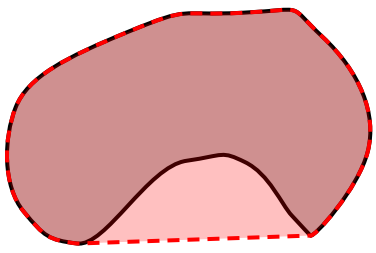


Convex Nonconvex

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.



There is another characterization of the convex hull when S is a finite set. Let $S = \{v_1, v_2, \ldots, v_n\}$. The *convex linear span* of S to is the set

$$\left\{ v = \sum_{j=1}^{n} t_j v_j \mid t_j \in [0,1] \text{ and } \sum_{j=1}^{n} t_j = 1 \right\}$$

We leave it as an exercise to show that when the subset S is finite, the convex hull and the convex linear span are the same set. We will primarily be interested in convex hulls of finite sets, so it is useful to have these alternate characterizations.

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).

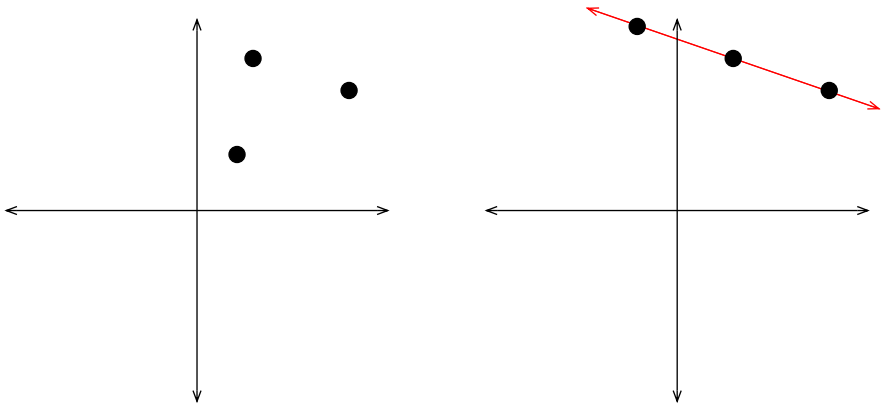


Figure 3.2: The figure on the left shows a set $\{x_0, x_1, x_2\}$ of 3 points in $\mathbb{R}^2$ which are in general position—any line can only contain 2 of the points. The figure on the right shows a set of points which are *not* in general position. The 1-dimensional affine subspace containing all 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.

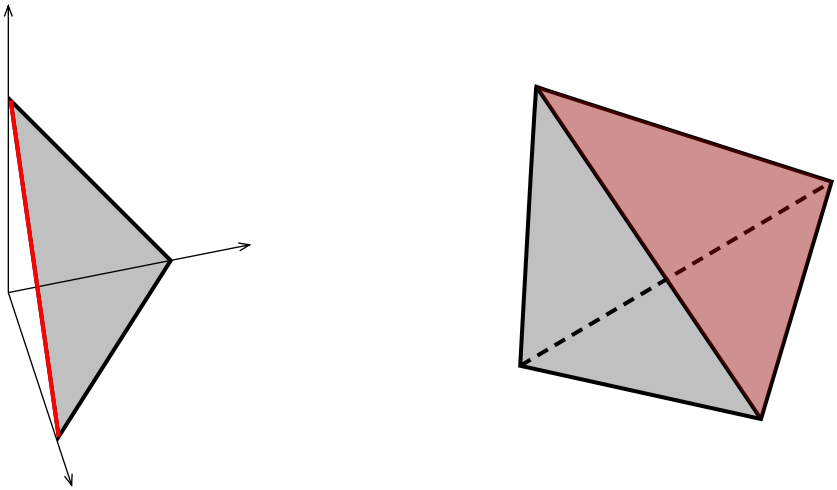
Example 3.2.1. It will be convenient to have a standard picture to refer to. The *standard n-dimensional simplex* is the simplex associated to the set

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

consisting of all points on the coordinate axes at Euclidean distance 1 from the origin. Equivalently, the standard n-simplex is the set

$$\{(x_0, x_1, \ldots, x_n) \in \mathbb{R}^{n+1} \mid 0 \leqslant x_i \leqslant 1 \text{ and } \sum_{i=0}^{n} x_i = 1\}.$$

In the figure below, the shape on the left is the standard 2-dimensional simplex with the 1-dimensional face (i.e., an edge) $\sigma((1,0,0),(0,0,1))$ highlighted. The shape on the right is a (nonstandard) 3-dimensional simplex embedded in $\mathbb{R}^3$ with one of its 2-dimensional faces highlighted.



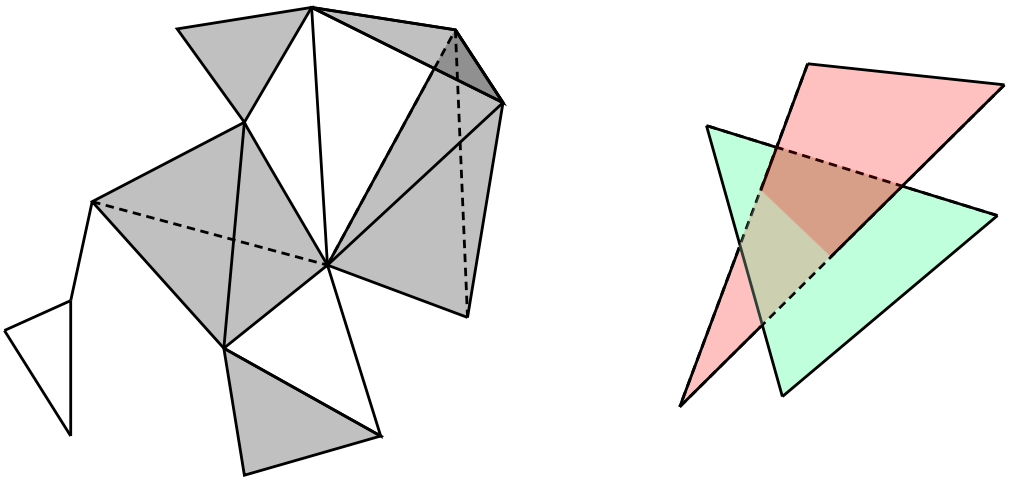
Simplicial Complexes

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 3.2.2. 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.



3.2.2 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. More generally, the metric structure of a simplicial complex induces a topological structure, and 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 3.2.3. 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 3.2.4. 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 3.2.1 (see the exercises)?

Simplicial Maps

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 3.2.3 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 3.2.2. *A geometric realization exists for any abstract simplicial complex* $X = (V(X), \Sigma(X))$.

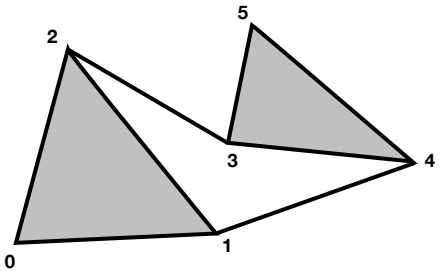
Proof. Assume $V(X)$ contains $n+1$ points. We embed the vertex set in $\mathbb{R}^{n+1}$ by mapping the elements of $V(X)$ to the vertices of the standard n-dimensional simplex. For each subset $\Sigma(X)$, we include the simplex in the geometric realization which is formed by taking the convex hull 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 3.2.5. 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 3.2.2.

Remark 3.2.3. *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 $\mathcal{X}$.*

k-Skeleta

A simplicial complex (either geometric or abstract) is stratified into collections of simplices of the same dimension. For a geometric simplicial complex $\mathcal{X}$, we define the *k-skeleton* of $\mathcal{X}$ to be the set

$$\mathcal{X}_k = \{\sigma \subset \mathcal{X} \mid \sigma \text{ is a } k\text{-simplex}\}.$$

Similarly, for an abstract simplicial complex $X = (V, \Sigma)$, we define the *k-skeleton* of X to be the set

$$X_k = \{\sigma \in \Sigma \mid \sigma = \{v_{i_0}, v_{i_1}, \ldots, v_{i_k}\}\}.$$

3.3 Topological Invariants of Simplicial Complexes

In this section we will study some examples of simplicial complexes in order to understand how we might pick out topological features using tools from linear algebra. Our discussion first requires a brief detour through a new concept from linear algebra.

3.3.1 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.

Definition of a Free Vector Space

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 3.3.1. *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. □

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

Basic Results on Free Vector Spaces

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

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

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

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

Proposition 3.3.3. *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. □

Notation

The formalism for free vector spaces described above is necessary to put the definition on firm footing, but is somewhat cumbersome in practice. Given a finite set $S = \{s_1, \ldots, s_n\}$, we will frequently drop the function notation and express elements of the free vector space $V_{\mathbb{F}}(S)$ as formal sums of the elements of S. That is, the vector

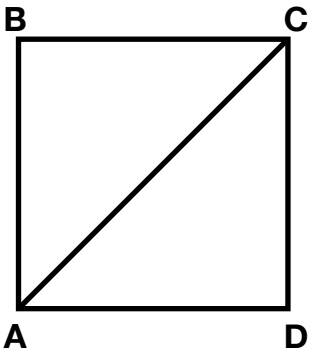
$$\lambda_1\phi_{s_1} + \lambda_2\phi_{s_2} + \cdots + \lambda_n\phi_{s_n}, \quad \lambda_j \in \mathbb{F}$$

will instead be written more succinctly as

$$\lambda_1 s_1 + \lambda_2 s_2 + \cdots + \lambda_n s_n.$$

3.3.2 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

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

and its edge set is

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

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 $C_0(X) := V_{F_2}(\text{Vert}(X))$, the free vector space over F_2 (the field with 2 elements) generated by $\text{Vert}(X)$. (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 3.3.1 and denoting the basis element of $C_1(X)$ assocated to the edge AB by ϕ_{AB}. In the simplified notation introduced in Section 3.3.1, this line is written as

$$\partial_1(AB) = A + B.$$

With respect to this basis, the matrix form of this linear map is given by

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

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

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

The first three columns of the reduced matrix clearly 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. Intuitvely, the space was formed by taking the (vector space associated to) the vertices of X and identifying vertices which lie on the boundary of an edge. That the quotient is 1-dimensional corresponds to the fact that the space $\mathcal{X}$ is connected! Let us prove a formal statement of this observation.

A *path component* of a topological space X is a maximal path connected subspace of X.

Proposition 3.3.4. *The number of path components of $\mathcal{X}$ is given by the dimension of the vector space $C_0(X, F_2)/\text{image}(\partial_1)$.*

Proof. Let $\mathcal{C} = \{\mathcal{X}^{(1)}, \ldots, \mathcal{X}^{(m)}\}$ denote the set of nonempty path components of X. Consider the linear map

$$L : C_0(X, F_2) \to V_{F_2}(\mathcal{C})$$

defined on a basis vector v by setting $L(v)$ to be the vector represented by the pathcomponent of v. Then L is surjective (each path component $\mathcal{X}_j$ contains some vertex, so the vector corresponding to $\mathcal{X}_j$ lies in the image of L). It follows that there is an induced isomorphism

$$C_0(X, F_2)/\ker(L) \xrightarrow{\approx} V_{F_2}(\mathcal{C}).$$

We wish to show that $\ker(L) = \text{image}(\partial_1)$.

First note that for any basis vector $vw \in \text{Edge}(X)$,

$$L(\partial_1(vw)) = L(v + w) = L(v) + L(w).$$

Since v and v lie in the same path component $\mathcal{X}^{(j)}$—the edge vw joining them is a path—we have

$$L(v) + L(w) = \mathcal{X}^{(j)} + \mathcal{X}^{(j)} = 0,$$

where the last equality follows since we are working over F_2. This proves that $\text{image}(\partial_1) \subset \ker(L)$.

It remains to show that $\ker(L) \subset \text{image}(\partial_1)$. Let $\mathbf{v} = v_1 + v_2 + \cdots + v_\ell$ be an element of $\ker(L)$. Then

$$0 = L(\vec{v}) = L(v_1) + \cdots + L(v_\ell)$$

implies that each path component appears an even number of time in this expression. That is, we can rename and rearrange our vertices so that

$$\mathbf{v} = v_1^{(1)} + v_2^{(1)} + \cdots + v_{2k_1}^{(1)} + v_1^{(2)} + v_2^{(2)} + \cdots + v_{2k_2}^{(2)} + \cdots + v_1^{(m)} + v_2^{(m)} + \cdots + v_{2k_m}^{(m)},$$

where

- $L(v_k^{(j)}) = \mathcal{X}^{(j)}$ for each $j = 1, \ldots, m$,
- Each k_j is a nonnegative integer ($k_j = 0$ is allowed since $\mathcal{X}^{(j)}$ might not be the path component of any vertex in $\vec{v}$.

Now we note that $v_1^{(1)}$ and $v_2^{(1)}$ are points in $\mathcal{X}^{(1)}$, meaning that they can be joined by some path in $\mathcal{X}_1$. Moreover, this path can be taken to have image in the 1-skeleton of $\mathcal{X}^{(1)}$ (any path joining vertices of a simplex can be homotoped to have image in the 1-skeleton of the simplex, and this can be done iteratively to homotope our path above to live in the 1-skeleton of $\mathcal{X}^{(1)}$). Thus we have a sequence of vertices

$$v_1^{(1)} = w_1, w_2, \ldots, w_p = v_2^{(1)}$$

such that $w_i w_{i+1}$ is an edge of X. Then

$$\partial_1(w_1w_2 + w_2w_3 + \cdots + w_{p-1}w_p) = w_1 + w_2 + w_2 + w_3 + \cdots + w_{p-1}w_p = w_1 + w_p = v_1^{(1)} + v_2^{(1)},$$

where we use the fact that we are working over F_2 to telescope the sum. We have shown that $v_1^{(1)} + v_2^{(1)} \in \text{image}(\partial_1)$. By the same arguments, we have that

$$v_3^{(1)} + v_4^{(1)}, v_5^{(1)} + v_6^{(1)}, \ldots, v_{2k_1-1}^{(1)} + v_{2k_1}^{(1)} \in \text{image}(\partial_1)$$

(using the even number of vertices), whence

$$v_1^{(1)} + v_2^{(1)} + \cdots + v_{2k_1}^{(1)} \in \text{image}(\partial_1).$$

Iterating the argument over the other groupings, we finally conclude that

$$v_1^{(1)} + \cdots + v_{2k_1}^{(1)} + v_1^{(2)} + \cdots + v_{2k_2}^{(2)} + \cdots + v_1^{(m)} + \cdots + v_{2k_m}^{(m)} \in \text{image}(\partial_1)$$

and this completes the proof. □

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 1.4.3) 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

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

or

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

in our simplified notation. 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

$$\{A, B\} \cup \{A, C\} \cup \{B, C\} \quad \text{and} \quad \{A, C\} \cup \{A, D\} \cup \{C, D\}, \tag{3.1}$$

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 (3.1). 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

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

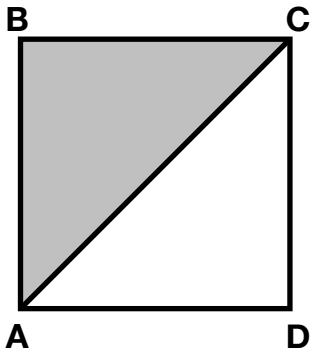
which can be expressed as the linear combination

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

(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 (3.1). We finally conclude that the dimension of the kernel of ∂_1 counts the loops in $\mathcal{X}$ which are *independent* in this sense.

3.3.3 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 3.3.4, 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

$$\partial_2 = \begin{array}{c} \\ AB \\ AC \\ AD \\ BC \\ CD \end{array} \begin{array}{c} ABC \\ \left[\begin{array}{c} 1 \\ 1 \\ 0 \\ 1 \\ 0 \end{array}\right] \end{array}.$$

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 3.3.5. *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.

3.4 Homology of Simplicial Complexes over F_2

3.4.1 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. Using the simplified notation from Section 3.3.1, we will frequently denote elements of C_k by

$$\lambda_1\sigma_1 + \cdots + \lambda_\ell\sigma_\ell \quad \lambda_j \in F_2,$$

where $\sigma_1, \ldots, \sigma_\ell \subset X_k$ are the k-simplices of X.

Remark 3.4.1. *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 3.5.2 for a brief discussion of these more general chain groups.*

Boundary Maps

Let X be a simplicial complex whose simplices are of dimension at most n. For each $k = 1, \ldots, n$, and for each $j = 0, \ldots, k$, we define the *jth boundary map*

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

by defining it on each basis element $\sigma = \{v_0, \ldots, v_k\}$ by

$$\partial_k^j \sigma = \{v_0, \ldots, \widehat{v_j}, \ldots, v_k\},$$

where $\widehat{v_j}$ denotes that the vertex v_j has been omitted from the list. Thus $\partial_k^j \sigma$ is a $(k-1)$-dimensional face of σ. The map ∂_k^j is defined on all of C_k by extending linearly. We then define the *boundary map* ∂_k on basis elements σ by

$$\partial_k \sigma = \sum_{j=0}^{k} \partial_k^j \sigma$$

and extend linearly to all of C_k.

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! It can also be expressed on basis elements by

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

3.4.2 Cycles, Boundaries and Homology

Main Property of ∂_k

The boundary maps ∂_k have a very important property:

Theorem 3.4.2. *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 $d^2 = 0$. 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

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

By linearity,

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

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

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 3.4.2.

Corollary 3.4.3. *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 3.4.2 implies that $\partial_k \circ \partial_{k+1}\psi$ is zero; i.e. $\partial_k \phi = 0$ and $\phi \in Z_k$. □

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 3.4.3.

Remark 3.4.4. *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 3.5.2.*

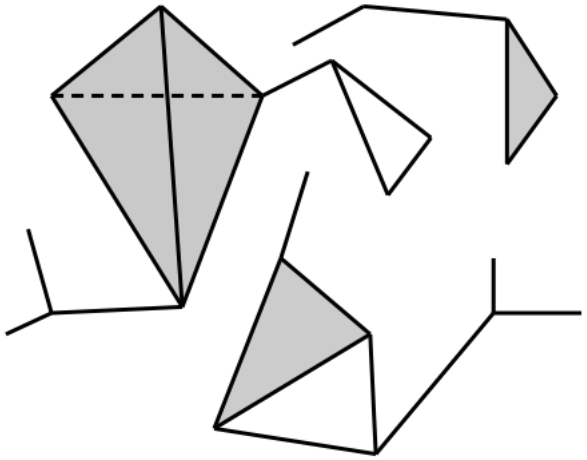
Betti Numbers

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 3.4.1. 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.



3.4.3 Functoriality

Induced Maps on Homology Vector Spaces

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 3.4.5. *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

$$\partial_k \left(C_k(f)(\phi_\sigma)\right) = \partial_k \phi_{f(\sigma)} = \sum \{\phi_\xi \mid \xi \text{ is a } (k-1)\text{-face of } f(\sigma)\}. \tag{3.2}$$

On the other hand,

$$\begin{aligned} C_{k-1}(f)(\partial_k(\sigma)) &= C_{k-1}(f)\left(\sum\{\phi_\tau \mid \tau \text{ is a } (k-1)\text{-face of } \sigma\}\right) \\ &= \sum\{\phi_{f(\tau)} \mid \tau \text{ is a } (k-1)\text{-face of } \sigma\}. \end{aligned} \tag{3.3}$$

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 (3.2) and (3.3) 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 $H_k(Y) = Z_k(Y)/B_k(Y)$. □

Categories and Functors

The property described by Theorem 3.4.5 shows that simplicial homology is a *functor*. This is a concept coming from *category theory*, a vast field of mathematics from which we will only describe some very basic ideas (see [1] for more details).

Roughly, a *category* is a pair $\mathrm{Cat} = (\mathcal{O}, \mathcal{M})$, where $\mathcal{O}$ is a collection of mathematical objects and $\mathcal{M}$ is a collection of morphisms, or allowable maps between the objects (satisfying certain axioms). Relevant examples are:

- The category Simp of simplicial complexes: the objects are simplicial complexes and the morphisms are simplicial maps between them.
- The category $\mathrm{Vec}_{\mathbb{F}}$ of vector spaces: the objects are vector spaces over a fixed field $\mathbb{F}$ and the morphisms are linear maps.
- The category Top of topological spaces: the objects are topological spaces and the morphisms are continuous maps.

A *functor* is a map $F : \mathrm{Cat}_1 \to \mathrm{Cat}_2$ between categories which takes each object X in Cat_1 to an object $F(X)$ in Cat_1. Moreover, if there is a morphism $f : X \to Y$ between a pair of objects in Cat_1, the functor must take it to a morphism $F(f) : F(X) \to F(Y)$.

Example 3.4.2. For each nonnegative integer k, H_k is a functor from Simp to Vec_{F_2}. Indeed, for each simplicial complex X, $H_k(X)$ is a vector space over F_2. Moreover, Theorem 3.4.5 says that for any morphism (simplicial map) $f : X \to Y$ of simplicial complexes, there is a morphism (linear map) $H_k(f) : H_k(X) \to H_k(Y)$.

Other Notions of Functoriality and a Black-Box Theorem

The notion of the functoriality of homology provided by Theorem 3.4.5 will be sufficient for our main purposes in the following chapters (using homology to compute signatures of datasets). However, the reader may be interested to know that there are much more general notions of homology and functoriality. These are explored to some extent in Section 3.5.2 (this section is independent from the following chapters and can be skipped, depending on the reader's interest).

To get intuition for computing homology and to complete some of the exercises, it will be useful to know the following fact. For now, this can be taken as a "black-box theorem"; that is, the reader can assume that it is true even if we will not prove it. Some clues on how the theorem can be proved are provided in Section 3.5.2.

Theorem 3.4.6. *Let K and K' be simplicial complexes. If their geometric realizations are homotopy equivalent, then their simplicial homology groups are isomorphic.*

In contrapositive form, this gives us a useful tool for distinguishing simplicial complexes up to homotopy equivalence.

Corollary 3.4.7. *If the homology groups of simplicial complexes K and K' are not all isomorphic, then the geometric realizations of K and K' are not homotopy equivalent.*

Another useful result is that we can define a simple version of homology for general topological spaces which are homotopy equivalent to simplicial complexes. A more sophisticated version of homology for a topological space (singular homology) is defined in Section 3.5.2.

Corollary 3.4.8. *If a topological space X is homotopy equivalent to a simplicial complex K, then we can define the homology of X to be*

$$H_n(X) := H_n(K),$$

for all n. Then $H_n(X)$ is well-defined up to isomorphism.

Proof. We need to show that if $X \sim_{h.e.} K$ and $X \sim_{h.e.} K'$, then $H_n(K) \approx H_n(K')$. This follows immediately from Theorem 3.4.6 and the fact that $\sim_{h.e.}$ is an equivalence relation (in particular, that it is transitive). □

3.5 More Advanced Topics in Homology

In this section, we outline some generalizations of the notion of homology of a simplicial complex over F_2. This section is essentially independent of the following chapters and can be skipped, depending on the interests of the reader.

3.5.1 Homology of a General Chain Complex

One might notice that as a purely algebraic construction, homology is a measurement which records information about a sequence of vector spaces with maps between them. In this section we make a short detour to introduce terminology for homology in a more general setting. This is meant to give the interested reader a starting point for studying the field of mathematics known as *homological algebra*.

Chain Complexes

Let $\mathbb{F}$ be a field. A *(non-negatively graded) chain complex over* $\mathbb{F}$ consists of:

- A collection $C_* = \{C_k\}_{k \geqslant 0}$ of vector spaces over $\mathbb{F}$, and
- Linear maps $\partial_k : C_k \to C_{k-1}$ for all $n \geqslant 1$ referred to as *boundary maps* which satisfy $\partial_k \circ \partial_{k+1} = 0$ for all $k \geqslant 1$.

Chain complexes are typically represented in the following form

$$\cdots \xrightarrow{\partial_{k+1}} C_k \xrightarrow{\partial_k} C_{k-1} \xrightarrow{d_{k-1}} \cdots \xrightarrow{d_2} C_1 \xrightarrow{d_1} C_0 \to 0.$$

Remark 3.5.1. *More generally, one could replace* $\mathbb{F}$ *with a* ring R *(a generalization of a field where multiplicative inverses don't necessarily exist) and consider chain complexes over* R *consisting of* R-modules *(essentially "vector spaces" over* R*)* C_k *and module homomorphisms* ∂_k*. Developing the machinery to this level of generality at this moment would take our discussion too far afield!*

Terminology from Homological Algebra

A sequence of vector spaces and maps

$$\cdots \xrightarrow{f_{k+1}} V_k \xrightarrow{f_k} V_{k-1} \xrightarrow{f_{k-1}} \cdots \xrightarrow{f_2} V_1 \xrightarrow{f_1} V_0 \to 0$$

is called *closed* if $f_m \circ f_{m+1} = 0$ for all m. This is equivalent to the statement that $\text{image}(f_{m+1}) \subset \ker(f_m)$. The sequence is called *exact* if $\text{image}(f_{m+1}) = \ker(f_m)$. An exact sequence of the form

$$0 \to V_2 \xrightarrow{d_2} V_1 \xrightarrow{d_1} V_0 \to 0$$

is called a *short exact sequence* . A sequence with more than three terms is called a *long exact sequence* .

Cycles, Boundaries and Homology

For a chain complex C_*, the subspace $Z_k = Z_k(C_*) = \ker(\partial_k) \subset C_k$ is called the space of k*-cycles* of C_* . The space $B_k = B_k(C_*) = \text{image}(\partial_{k+1})$ is called the space of k*-boundaries* of C_* . We then define the *kth homology group* of C_* to be

$$H_k(C_*) = Z_k / B_k.$$

Such a construction makes sense because the assumption that each ∂_k is closed implies that $B_k \subset Z_k$ for all k.

So we see that the chain groups defined above for simplicial complexes are examples of a much more general theory. We will see some other natural examples of chain complexes and homology in the next section.

Chain Maps

Let (C_*, d_*^C) and (D_*, d_*^D) be chain complexes. The correct notion of a map between these complexes is referred to as a *chain map* . This is a collection ϕ_* of linear maps $\phi_k : C_k \to D_k$ which preserve the boundary maps in the sense that the following diagram commutes:

$$\begin{array}{ccc} C_k & \xrightarrow{d_k^C} & C_{k-1} \\ {\scriptstyle \phi_k}\downarrow & & \downarrow{\scriptstyle \phi_{k-1}} \\ D_k & \xrightarrow{d_k^D} & D_{k-1} \end{array}$$

Chain maps are useful in that they induce well-defined maps on homology. We leave the proof of the following theorem as an exercise. We use the notation Z_n^C for the space of n-cycles of the complex (C_*, d_*^C) and B_n^C for the space of n-boundaries of (C_*, d_*^C).

Theorem 3.5.2. *A chain map $\phi_* : (C_*, d_*^C) \to (D_*, d_*^D)$ maps $Z_n^C \xrightarrow{\phi_n} Z_n^D$ and $B_n^C \xrightarrow{\phi_n} B_n^D$ (i.e. a chain map takes cycles to cycles and boundaries to boundaries). It follows that ϕ_* induces a linear map on each homology group $H_n(C_*) \xrightarrow{\phi_n} H_n(D_*)$.*

We will show in the next section that chain maps arise naturally from maps of simplicial complexes.

Short Exact Sequences of Chain Complexes

Let C_*, D_* and E_* be chain complexes. A *short exact sequence of chain complexes* is a pair of chain maps

$$0 \to C_* \xrightarrow{\phi_*} D_* \xrightarrow{\psi_*} E_* \to 0$$

such that in each degree k, the chain complex

$$0 \to C_k \xrightarrow{\phi_k} D_k \xrightarrow{\psi_k} E_k \to 0$$

is a short exact sequence.

The following is one of the fundamental theorems of homology theory. The proof will be left as a guided exercise.

Theorem 3.5.3. *A short exact sequence of chain complexes induces a long exact sequence of homology*

$$\cdots \to H_n(C_*) \xrightarrow{f_n} H_n(D_*) \xrightarrow{g_n} H_n(E_n) \xrightarrow{\partial_n} H_{n-1}(C_*) \xrightarrow{f_{n-1}} H_{n-1}(D_*) \xrightarrow{g_{n-1}} H_{n-1}(E_*) \xrightarrow{\partial_{n-1}} \cdots .$$

3.5.2 Variants of 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 [4].

3.5.3 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 over $\mathbb{Z}$

The most common convention is to define homology over the integers $\mathbb{Z}$, so we spend some extra time to develop this idea here, somewhat informally. Let X be a simplicial complex. We define the *kth chain group over $\mathbb{Z}$ of X* to be the collection of formal sums

$$C_k(X;\mathbb{Z}) = \{\lambda_1\sigma_1 + \cdots + \lambda_m\sigma_m \mid \lambda_j \in \mathbb{Z}\},$$

where $\sigma_1, \ldots, \sigma_m$ are the k-dimensional simplices of X. Formal sums of this form are referred to as *k-chains*. Note that this is in direct analogy to the kth chain group of X over F_2. The important difference here is that $C_k(X;\mathbb{Z})$ is not a vector space! It is a more general object called a *free Abelian group*. The formalism involved here is in the domain of abstract algebra; this is a huge field of mathematics, and we will avoid algebraic

technicalities as much as possible in this discussion. For our purposes, it is sufficient to consider $C_k(X;\mathbb{Z})$ as the space of formal sums together with the addition rule

$$(\lambda_1\sigma_1 + \cdots + \lambda_m\sigma_m) + (\lambda'_1\sigma_1 + \cdots + \lambda'_m\sigma_m) = (\lambda_1 + \lambda'_1)\sigma_1 + \cdots + (\lambda_m + \lambda'_m)\sigma_m.$$

The boundary maps $d_n : C_n(X;\mathbb{Z}) \to C_{n-1}(X;\mathbb{Z})$ are defined them on simplices and extending linearly. For a n-simplex $\sigma = \{v_0, \ldots, v_n\}$ in X, we define $\partial_j\sigma$ by

$$\partial_j\sigma = \{v_0, \ldots, \widehat{v_j}, \ldots, v_n\}$$

and the boundary map d_n by

$$d_n\sigma = \sum_{j=0}^{n}(-1)^j\partial_j\sigma.$$

The boundary maps satisfy the fundamental property of a chain complex:

Theorem 3.5.4. *For each n, the boundary maps $d_n : C_n(X;\mathbb{Z}) \to C_{n-1}(X;\mathbb{Z})$ have the property that*

$$d_{n-1} \circ d_n : C_n(X;\mathbb{Z}) \to C_{n-2}(X;\mathbb{Z})$$

is the zero map sending each n-chain to the chain with all zero coefficients.

We leave the proof of this theorem as an exercise.

For these collections of formal sums, it still makes sense to define kernels and images. The *kernel* of d_n is the collection of n-chains in $C_n(X;\mathbb{Z})$ which map under d_n to the $(n-1)$-chain with all coefficients equal to zero. The *image* of d_n is simply the image of d_n, considered as a function. By the theorem above, we get a sequence of well-defined sets called the *homology groups over* $\mathbb{Z}$, defined just as in the F_2 by

$$H_n(X;\mathbb{Z}) = \ker(d_n)/\text{image}(d_{n+1}).$$

The quotient in this definition can be understood at the level of quotienting the set $\ker(d_n)$ by the equivalence relation $c \sim c'$ if and only if $c - c' \in \text{image}(d_{n+1})$.

3.5.4 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 3.3.)

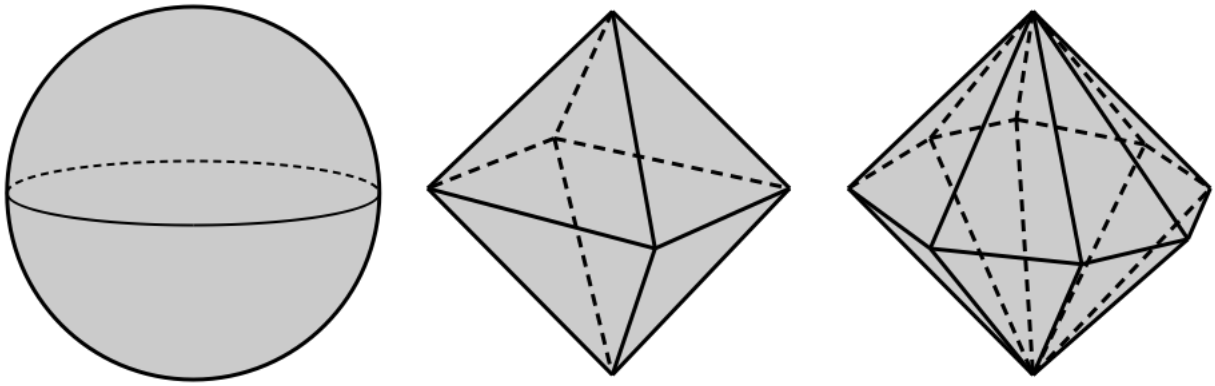


Figure 3.3: A 2-sphere $S^2 \subset \mathbb{R}^3$ and a pair of distinct triangulations.

- 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 [7] 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!

3.5.5 Functoriality Revisited

We saw in Theorem 3.4.5 that simplicial homology defines a functor from the category of simplicial complexes to the category of vector spaces over F_2. Similarly, we have the corresponding theorem for singular homology.

Theorem 3.5.5. *For each nonnegative integer k, singular homology $H_k^{sing}(\cdot; F_2)$ defines a functor from the category of topological spaces to the category of vector spaces over F_2.*

We omit the proof of the theorem, although it is not significantly more difficult than the proof of Theorem 3.4.5 (the interested reader should try to give the proof themselves!). With more work, one is able to prove the following generalization of our "black-box theorem", Theorem 3.4.6.

Theorem 3.5.6. *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.*

We also omit the proof of this theorem. The interested reader can find a proof in [4, Theorem 2.10]. Surprisingly, by using greater generality and considering topological spaces and singular homology (rather than simplicial complexes and simplicial maps), the proof of this theorem becomes significantly easier than that of Theorem 3.4.6. To derive Theorem 3.4.6 from the more general Theorem 3.5.6, one can use the following result.

Theorem 3.5.7. *Let K be an abstract simplicial complex. Then the singular homology of its geometric realization is isomorphic to its simplicial homology: $H_k^{sing}(\mathcal{K}; F_2) \approx H_k(K; F_2)$ for all k.*

The proof of this theorem is fairly technical and is treated in [4, Theorem 2.27] (the theorem there treats the slightly more general setting of Δ-complexes). Assuming the above theorems, one is able to derive Theorem 3.4.6 as a corollary.

Proof of Theorem 3.4.6. Combining the two theorems above, we have

$$H_n(K; F_2) \approx H_n^{sing}(K; F_2) \approx H_n^{sing}(K'; F_2) \approx H^n(K'; F_2).$$

□

3.6 Exercises

1. Let $S \subset \mathbb{R}^k$ be a set. Show that $\mathrm{cvx}(S)$ is equal to the set

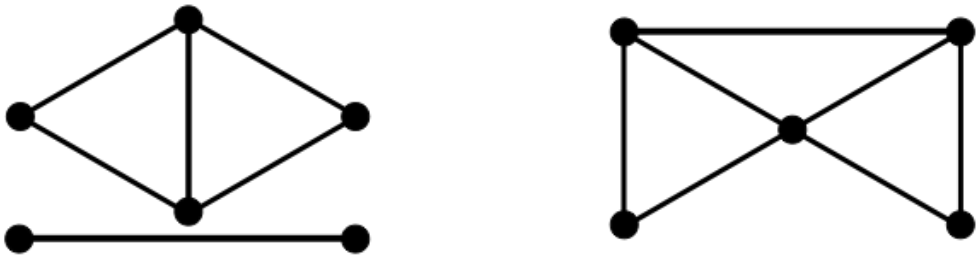
$$\bigcup \{(1-t)x + ty \mid t \in [0,1],\ x, y \in S\}.$$

2. Let $S \subset \mathbb{R}^k$ be a finite set. Show that $\mathrm{cvx}(S)$ is equal to the convex linear span of the points of S.

3. Write down the abstract simplicial complex associated to the geometric simplicial complex shown below.

4. Prove Proposition 3.3.2.

5. Compute the dimensions of the homology vector spaces $H_j(X)$ and $H_j(Y)$ over $\mathbb{F}_2$ for the 1-dimensional simplicial complexes shown below, where $j \in \{0, 1\}$.



X $\qquad$ Y

6. Compute the Betti numbers of the abstract simplicial complex X with vertex set

$$V(X) = \{A, B, C, D, E, F\},$$

and $1-$ and $2-$simplices given by the respective sets

$$\{AB, AD, AE, BC, BD, BE, CD, DE\} \quad \text{and} \quad \{ABC\}.$$

Can you draw a picture of a geometric realization of this complex?

7. For each of the simplicial complexes shown in Figure 3.4, 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 $\ker(\partial_1)/\text{image}(\partial_2)$ in each case. Does your answer make sense intuitively?

8. Let $\mathcal{X}$ be a simplicial complex and let $X = (V(X), \Sigma(X))$ be as defined in Example 3.2.3. Show that X defines an abstract simplicial complex.

9. Let $\mathcal{X}$ be the standard n-simplex defined in Example 3.2.1 and let X be the abstract simplex associated to $\mathcal{X}$ (see Example 3.2.3). Show that X is simplicially isomorphic to the standard abstract n-simplex defined in Example 3.2.4.

10. Prove that the spaces from Example 3.1.4 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 3.4.7), conclude that the spaces are not homotopy equivalent, hence not homeomorphic.

11. Prove Theorem 3.5.2.

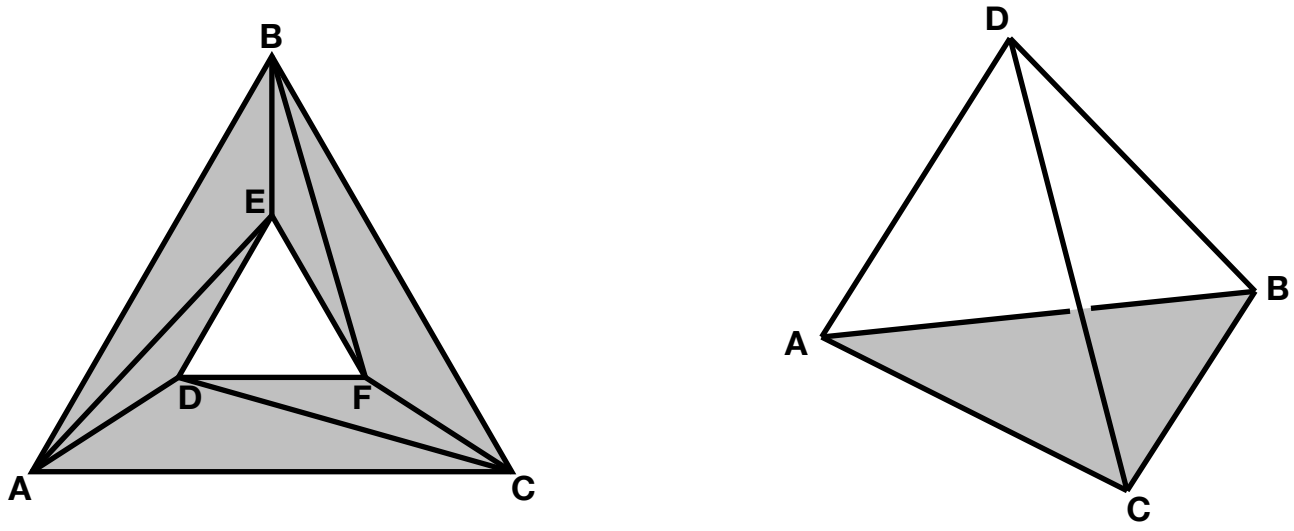


Figure 3.4: The space on the left is a simplicial complex in $\mathbb{R}^2$. The space on the right is a simplicial complex in $\mathbb{R}^3$. It has the 1-skeleton of a 3-simplex, with a single face $\{A, B, C\}$ filled in.

12. Show that any triangulation of the 2-sphere must involve at least four 2-simplices.

13. a) Show that a chain map $\phi_* : (C_*, d_*^C) \to (D_*, d_*^D)$ maps $Z_n^C \xrightarrow{\phi_n} Z_n^D$ and $B_n^C \xrightarrow{\phi_n} B_n^D$ (i.e. a chain map takes cycles to cycles and boundaries to boundaries).

    b) Show that ϕ_* induces a linear map on each homology group $H_n(C_*) \xrightarrow{\phi_n} H_n(D_*)$. Note: you need to define the map ϕ_n on homology vector spaces. Since your map is between quotient vector spaces, it is highly likely that you will have to prove that your map is well-defined!

14. Let K be a simplicial complex, K^1 and K^2 subcomplexes (subsets which are simplicial complexes themselves, with vertices in the vertex set of K) such that $K^1 \cup K^2 = K$ and $K^1 \cap K^2 = L$ is another subcomplex. There are inclusion maps $i^1 : L \hookrightarrow K^1$, $i^2 : L \hookrightarrow K^2$, $j^1 : K^1 \hookrightarrow K$ and $j^2 : K^2 \hookrightarrow K$.

    a) Show that these inclusion maps are simplicial.

    b) Since the maps are simplicial, they induce chain maps (by the previous problem). Moreover, we can construct a map $C_*(L) \xrightarrow{i^1 \oplus i^2} C_*(K^1) \oplus C_*(K^2)$ defined by

$$i^1 \oplus i^2(\sigma) = (i^1(\sigma), i^2(\sigma))$$

    and a map $C_*(K^2) \xrightarrow{j^1 + j^2} C_*(K)$ defined by

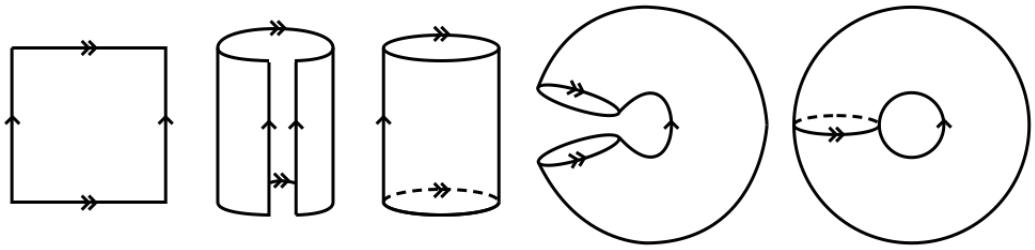
$$(j^1 + j^2)(\sigma_1, \sigma_2) = j^1(\sigma_1) + j^2(\sigma_2).$$

Show that

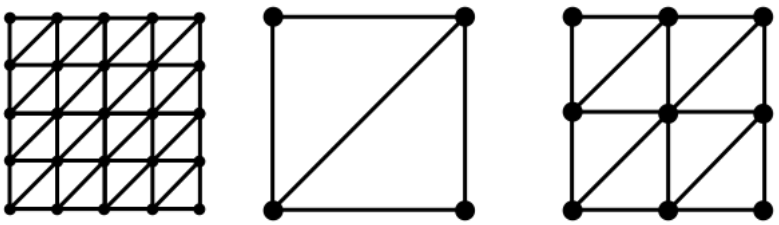
$$0 \to C_*(L) \xrightarrow{i^1 \oplus i^2} C_*(K^1) \oplus C_*(K^2) \xrightarrow{j^1 + j^2} C_*(K) \to 0.$$

is a short exact sequence.

15. A torus is the topological space which is homeomorphic to the surface of a donut. This space can be obtained by identifying opposite sides of a square—see the figure below.



The figure on the left below can then be used to give a triangulation of the torus. Explain why the two figures shown below on the right *do not* give triangulations of the torus (i.e., explain why they don't define simplicial complexes after the identification procedure is performed).



16. Show that $d_{n-1} \circ d_n = 0$ for chain groups over $\mathbb{Z}$.

17. Prove that an n-simplex Δ^n is contractible (i.e., homotopy equivalent to a point). Use this to justify the claim that we made in class that the homology of Δ^n is

$$H_k(\Delta^n; \mathbb{F}_2) = \begin{cases} \mathbb{F}_2 & k = 0 \\ 0 & k \neq 0. \end{cases}$$

For this problem, you are allowed to use the "black box theorems" and their corollaries from Section 3.4.3.

18. Compute the homology vector spaces over $\mathbb{F}_2$ of a *punctured torus* (a torus with a single point removed).

Hint: Recall that the torus is realized as a square with opposite edges identified. What does the punctured torus look like from this perspective? Try homotoping to a simpler space.

3.6.1 Extended Exercise: Classification of 1-complexes

In this section, we outline a proof of a "classification theorem" for 1-dimensional simplicial complexes. The goal of the exercise is to fill in the proofs of the lemmas.

Theorem 3.6.1. *For any connected* 1*-dimensional simplicial complex* K*,*

$$H_k(K; F_2) \approx \begin{cases} F_2 & k = 0 \\ F_2^{1+E-V} & k = 1 \\ \{0\} & k \neq 0, 1, \end{cases}$$

where V *denotes the number of* 0*-simplices (vertices) in* K *and* E *denotes the number of* 1*-simplices (edges) in* K*.*

Note that a 1-dimensional simplicial complex is a geometric realization of a graph. Prove the following lemmas:

Lemma 3.6.2. *Any graph* G *contains a subgraph which is a tree and which contains every vertex of* G*. Such a subgraph is called a* maximal subtree *of* G*.*

Lemma 3.6.3. *Any (geometric realization of a) tree is contractible.*

Lemma 3.6.4. *For a tree with* V *vertices and* E *edges,* $V - E = 1$*.*

Lemma 3.6.5. *Let* G *be a graph with* V *vertices and* E *edges. The number of edges not contained in a maximal subtree* T *of a graph* G *is* $1 + E - V$*.*

A *wedge of* n *circles* is the topological space obtained by taking n disjoint circles S^1 and a disjoint 1-point space $\{x\}$, choosing a point from each copy of S^1, and attaching the chosen point in each circle to $\{x\}$.

Lemma 3.6.6. *The homology of a wedge of* n*-circles* X *is*

$$H_k(X; F_2) \approx \begin{cases} F_2 & k = 0 \\ F_2^n & k = 1 \\ \{0\} & k \neq 0, 1. \end{cases}$$

Lemma 3.6.7. *A* 1*-dimensional simplicial complex* K *with* V *vertices and* E *edges is homotopy equivalent to a wedge of* $1 + E - V$ *circles.*

Combining these lemmas together with the "black box theorems" of Section 3.4.3, we have proved Theorem 3.6.1. The theorem has the following immediate corollary.

Corollary 3.6.8. *Let* K *and* K' *be connected* 1*-dimensional simplicial complexes. Suppose* K *has* V *vertices and* E *edges and that* K' *has* V' *vertices and* K' *edges. Then* K *and* K' *are homotopy equivalent if and only if* $E - V = E' - V'$*.*

4 Persistent Homology

As we stated in Chapter 2, 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 4.1. 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 the so-called *persistent homology* of their associated *Vietoris-Rips complexes*. We have not yet defined these technical terms, but 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 $\mathsf{VR}(X)$ 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.

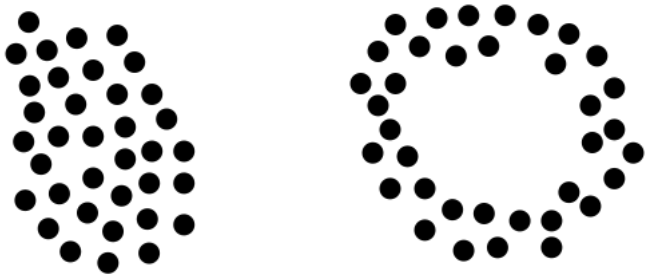


Figure 4.1: A pair of point clouds with the same number of points.

4.1 Filtered Simplicial Complexes

4.1.1 Definitions

Filtered Simplicial Complexes

A *filtered simplicial complex* is a collection $\mathsf{K} = \{K_r\}_{r \geqslant 0}$ of (finite) simplicial complexes K_r together with a family of simplicial maps

$$f_{r,r'} : V(K_r) \to V(K_{r'})$$

for each pair (r, r') with $r \leqslant r'$. The simplicial maps are required to satisfy the compatibility condition that for all $r \leqslant r' \leqslant r''$, we have

$$f_{r,r''} = f_{r'',r'} \circ f_{r',r}.$$

In applications, filtered simplicial complexes arise as multiscale approximations of the underlying topology of discrete data sets. They are constructed using a variety of algorithms, some of which are introduced in the following sections.

Remark 4.1.1. *We are indexing the filtered simplicial complex over the set of positive reals $r \geqslant 0$. This is only a convenient convention, and most of what we will do would make sense if we indexed over any interval in $\mathbb{R}$. For example $r \geqslant r_0$ for some fixed real r_0 or the entire real line. These alternative indexing sets are used in some examples. This convention will continue when we move on to* persistence vector spaces*, which are essentially parameterized families of vector spaces.*

Finite Filtered Simplicial Complexes

Frequently in the cases we are interested in, the full information of a filtered simplicial complex can be described in a finite sequence of "snapshots". To make this precise, we define a *finite filtered simplicial complex* to be a finite sequence $\mathsf{K} = \{K_{r_j}\}_{j=1}^{n}$ of simplicial complexes, where $0 \leqslant r_1 < r_2 < \cdots < r_n$ is an increasing finite sequence of real numbers, together with simplicial maps

$$f_{r_i,r_j} : K_{r_i} \to K_{r_j}$$

for all $r_i \leqslant r_j$. We continue to require a compatibility condition on the simplicial maps, so it suffices to specify the simplicial maps $f_{r_i,r_{i+1}}$ for each i, which can simply be denoted f_{r_i}.

It will sometimes be the case that the index parameters in a finite filtered simplicial complex are not of particular interest. In this case, we express a finite filtered simplicial complex generically as

$$K_1 \xrightarrow{f_1} K_2 \xrightarrow{f_2} K_3 \xrightarrow{f_3} \cdots \xrightarrow{f_{n-1}} K_n, \tag{4.1}$$

or even more simply as

$$K_1 \to K_2 \to K_3 \to \cdots \to K_n$$

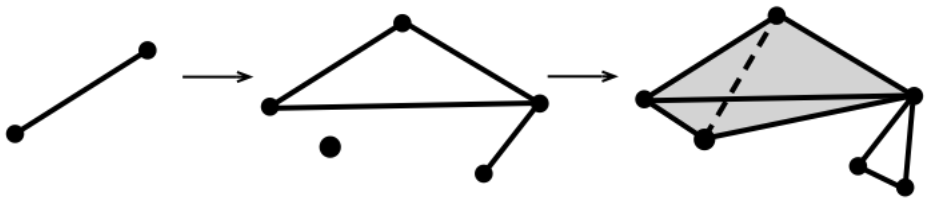
when the maps are clear from context. In our applications, the simplicial maps will be inclusions on a common vertex set.

Remark 4.1.2. *There are two abuses of notation here:*

1. *The maps f_i in* (4.1) *are really maps on the* vertex sets *of the simplicial complexes.*

2. *We are using the same notation* K *for a filtered simplicial complex and for a* finite *filtered simplicial complex. The filtered simplicial complexes we are primarily interested in are completely described by their finite counterparts. As such, we frequently conflate the two notions when no confusion should arise. To distinguish them from their finite counterparts, a filtered simplicial complex* $\mathsf{K} = \{K_r\}_{r \geqslant 0}$ *will sometimes be referred to as an* $\mathbb{R}$-indexed filtered simplicial complex*, or* $\mathbb{R}$-complex *for short.*

An example of a finite filtered simplicial complex is shown below, where all simplicial maps are inclusions of the vertex sets.



4.1.2 Finitely-Presented Filtered Simplicial Complexes

Given a finite simplicial complex $\mathsf{K} = \{K_{r_j}\}_{j=1}^{n}$, we can construct an $\mathbb{R}$-complex $\{K_r\}_{r \geqslant r_0}$ as follows:

$$K_r = \begin{cases} K_{r_i} & r \in [r_i, r_{i+1}) \\ K_{r_n} & r \geqslant r_n, \end{cases}$$

with maps $f_{r,r'}$ $(r \leqslant r')$ defined by

$$f_{r,r'} = \begin{cases} f_{r_i,r_j} & r \in [r_i, r_{i+1}),\ r' \in [r_j, r_{j+1}),\ i < j \\ id_{V(K_{r_i})} & r, r' \in [r_i, r_{i+1}) \end{cases}$$

(for simplicity, take $r_{n+1} = +\infty$). If desired, the $\mathbb{R}$-complex could be extended to be defined over all $r \geqslant 0$ by setting $K_r = \varnothing$ if $r < r_0$.

If an $\mathbb{R}$-complex can be obtained by applying the above construction to a finite filtered simplicial complex, then the $\mathbb{R}$-complex is said to be *finitely-presented*. Essentially all of the examples that we are interested in will be finitely-presented.

4.2 Vietoris-Rips Complexes

4.2.1 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 at parameter* r , denoted

$\mathsf{VR}(X, r)$, as follows. Writing $\mathsf{VR}(X, r) = (V_r, \Sigma_r)$, we define $V_r = X$ for all r. The simplex set Σ_r is defined by

$$\Sigma_r = \{\sigma \subset X \mid d(x, y) \leqslant r \ \ \forall\, x, y \in \sigma\}.$$

Observe that $\Sigma_r \subset \Sigma_s$ whenever $r \leqslant s$. Indeed,

$$x, y \in \sigma \in \Sigma_r \Rightarrow d(x, y) \leqslant r \leqslant s \Rightarrow \sigma \in \Sigma_s.$$

Therefore the identity map on the vertex set X induces a simplicial map $f_{r,s}$ from $\mathsf{VR}(X, r)$ to $\mathsf{VR}(X, s)$ for each pair $r \leqslant s$. It follows that the Vietoris-Rips complex of a finite metric space is a filtered simplicial complex, which we denote $\mathsf{VR}(X) = \{\mathsf{VR}(X, r)\}_{r \geqslant 0}$.

Note that there are only finitely many parameter values r where the structure of $\mathsf{VR}(X, r)$ can change. Indeed, if $X = \{x_1, \ldots, x_n\}$, then we can represent the metric structure of X by its *distance matrix* , a symmetric matrix of the form

$$d = (d(x_i, x_j))_{i,j} = \begin{array}{c} \\ x_1 \\ x_2 \\ \vdots \\ x_n \end{array} \begin{array}{c} \begin{array}{cccc} x_1 & x_2 & \cdots & x_n \end{array} \\ \begin{bmatrix} 0 & d(x_1, x_2) & \cdots & d(x_1, x_N) \\ d(x_2, x_1) & 0 & \cdots & d(x_2, x_N) \\ \vdots & \vdots & \vdots & \vdots \\ d(x_N, x_1) & d(x_N, x_2) & \cdots & 0 \end{bmatrix} \end{array}.$$

The (finitely many) entries of this matrix $0 = r_0 < r_1 < \cdots < r_N$ give all possible parameters r where $\mathsf{VR}(X, r)$ can change. It follows that $\mathsf{VR}(X, r)$ is completely described by a finite filtered simplicial complex

$$\mathsf{VR}(X, r_0) \to \mathsf{VR}(X, r_1) \to \cdots \mathsf{VR}(X, r_N).$$

That is, $\mathsf{VR}(X)$ is finitely-presented (see the previous section).

4.2.2 A Simple Example

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

Example 4.2.1. 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 $\mathsf{VR}(X, r)$ for any choice of $r \geqslant 0$. The

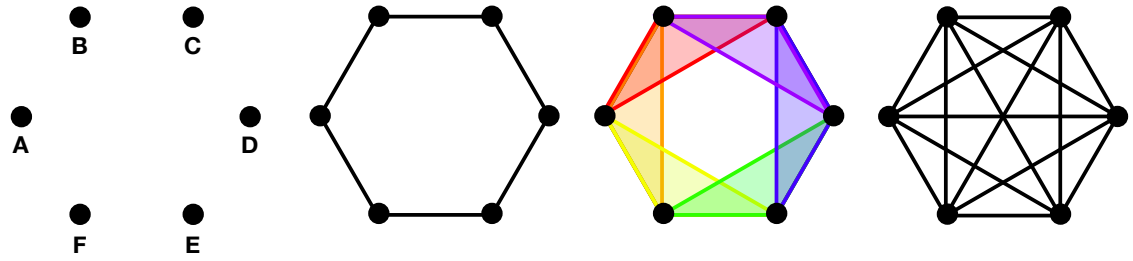
distance matrix of X is given by

$$d = (d(x_i, x_j))_{i,j} = \begin{array}{c} \\ A \\ B \\ C \\ D \\ E \\ F \end{array} \begin{array}{c} \begin{array}{cccccc} A & B & C & D & E & F \end{array} \\ \begin{bmatrix} 0 & 1 & \sqrt{3} & 2 & \sqrt{3} & 1 \\ 1 & 0 & 1 & \sqrt{3} & 2 & \sqrt{3} \\ \sqrt{3} & 1 & 0 & 1 & \sqrt{3} & 2 \\ 2 & \sqrt{3} & 1 & 0 & 1 & \sqrt{3} \\ \sqrt{3} & 2 & \sqrt{3} & 1 & 0 & 1 \\ 1 & \sqrt{3} & 2 & \sqrt{3} & 1 & 0 \end{bmatrix} \end{array}.$$

The numbers appearing in this distance matrix are $0 < 1 < \sqrt{3} < 2$. The Vietoris-Rips complexes of X is therefore described by the finite filtered simplicial complex

$$\mathsf{VR}(X, 0) \to \mathsf{VR}(X, 1) \to \mathsf{VR}(X, \sqrt{3}) \to \mathsf{VR}(X, 2),$$

where the complex at each parameter is described below.



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

2. For $1 \leqslant r < \sqrt{3}$, $\mathsf{VR}(X, r) = \mathsf{VR}(X, 1)$ 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 $\mathsf{VR}(X, r)$, 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 $\mathsf{VR}(X, r)$.

3. For $\sqrt{3} \leqslant r < 2$, $\mathsf{VR}(X, r) = \mathsf{VR}(X, \sqrt{3})$, and each triple of consecutive vertices forms a 2-simplex in $\mathsf{VR}(X, r)$. 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 $\mathsf{VR}(X, r)$. 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 $\mathsf{VR}(X, r)$. The figure second from the right shows the 2-dimensional simplices in $\mathsf{VR}(X, r)$. 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 $\mathsf{VR}(X, r)$ as an abstract simplicial complex.

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

4.2.3 Observations

From Example 4.2.1, we can immediately make some observations about $\mathsf{VR}(X, r)$ for any finite metric space (X, d).

Proposition 4.2.1. *For $r < \min\{d(x, x') \mid x, x' \in X, x \neq x'\}$, $\mathsf{VR}(X, r)$ is homeomorphic to X.*

Proof. If $r < \min\{d(x, x') \mid x, x' \in X, x \neq x'\}$, then for all $x, x' \in X$ we have

$$d(x, x') \leqslant r \Leftrightarrow x = x'.$$

It follows that the set of simplicies of $\mathsf{VR}(X, r)$ is

$$\{\sigma \subset X \mid d(x, x') \leqslant r \text{ for all } x, x' \in X\} = \{\{x\} \mid x \in X\} \approx X.$$

□

Let S be a finite set. The *complete simplicial complex on S* is the simplicial complex containing a simplex for every subset of S. It is implicially isomorphic to the standard simplex Δ^n, with $n = |X| - 1$. We leave the proof of the following proposition as an exercise.

Proposition 4.2.2. *For $r \geqslant \max\{d(x, x') \mid x, x' \in X\}$, $\mathsf{VR}(X, r)$ is the complete simplicial complex on X.*

4.2.4 Other Complexes

Besides the Vietoris-Rips complex, there are a variety of other ways to associate a simplicial complex to a finite metric space. We present some examples (α-complexes and filtrations induced by a function) below. In general, the correct filtration to use depends highly on the particular data analysis task at hand. Designing a useful filtration is part of the art of successfully applying tools from topological data analysis.

α-Complexes

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

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

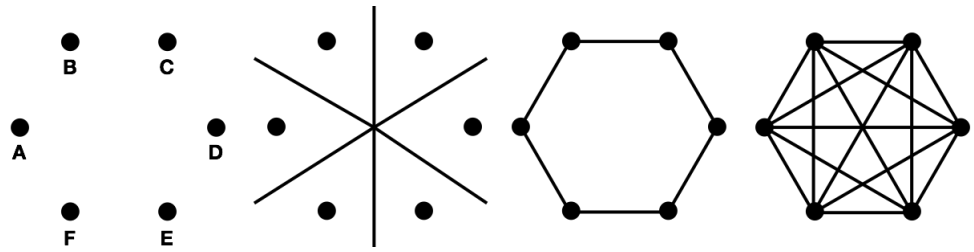


Figure 4.2: A point cloud, its Voronoi decomposition and its α-complexes at scale parameters $r = 3/4$ and $r = 5/4$.

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

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

An example is shown in Figure 4.2. 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.

Sublevel Set Filtrations

Let X be a simplicial complex and let $f : \Sigma(X) \to \mathbb{R}$ be a function. We define a filtered simplicial complex K^f from f called the *sublevel set filtration of X* as follows. Let $K_r^f = (V_r^f, \Sigma_r^f)$, with

$$V_r^f = \{v \in V(X) \mid f(\{v\}) \leqslant r\}$$

and

$$\Sigma_r^f = \{\sigma \in \Sigma(X) \mid f(\sigma) \leqslant r\}.$$

Then K^f is a finite filtered simplicial complex with simplicial maps $f_{r,s}$ given by inclusions on the vertex set.

Sublevel set filtrations are quite flexible. Any filtered simplicial complex whose simplicial maps are inclusions can be realized as a sublevel set filtration. This includes Vietoris-Rips complexes, as demonstrated by the following example.

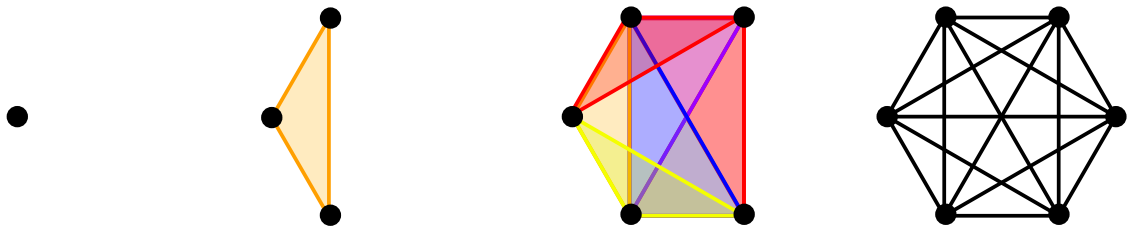
Example 4.2.2. The Vietoris-Rips complex of a finite metric space (X, d) is a special example of a sublevel set filtration. We treat the points X as the vertices of a complete

simplex δ^n with $n = |X| - 1$. We define a function f on $\Sigma(\Delta^n)$ by

$$f(\sigma) = \max\{d(x, y) \mid x, y \in \sigma\}.$$

The resulting sublevel set filtration K^f is exactly $\mathsf{VR}(X)$.

Example 4.2.3. Returning to the space X consisting of the vertices of a regular hexagon, we consider the points of X as the vertices of a complete 5-dimensional simplex. We define a filtration function $f : \Sigma(X) \to \mathbb{R}$ by first defining $f(\{V\})$ to be the distance of the arbitrary vertex V to the fixed vertex A, then declaring $f(\sigma) = \max\{f(x) \mid x \in \sigma\}$. The filtered simplicial complex K^f is pictured below (the final picture denotes the complete simplex).



4.3 Persistence Homology

An important phenomenon illustrated by Example 4.2.1 is that the topology of $\mathsf{VR}(X, r)$ 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 $1 \leqslant r < 2$, and the loop finally dissappears when $r = 2$ as it is filled in by higher-dimensional simplices. We have spent a great deal of energy studying an algorithm for quantifying topological features of simplicial complexes—simplicial homology! The key idea of persistent homology, which we are now prepared to define, is that one can track the appearance and disappearance of topological features in a filtered simplicial complex over time.

We now arrive at the main object of study in the theory of persistent homology. Let $\mathsf{K} = \{K_r\}_{r \geqslant 0}$ be a filtered simplicial complex. The *k-th persistence homology vector space* of a filtered simplicial complex is the collection of vector spaces

$$PH_k(\mathsf{K}) := \{PH_k(\mathsf{K})_r\}_{r \geqslant 0},$$

where

$$PH_k(\mathsf{K})_r := H_k(K_r).$$

Since the data of K includes simplicial maps $f_{r,s} : K_r \to K_s$ for all $r \leqslant s$. By the functoriality theorem for homology of simplicial complexes (Theorem 3.4.5), there are induced linear maps

$$H_k(f_{r,s}) : H_k(K_r) \to H_k(K_s)$$

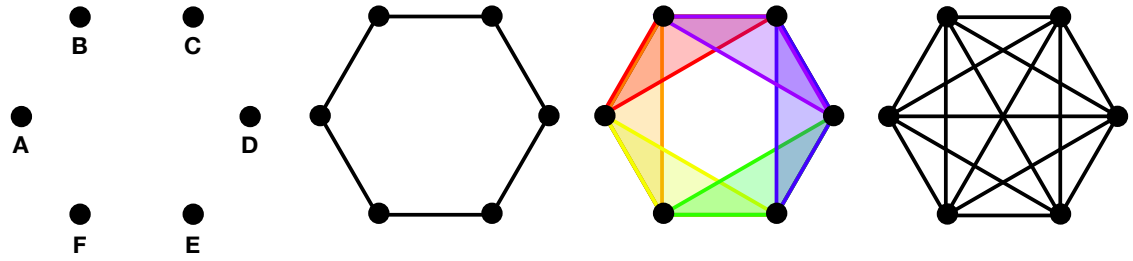
for all $r \leqslant s$. These linear maps should be considered as part of the data in the definition of $PH_k(\mathsf{K})$.

For a finite filtered simplicial complex $\mathsf{K} = K_1 \to K_2 \to \cdots \to K_n$, $PH_k(\mathsf{K})$ can be described completely by considering the homology vector space at finitely many values of r; that is, we define

$$PH_k(\mathsf{K}) = H_k(K_1) \to H_k(K_2) \to \cdots \to H_k(K_n).$$

4.3.1 The Hexagon Example

Consider the finite filtered simplicial complex arising as a Vietoris-Rips complex from Example 4.2.1. The parameter values depicted are $r = 0, 1, \sqrt{3}, 2$.



The $k = 0$ persistence homology is

$$F_2^6 \to F_2 \to F_2 \to F_2,$$

where the first map has 5-dimensional kernel and the other two maps are identities. The $k = 1$ persistence homology is

$$0 \to F_2 \to F_2 \to 0,$$

where the map in the middle is the identity map. Higher order homology of $\mathsf{VR}(X)$ is constantly zero.

Notice that the $k = 0$ homology can be decomposed in an interesting way. Let $\mathsf{I}[1, 4]$ denote the persistence vector space

$$F_2 \to F_2 \to F_2 \to F_2$$

and let $\mathsf{I}[1, 2)$ denote

$$F_2 \to 0 \to 0 \to 0.$$

Then the $k = 0$ persistence homology can be expressed as some combination of $\mathsf{I}[1, 4]$ together with five copies of $\mathsf{I}[1, 2)$. This informal observation gives a convenient way to record the variation in topology of the Vietoris-Rips complex of the space. This type of decomposition is known in general as a *barcode decomposition*. A fundamental result of the theory of persistence homology is that *any* finite persistence vector space can be decomposed in a similar manner. Precise definitions of these concepts and this result will be the main subjects of the rest of the chapter.

4.4 Persistence Vector Spaces

Our goal is now to study the structure of persistence homology vector spaces of a filtered simplicial complex. At its core, the persistent homology of K is an indexed collection of vector spaces, together with linear maps joining the spaces. As is typical in mathematics, it will be enlightening to find an abstract description of this structure and to study this abstraction in its own right. This leads to the notion of a persistence vector space, which we now begin to study.

4.4.1 Persistence Vector Spaces Indexed Over $\mathbb{R}$

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 $\mathsf{V} = \{V_r\}_{r \geqslant 0}$ 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 $L_{r,r''} = L_{r',r''} \circ L_{r,r'}$.

The definition of a persistence vector would make sense if we parameterized over any interval in $\mathbb{R}$; as in the case of filtered simplicial complexes, we fix positive reals as our indexing set as a convention. We will denote a persistence vector space by $\mathsf{V} = \{V_r\}$, with the understanding that persistence vector spaces are always parameterized over positive real numbers $r \geqslant 0$, unless specifically noted otherwise. When talking about multiple persistence vector spaces, we will need to distinguish their families of linear maps. In this case, we will use the notation $L^{\mathsf{V}}_{r,r'}$ for the maps associated to V.

Linear Transformation Between Persistence Vector Spaces

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

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

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,

$$\ell_{r'} \circ L^V_{r,r'} = L^W_{r,r'} \circ \ell_r.$$

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 $\mathsf{U} = \{U_r\}$ of linear subspaces $U_r \subset V_r$ such that $L^{\mathsf{V}}_{r,r'}(U_r) \subset U_{r'}$ holds for each $r \leqslant r'$.

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

$$\ker(\ell) = \{\ker(\ell_r)\}$$

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

$$\mathrm{im}(\ell) = \{\mathrm{im}(\ell_r)\}.$$

Quotient Persistence Vector Space

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

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

Direct Sum of Persistence Vector Spaces

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

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

given by

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

4.4.2 Finite Persistence Vector Spaces

We saw in Section 4.1 two representations of filtered simplicial complexes:

1. $\mathbb{R}$-complexes $\mathsf{K} = \{K_r\}_{r\geqslant 0}$, indexed over a continuous set,
2. Finite filtered simplicial complexes $\mathsf{K} = \{K_{r_j}\}_{j=1}^{n}$, indexed over a discrete set.

Moreover, we saw in Section 4.1.2 a recipe to construct an $\mathbb{R}$-complex from a filtered simplicial complex. We defined an $\mathbb{R}$-complex to be *finitely-presented* if it can be obtained by this construction. We saw in Section 4.3.1 that a finitely-presented filtered simplicial complex leads to a persistent homology vector space which can also be represented by a finite amount of data. We therefore wish to define a notion of a finitely-presented persistence vector space.

Definitions

A *finite persistence vector space* over $\mathbb{F}$ is a finite collection $\mathsf{V} = \{V_{r_j}\}_{j=1}^n$ of finite-dimensional vector spaces over $\mathbb{F}$, where $r_1 < r_2 < \cdots < r_n$ is some increasing sequence of real numbers, together with linear maps

$$L_{r_i,r_j} : V_{r_i} \to V_{r_j}$$

for each $r_i \leqslant r_j$ satisfying the compatibility condition $L_{r_j,r_k} \circ L_{r_i,r_j} = L_{r_i,r_k}$ whenever $r_i \leqslant r_j \leqslant r_k$. This data is completely described by the diagram

$$V_{r_1} \xrightarrow{L_{r_1,r_2}} V_2 \xrightarrow{L_{r_2,r_3}} \cdots \xrightarrow{L_{r_n,r_{n-1}}} V_{r_n};$$

i.e., all maps L_{r_i,r_j} are determined by knowing $L_{r_i,r_{i+1}}$. The maps may be decorated for clarity, e.g., $L^{\mathsf{V}}_{r_i,r_j}$. The usual constructions for standard vector spaces generalize to finite persistence vector spaces, just as they did for the parameterized persistence vector spaces introduced above.

Sometimes it may be the case that the particular index values r_j are not important to us, and we use the generic notation

$$V_1 \xrightarrow{L_1} V_2 \xrightarrow{L_2} \cdots \xrightarrow{L_{n-1}} V_n.$$

Since we now have a need to make a distinction, we refer to a persistence vector space $\mathsf{V} = \{V_r\}_{r\geqslant 0}$ as an $\mathbb{R}$-*index persistence vector space* or an $\mathbb{R}$-*PVS*. Similarly, the terminology for a finite persistence vector space will sometimes be shortened to *finite-PVS*.

Finitely-Presented $\mathbb{R}$-PVS

From a finite-PVS $\mathsf{V} = \{V_{r_j}\}_{j=1}^n$, we obtain an $\mathbb{R}$-PVS $\overline{\mathsf{V}} = \{\overline{V}_r\}_{r\geqslant 0}$ by the following formula:

$$\overline{V}_r = \begin{cases} 0 & r < r_1 \\ V_1 & r \in [r_1, r_2) \\ V_2 & r \in [r_2, r_3) \\ \vdots & \vdots \\ V_{n-1} & r \in [r_{n-1}, r_n) \\ V_n & r \geqslant r_n. \end{cases}$$

The linear maps $L^{\overline{\mathsf{V}}}_{r,r'}$ $(r \leqslant r')$ are given by

$$L^{\overline{\mathsf{V}}}_{r,r'} = \begin{cases} 0 & r < r_1 \text{ or } r' \geqslant r_n \\ L^{\mathsf{V}}_{r_i,r_j} & r \in [r_i, r_{i+1}) \text{ and } r' \in [r_j, r_{j+1}) \text{ and } i < j \\ \mathrm{Id}_{V_{r_i}} & r, r' \in [r_i, r_{i+1}). \end{cases}$$

An $\mathbb{R}$-PVS which arises from a finite-PVS via the above construction will be called *finitely-presented*. Our main results will treat finitely-presented PVS, since they arise

most naturally in our data analysis applications, as evidenced by the following proposition.

Proposition 4.4.1. *If* K *is a finitely-presented filtered simplicial complex, then its persistent homology* $PH_k(\mathsf{K})$ *is a finitely-presented persistence vector space.*

Proof. Indeed, if K is constructed from the finite filtered simplicial complex $\{K_{r_j}\}_{j=1}^n$, then $PH_k(\mathsf{K})$ is constructed from the finite persistence vector space $\{H_k(K_{r_j})\}_{j=1}^n$. □

4.4.3 Finite Persistence Vector Space Constructions

For each notion from linear algebra which was generalized to the $\mathbb{R}$-PVS setting in Section 4.4.1, there is a corresponding generalization in the finite-PVS setting. Throughout this subsection, we will make a strong assumption: we assume that all finite-PVS being considered have the same fixed indexing set $r_1 < r_2 < \cdots < r_n$. This restriction is certainly not required to make the definitions work, but it will greatly simplify notation. For starters, it allows us to use the simplified notation $\mathsf{V} = \{V_j\}_{j=1}^n$ and $\mathsf{W} = \{W_j\}_{j=1}^n$ for our finite persistence vector spaces. Let L_j^{V} and L_j^{W} denote the linear maps for each finite-PVS. We remark that nothing is lost by making the simplification of fixing an overall indexing set, since this will be the relevant setting for the theorems we want to prove later.

Direct Sums

The *direct sum* $\mathsf{V} \oplus \mathsf{W}$ is the finite-PVS $\{V_j \oplus W_j\}_{j=1}^n$ and linear maps

$$L_j^{\mathsf{V}\oplus\mathsf{W}} = L_j^{\mathsf{V}} \oplus L_j^{\mathsf{W}}.$$

Morphisms

A *morphism* from V to W is a collection $\ell = \{\ell_j\}$ of linear maps $\ell_j : V_j \to W_j$ such that the following diagram commutes for each j:

$$\begin{array}{ccc} V_j & \xrightarrow{L_j^{\mathsf{V}}} & V_{j+1} \\ {\scriptstyle \ell_j}\downarrow & & \downarrow{\scriptstyle \ell_{j+1}} \\ W_j & \xrightarrow{L_j^{\mathsf{W}}} & W_{j+1} \end{array}$$

A morphism ℓ is called an *isomorphism* if every ℓ_j is a linear isomorphism. We leave the proof of the following basic proposition as an exercise.

Proposition 4.4.2. *Isomorphism is an equivalence relation on the collection of finite persistence vector spaces with the same indexing set.*

Subspaces

A *subspace* of a V is a finite-PVS W (with the same indexing set) where each W_j is a subspace of V_j and the restricted maps satisfy

$$\text{image}\left(L_j^{\mathsf{V}}|_{W_j}\right) \subset W_{j+1}.$$

Quotients

Let $\mathsf{V} = \{V_j\}_{j=0}^N$ be a finite-PVS and W a subspace of V. The *quotient persistence vector space* is the finite-PVS $\mathsf{V}/\mathsf{W} = \{V_j/W_j\}_{j=0}^N$ with linear maps

$$L_j^{\mathsf{V}/\mathsf{W}} : V_j/W_j \to V_{j+1}/W_{j+1}$$

defined by

$$L_j^{\mathsf{V}/\mathsf{W}}([v]) = \left[L_j^{\mathsf{V}}(v)\right].$$

These maps are well-defined by the definition of subspace.

Kernels, Images and Cokernels

Let $\ell : \mathsf{V} \to \mathsf{W}$ be a morphism of persistence vector spaces. There are two natural subspaces associated to ℓ. The *kernel* is the subspace of V defined by

$$\ker(\ell_1) \xrightarrow{L_1^{\mathsf{V}}|_{\ker(\ell_1)}} \ker(\ell_2) \xrightarrow{L_2^{\mathsf{V}}|_{\ker(\ell_2)}} \cdots \xrightarrow{L_{n-1}^{\mathsf{V}}|_{\ker(\ell_{n-1})}} \ker(\ell_n)$$

The *image* is the subspace of W defined by

$$\text{im}(\ell_1) \xrightarrow{L_1^{\mathsf{W}}|_{\text{im}(\ell_1)}} \text{im}(\ell_2) \xrightarrow{L_2^{\mathsf{W}}|_{\text{im}(\ell_2)}} \cdots \xrightarrow{L_{n-1}^{\mathsf{W}}|_{\text{im}(\ell_{n-1})}} \text{im}(\ell_n)$$

We also define the *cokernel* of ℓ to be the persistence vector space $\mathsf{W}/\text{im}(\ell)$.

We leave it as an exercise to check that each of these persistence vector spaces is well-defined.

4.5 Structure Theorem for Finite Persistence Vector Spaces

Recall Theorem 1.3.6: finite-dimensional vector spaces V and W (over the same field) are linearly isomorphic if and only if $\dim(V) = \dim(W)$. This is an example of a *classification theorem*—up to a natural notion of equivalence, vector spaces are completely determined by a single positive integer. Our goal for this section is to prove similar theorems in the setting of persistence vector spaces. The basic question is: is there a finite list of (integer-valued?) invariants which completely characterize a persistence vector space, up to isomorphism of persistence vector spaces?

The answer to the basic question is "yes" for finite and finitely-presented persistence vector spaces. The theorem stating this is sometimes referred to as the *Fundamental Theorem of Persistent Homology*. This chapter will be devoted to proving the theorem.

4.5.1 Statement of the Theorem

Finitely-Presented Persistence Vector Spaces

For $0 \leqslant b < d$, let $\mathsf{I}[b, d)$ denote the *interval module* determined by b and d. This is the ($\mathbb{R}$-indexed) persistence vector space over the field $\mathbb{F}$ with

$$\mathsf{I}[b,d)_r = \begin{cases} \mathbb{F} & r \in [b,d) \\ \{0\} & \text{otherwise.} \end{cases}$$

The linear maps $L_{r,s}$ of the interval module are given by

$$L_{r,s} = \begin{cases} \mathrm{id}_{\mathbb{F}} : \mathbb{F} \to \mathbb{F} & r, s \in [b,d) \\ 0 & \text{otherwise.} \end{cases}$$

The main theorem of this chapter, and one of the most important theorems in the theory of persistent homology is the following.

Theorem 4.5.1 (Classification Theorem for Finitely-presented Persistence Vector Spaces)**.** *Let $\mathsf{V} = \{V_r\}_{r \geqslant 0}$ be a finitely-presented persistence vector space. Then there exists a collection of $0 \leqslant b_j < d_j$, $j \in \{1, \ldots, M\}$, such that*

$$\mathsf{V} \approx \mathsf{I}[b_1, d_1) \oplus \mathsf{I}[b_2, d_2) \oplus \cdots \oplus \mathsf{I}[b_M, d_M).$$

Moreover, this decomposition is unique up to reordering the summands.

Our main example of a finitely-presented persistence vector space is the kth persistent homology of a finitely-presented filtered simplicial complex. We have the following corollary.

Corollary 4.5.2. *Let K be a finitely-presented filtered simplicial complex. For any k, the k-th persistent homology $PH_k(\mathsf{K})$ can be represented uniquely (up to order of the summands) as*

$$PH_k(\mathsf{K}) \approx \mathsf{I}[b_1, d_1) \oplus \mathsf{I}[b_2, d_2) \oplus \cdots \oplus \mathsf{I}[b_M, d_M)$$

for some collection of $0 \leqslant b_j < d_j$.

The decomposition guaranteed by this corollary is called the *barcode* of $PH_k(\mathsf{K})$. As an example, if $\mathsf{K} = \mathsf{VR}(X)$ for some finite metric space X, then the barcode gives a representation of the multiscale topology of the metric space. Moreover, the representation is surprisingly simple: it is encoded by a finite collection of pairs of points (b_j, d_j).

Finite Persistence Vector Spaces

Since Theorem 4.5.1 treats finitely-presented persistence vector spaces, it has a corresponding formulation in terms of finite persistence vector spaces of the form $\mathsf{V} = \{V_{r_j}\}_{j=1}^{N}$.

For the rest of this chapter, the particular indexing set is not especially relevant. We thus represented each finite-PVS with the shorthand

$$\mathsf{V} = V_1 \xrightarrow{L_1} V_2 \xrightarrow{L_2} \cdots \xrightarrow{L_{N-1}} V_N$$

and use $L_{j,k}$ to denote the map

$$L_{k-1} \circ \cdots \circ L_{j+1} \circ L_j : V_j \to V_k$$

for each $j < k$.

For integers $1 \leqslant b < d \leqslant N$, we define the *finite interval module* $\mathsf{I}[b, d)$ to be the persistence vector space over $\mathbb{F}$ given by

$$\mathsf{I}[b, d) = 0 \to 0 \to \cdots \to 0 \to \mathbb{F} \to \mathbb{F} \to \cdots \to \mathbb{F} \to 0 \to 0 \to \cdots \to 0,$$

where the first $\mathbb{F}$ appears at index b and the last appears at index $d-1$ (so $\mathbb{F}$ is sent to $\{0\}$ at index d, where it "dies"). All maps between nonzero vector spaces are the identity map. It will be convenient to use the notation

$$\mathsf{I}[b, d+1) = 0 \to 0 \to \cdots \to 0 \to \mathbb{F} \to \mathbb{F} \to \cdots \to \mathbb{F}.$$

That is, $\mathbb{F}$ appears at every index greater than or equal to b.

We have the following finite reformulation of Theorem 4.5.1.

Theorem 4.5.3 (Classification Theorem for Finite Persistence Vector Spaces)**.** *Let* $\mathsf{V} = \{V_j\}_{j=1}^N$ *be a finite persistence vector space. Then there exists a collection of* $b_j, d_j \in \{1, \ldots, N+1\}$ *with* $b_j < d_j$ *and* $j = 1, \ldots, M$*, such that*

$$\mathsf{V} \approx \mathsf{I}[b_1, d_1) \oplus \mathsf{I}[b_2, d_2) \oplus \cdots \oplus \mathsf{I}[b_M, d_M).$$

Moreover, this decomposition is unique up to reordering the summands.

This is the version of the theorem that we will prove. Before proceeding, we note the following corollary. This shows that finite-PVS are classified by a list of integers (cf. the standard classification of finite-dimensional vector spaces, Proposition 1.3.6).

Corollary 4.5.4. *A length-N persistence vector space* $\mathsf{V} = V_1 \xrightarrow{L_1} \cdots \xrightarrow{L_{N-1}} V_N$ *is classified up to isomorphism by the* $\frac{N(N-1)}{2}$ *integers in the set*

$$\{\operatorname{rank}(L_{i,j}) \mid i \leqslant j\},$$

where $\operatorname{rank}(L_{i,i})$ *is taken to be* $\dim(V_i)$.

Proof. Using Theorem 4.5.3, we write

$$\mathsf{V} \approx \bigoplus_{j=1}^{M} \mathsf{I}[b_j, d_j).$$

Any such decomposition (unique up to rearranging the summands) determines the list of ranks $\{\text{rank}(L_{i,j}) \mid i \leqslant j\}$. We leave it as an exercise to determine a closed-form expression for the rank list from the decomposition. □

Example 4.5.1. Let V be the persistence vector space $V_1 \xrightarrow{L} V_2$ over $\mathbb{R}$ with $V_1 = R^2$, $V_2 = \mathbb{R}^3$ and

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

with the matrix expressed in standard bases. Then

$$\mathsf{V} \approx \mathsf{I}[1,3) \oplus \mathsf{I}[1,2) \oplus \mathsf{I}[2,3) \oplus \mathsf{I}[2,3).$$

Let $\mathbf{e}_j$ denote the standard basis vectors for $\mathbb{R}^k$. The first summand in this decomposition represents the fact that $L(\mathbf{e}_1) = \mathbf{e}_1 \neq \vec{0}$; said differently, it corresponds to the image of L. The second summand represents the kernel of L, i.e., $L(\mathbf{e}_2) = \vec{0}$. Finally, the last two summands correspond to the fact that the vectors $\mathbf{e}_2$ and $\mathbf{e}_3$ in V_2 are not in the image of L. Thus they represent the cokernel of L.

Example 4.5.2. In general, for a length-2 persistence vector space $\mathsf{V} = V_1 \xrightarrow{L} V_2$, we have

$$\mathsf{V} = \bigoplus_{j=1}^{\ell} \mathsf{I}[1,3) \oplus \bigoplus_{j=1}^{m} \mathsf{I}[1,2) \oplus \bigoplus_{j=1}^{n} \mathsf{I}[2,3),$$

where $\ell = \dim(\text{image}(L))$, $m = \dim(\ker(L))$ and $n = \dim(\text{coker}(L))$. In accordance with the corollary, this data can be recovered from the three integers $\dim(V_1)$, $\dim(V_2)$ and $\text{rank}(L)$.

4.5.2 Proof of the Classification Theorem

Throughout this section, let

$$\mathsf{V} = V_1 \xrightarrow{L_1} V_2 \xrightarrow{L_2} \cdots \xrightarrow{L_{N-1}} V_N$$

be a finite persistence vector space over the field $\mathbb{F}$.

Preliminary Lemmas

We will require some basic lemmas about morphisms of finite persistence vector spaces.

Lemma 4.5.5. *Let $\mathsf{V}^{(1)}, \mathsf{V}^{(2)}, \mathsf{W}^{(1)}, \mathsf{W}^{(2)}$ denote length-N finite persistence vector spaces. If $\mathsf{V}^{(j)} \approx \mathsf{W}^{(j)}$ for $j = 1, 2$, then $\mathsf{V}^{(1)} \oplus \mathsf{V}^{(2)} \approx \mathsf{W}^{(1)} \oplus \mathsf{W}^{(2)}$.*

Proof. Choose isomorphisms $\ell^{(j)} : \mathsf{V}^{(j)} \to \mathsf{W}^{(j)}$ for $j = 1, 2$. It is easy to check that $\ell^{(1)} \oplus \ell^{(2)}$ gives an isomorphism of the direct sums. □

We leave the proof of the following lemma as an exercise.

Lemma 4.5.6. *An isomorphism $\ell : \mathsf{V} \to \mathsf{W}$ of finite persistence vector spaces induces isomorphisms* $\ker(L^{\mathsf{V}}_{i,j}) \approx \ker(L^{\mathsf{W}}_{i,j})$, $\operatorname{im}(L^{\mathsf{V}}_{i,j}) \approx \operatorname{im}(L^{\mathsf{W}}_{i,j})$ *and* $\operatorname{coker}(L^{\mathsf{V}}_{i,j}) \approx \operatorname{coker}(L^{\mathsf{W}}_{i,j})$.

Splitting Lemma

The proof of Theorem 4.5.3 will follow from a main technical lemma.

Lemma 4.5.7 (Splitting Lemma)**.** *Let $\boldsymbol{v} \in V_1$ and let $\mathsf{U} = \{U_j\}_{j=1}^N$ denote the sub-PVS with*

$$U_j = \operatorname{span}\{L_{1,j}(\boldsymbol{v})\}.$$

Then

$$\mathsf{V} \approx \mathsf{U} \oplus \mathsf{V}/\mathsf{U}.$$

Our proof the Splitting Lemma will use its own technical lemma.

Lemma 4.5.8. *Let V, $\boldsymbol{v}$ and U be as in the statement of Lemma 4.5.7. Let*

$$d = \min\{j \mid L_{1,j}(\boldsymbol{v}) = \vec{0}\},$$

if $L_{1,N}(\boldsymbol{v}) = \vec{0}$; otherwise, we set $d = N + 1$. There exists a bilinear form $\langle \cdot, \cdot \rangle_j$ on each V_j such that

1. *For all $j < d - 1$ and for all $\boldsymbol{w} \in V_j$,*

$$\langle \boldsymbol{w}, L_{1,j}(\boldsymbol{v}) \rangle_j = \langle L_j(\boldsymbol{w}), L_{1,j+1}(\boldsymbol{v}) \rangle_{j+1},$$

2. *For all j,*

$$\langle L_{1,j}(\boldsymbol{v}), L_{1,j}(\boldsymbol{v}) \rangle_j = 0 \Leftrightarrow L_{1,j}(\boldsymbol{v}) = \vec{0}.$$

Proof. For $j \geqslant d - 1$, choose arbitrary inner products $\langle \cdot, \cdot \rangle_j$ on V_j. For $j < d - 1$, define the bilinear form on $\mathbf{u}, \mathbf{w} \in V_j$ by

$$\langle \mathbf{u}, \mathbf{w} \rangle_j = \langle L_{j,d-1}(\mathbf{u}), L_{j,d-1}(\mathbf{w}) \rangle_{d-1}.$$

We need to check that these bilinear forms satisfy the desired properties.

Let $j < d - 1$ and let $\mathbf{w} \in V_j$. Then

$$\begin{aligned} \langle \mathbf{w}, L_{1,j}(\mathbf{v}) \rangle_j &= \langle L_{j,d-1}(\mathbf{w}), L_{j,d-1} \circ L_{1,j}(\mathbf{v}) \rangle_{d-1} \\ &= \langle L_{j+1,d-1} \circ L_j(\mathbf{w}), L_{j+1,d-1} \circ L_{1,j+1}(\mathbf{v}) \rangle_{d-1} \\ &= \langle L_j(\mathbf{w}), L_{1,j+1}(\mathbf{v}) \rangle_{j+1}. \end{aligned}$$

Thus our bilinear forms satisfy the first property. To prove that they satisfy the second property, note that $L_{1,j}(\mathbf{v}) = \vec{0}$ if and only if $j \geqslant d$ and

$$\langle L_{1,j}(\mathbf{v}), L_{1,j}(\mathbf{v}) \rangle_j = \begin{cases} \langle L_{1,d-1}(\mathbf{v}), L_{1,d-1}(\mathbf{v}) \rangle_{d-1} & j \leqslant d - 1 \\ 0 & j \geqslant d, \end{cases}$$

and this is 0 if and only if $j \geqslant d$, since $\langle \cdot, \cdot \rangle_{d-1}$ is an inner product. □

We now proceed with the proof of the main technical lemma.

Proof of Lemma 4.5.7. Choose bilinear forms $\langle \cdot, \cdot \rangle_j$ on each V_j satisfying the properties of Lemma 4.5.8. For each j, we define a linear map $\ell_j : V_j \to U_j \oplus V_j/U_j$ by

$$\ell_j(\mathbf{w}) = \left(\langle \mathbf{w}, L_{1,j}(\mathbf{v}) \rangle_j \, L_{1,j}(\mathbf{v}), [\mathbf{w}] \right).$$

We claim that each such linear map is a linear isomorphism. Indeed, this is clear when $j \geqslant d$ as it is just the map

$$\mathbf{w} \mapsto (\vec{0}, [\mathbf{w}]) \in \{0\} \oplus V_j/\{0\}.$$

For the $j < d$ case, suppose that $\mathbf{w} \in \ker(\ell_j)$. Then

$$\langle \mathbf{w}, L_{1,j}(\mathbf{v}) \rangle_j \, L_{1,j}(\mathbf{v}) = \vec{0} \tag{4.2}$$

and

$$[\mathbf{w}] = \vec{0}. \tag{4.3}$$

Equation (4.3) tells us that $\mathbf{w} \in U_j = \operatorname{span}\{L_{1,j}(\mathbf{v})\}$, so $\mathbf{w} = \lambda L_{1,j}(\mathbf{v})$. Plugging this into (4.3) yields

$$\vec{0} = \lambda \, \langle L_{1,j}(\mathbf{v}), L_{1,j}(\mathbf{v}) \rangle_j \, L_{1,j}(\mathbf{v}),$$

which implies $\lambda = 0$ by our assumption that $j < d$ (so $L_{1,j}(\mathbf{v}) \neq \vec{0}$) and by the second property of $\langle \cdot, \cdot \rangle_j$. Thus $\mathbf{w} = \vec{0}$ and ℓ_j is injective. By dimension counting, we see that ℓ_j is an isomorphism.

It remains to show that the collection $\ell = \{\ell_j\}$ defines a morphism of persistence vector spaces. This means we need to show that the commutativity property

$$L_j^{\mathsf{U} \oplus \mathsf{V}/\mathsf{U}} \circ \ell_j = \ell_{j+1} \circ L_j$$

holds for all j. We first prove the claim for $j < d - 1$. For any $\mathbf{w} \in V_j$, we have

$$\begin{aligned}
L_j^{\mathsf{U} \oplus \mathsf{V}/\mathsf{U}} \circ \ell_j(\mathbf{w}) &= L_j|_{U_j} \oplus L_j^{\mathsf{V}/\mathsf{U}} \left(\langle \mathbf{w}, L_{1,j}(\mathbf{v}) \rangle_j \, L_{1,j}(\mathbf{v}), [\mathbf{w}] \right) \\
&= \left(\langle \mathbf{w}, L_{1,j}(\mathbf{v}) \rangle_j \, L_{1,j+1}(\mathbf{v}), [L_j(\mathbf{w})] \right) \\
&= \left(\langle L_j(\mathbf{w}), L_{1,j+1}(\mathbf{v}) \rangle_{j+1} \, L_{1,j+1}(\mathbf{v}), [L_j(\mathbf{w})] \right) \\
&= \ell_{j+1} \circ L_j(\mathbf{w}),
\end{aligned}$$

where the second-to-last equality follows by the first property of $\langle \cdot, \cdot \rangle_j$. In the remaining case $j \geqslant d-1$, we have $L_{1,j}(\mathbf{v}) = \vec{0}$. It follows that for any $\mathbf{w} \in V_j$,

$$\begin{aligned} L_j^{\mathsf{U} \oplus \mathsf{V}/\mathsf{U}} \circ \ell_j(\mathbf{w}) &= L_j|_{U_j} \oplus L_j^{\mathsf{V}/\mathsf{U}} \left(\langle \mathbf{w}, L_{1,j}(\mathbf{v}) \rangle_j \, L_{1,j}(\mathbf{v}), [\mathbf{w}] \right) \\ &= \left(\vec{0}, [L_j(\mathbf{w})] \right) \\ &= \ell_{j+1} \circ L_j(\mathbf{w}). \end{aligned}$$

□

Corollary 4.5.9. *Suppose $V_1 \neq \{0\}$. For any nonzero $\boldsymbol{v} \in V_1$, let d be the smallest index such that $L_{1,d}(\boldsymbol{v}) = \vec{0}$ (taking $d = N+1$ if $L_{1,N}(\boldsymbol{v}) \neq \vec{0}$). With U as defined in Lemma 4.5.7,*

$$\mathsf{V} \approx \mathsf{I}[1, d) \oplus \mathsf{V}/\mathsf{U}.$$

Proof. We observe that $\mathsf{U} \approx \mathsf{I}[1, d)$ via the morphism defined by $\ell_j(L_{1,j}(\mathbf{v})) = 1_{\mathbb{F}}$ when $j \leqslant d$ and $\ell_j = 0$ otherwise. The result then follows immediately by combining Lemmas 4.5.7 and 4.5.5. □

Proof of the Main Theorem

We now prove Theorem 4.5.3.

Proof. We will induct on the *total dimension* of V; that is, on the quantity

$$D = \dim(V_1) + \dim(V_2) + \cdots + \dim(V_n).$$

For the base case $D = 1$, the only possible length-N persistence vector space is isomorphic to $\mathsf{I}[b, b+1)$ for some $b \in \{1, \ldots, N\}$. Now assume that the claim holds for all persistence vector spaces with total dimension less than or equal to $D-1$ and let V have total dimension D. Without loss of generality, suppose that $\mathbf{v} \in V_1$ is nonzero (if $V_1 = \{0\}$, the rest of the proof can be adapted by shifting to begin at the first index where a nonzero vector space appears). Let $d \in \{1, \ldots, N+1\}$ and U be as defined in Corollary 4.5.9 and Lemma 4.5.7, respectively. By the corollary,

$$\mathsf{V} \approx \mathsf{I}[1, d) \oplus \mathsf{V}/\mathsf{U}.$$

The persistence vector space $\mathsf{V}/\mathsf{U} = \{V_j/U_j\}_{j=1}^N$ as total dimension at most $D-1$. Indeed, $\dim(V_1/\mathrm{span}(\mathbf{v})) = \dim(V_1) - 1$ and for all $j > 1$, $\dim(V_j/U_j) \leqslant \dim(V_j)$. By the inductive hypothesis,

$$\mathsf{V}/\mathsf{U} \approx \bigoplus_{j=1}^{M} \mathsf{I}[b_j, d_j)$$

for some b_j, d_j. The existence part of the theorem therefore follows by induction after applying Lemma 4.5.5.

We now wish to prove the uniqueness part of the theorem. Suppose that

$$\mathsf{V} \approx \bigoplus_{j=1}^{M} \mathsf{I}[b_j, d_j) \approx \bigoplus_{j=1}^{M'} \mathsf{I}[b'_j, d'_j). \tag{4.4}$$

We continue to assume (without loss of generality) that $V_1 \neq \{0\}$ and we will once again induct on total dimension. The base case $D = 1$ is already covered, as this implies $\mathsf{V} \approx \mathsf{I}[b, b+1)$ and it is clear that no other direct sum of interval modules is isomorphic to $I[b, b+1)$.

Suppose the claim holds for modules with total dimension at most $D-1$ and consider a persistence vector space V of total dimension D satisfying (4.4). Consider the "first death times"

$$d = \min\{d_j \mid b_j = 1\} \quad \text{and} \quad d' = \min\{d'_j \mid b'_j = 1\}.$$

These quantities are forced to be equal, since they can be intrinsically expressed as

$$d = d' = \min\{j \mid \ker(L_{1,j}) \neq \{0\}\},$$

and isomorphisms of persistence vector spaces induce isomorphisms of kernels of the maps $L_{i,j}$ (Lemma 4.5.6). Next we wish to show that the two decompositions of V contain the same number of copies of $\mathsf{I}[1, d)$; that is, we can rearrange the ordering of our summands to write

$$\mathsf{V} \approx \bigoplus_{j=1}^{k} \mathsf{I}[1, d) \oplus \bigoplus_{j=k+1}^{M} \mathsf{I}[b_j, d_j) \approx \bigoplus_{j=1}^{k'} \mathsf{I}[1, d) \oplus \bigoplus_{j=k'+1}^{M'} \mathsf{I}[b'_j, d'_j),$$

and we once again have an intrinsic characterization of k, k' as

$$k = k' = \text{null}(L_{1,d}).$$

Now consider

$$\bigoplus_{j=k+1}^{M} \mathsf{I}[b_j, d_j) \approx \bigoplus_{j=k'+1}^{M'} \mathsf{I}[b'_j, d'_j).$$

Both of these persistence vector spaces have total dimension strictly less than D. By the inductive hypothesis, these spaces only differ up to a rearrangement of summands and this completes the proof. □

4.6 Barcodes and Persistence Diagrams

Barcodes

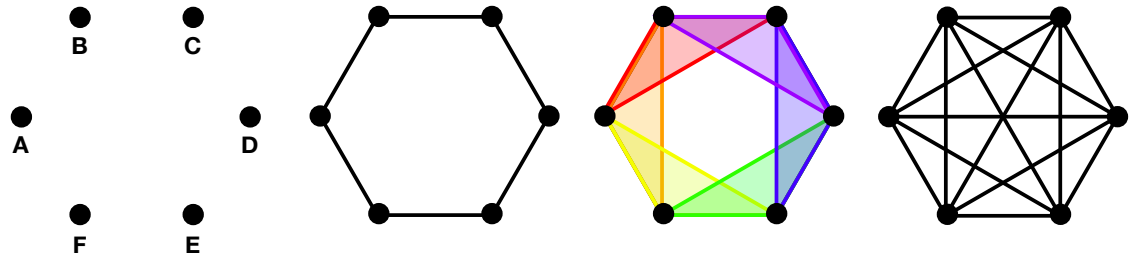
We can now associate to any finitely-presented filtered simplicial complex $\mathsf{K} = \{K_r\}$ a topological signature called a *barcode* for each integer $k \geqslant 0$. To do so, we calculate the persistence homology of $PH_k(\mathsf{K})$ then apply Theorem 4.5.3 to obtain a (unique up to reordering) decomposition

$$PH_k(\mathsf{K}) \approx \mathsf{I}[b_1, d_1) \oplus \mathsf{I}[b_2, d_2) \oplus \cdots \oplus \mathsf{I}[b_M, d_M).$$

Recording the pairs (b_j, d_j), we can represent this decomposition as a multiset $\{(b_j, d_j)\}_{j=1}^{M}$. That this is a multiset means that the points must be counted with multiplicity (we are abusing notation and denoting it with set notation, however). The barcode is a graphical representation of the multiset, consisting of a collection of horizontal line segments above an r-axis. Each line segment (or *bar*) in the barcode represents a point in the multiset: the bar representing (b_j, d_j) has its left endpoint at $r = b_j$ and its right endpoint at $r = d_j$.

Examples of Barcodes

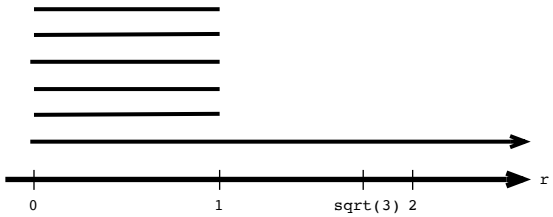
For the first example, let X denote the vertices of a regular hexagon. The Vietoris-Rips complex of X is pictured below at the relevant parameters $r = 0, 1, \sqrt{3}, 2$



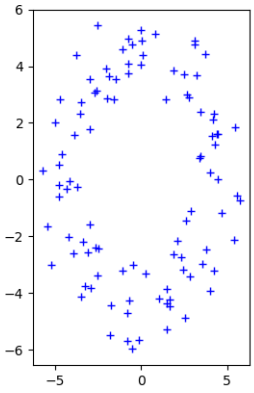
The degree-0 persistent homology decomposes as

$$\mathsf{I}[0,\infty) \oplus \mathsf{I}[0,1) \oplus \mathsf{I}[0,1) \oplus \mathsf{I}[0,1) \oplus \mathsf{I}[0,1) \oplus \mathsf{I}[0,1).$$

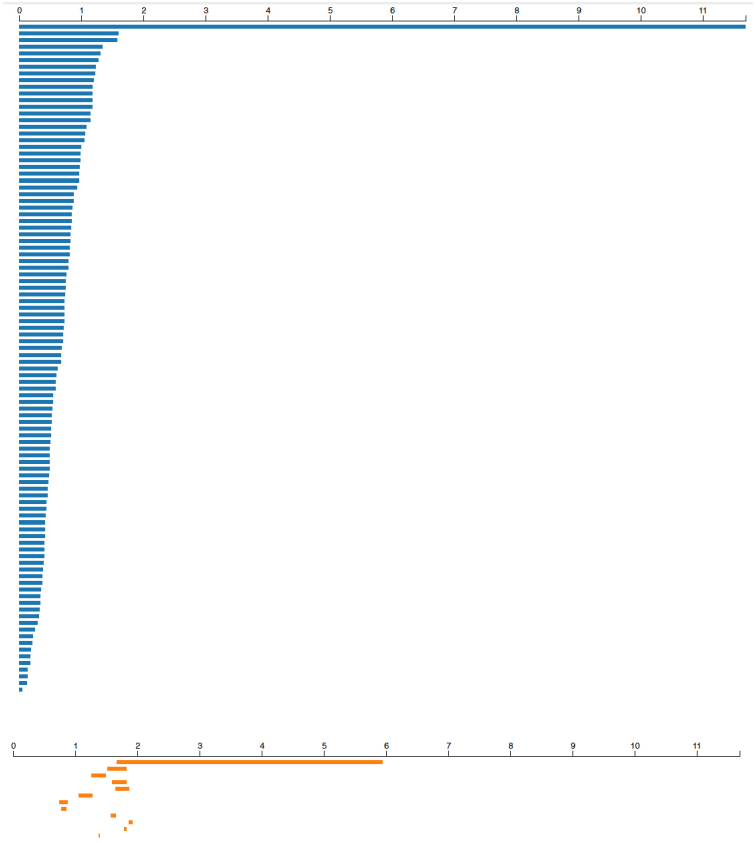
The barcode for this example is shown below.



As another example, consider the point cloud shown below.



There are too many points to compute the Vietoris-Rips complex and the barcodes for this space by hand. We use the software *Ripser* [2] to obtain the following barcodes for degree 0 and degree 1 persistent homology, respectively.



Observe that the original point cloud was essentially a "noisy circle". The barcodes pick up the fact that there is a single persistent homology class in each dimension.

4.6.1 Persistence Diagrams

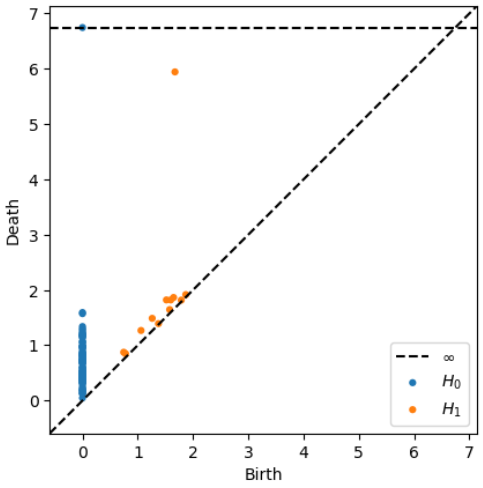
The circle example above shows that a more compact visual representation of persistent homology could be useful. The *persistence diagram* of a finitely-presented persistence vector space

$$\mathsf{V} \approx \bigoplus_{j=1}^{M} \mathsf{I}[b_j, d_j)$$

is obtained from the multiset $\{(b_j, d_j)\}_{j=1}^{M}$ by plotting each point in the xy-plane (recording multiplicity as necessary). Since $b_j < d_j$ for each j, the points all appear above the diagonal line $x = y$.

Examples of Persistence Diagrams

The persistence diagram for the noisy circle example is shown below. The diagram for both degree-0 and degree-1 are plotted on the same axes.



4.7 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.

2. Show that if X is a finite metric space with N points, then $\mathsf{VR}(X, r)$ is simplicially isomorphic to Δ^N for all $r \geqslant \operatorname{diam}(X)$.

3. Show that isomorphism of length-n persistence modules is an equivalence relation (Proposition 4.4.2).

4. Show that the kernel, image and cokernel of a morphism of persistence vector spaces is well-defined.

5. Prove that an isomorphism of length-2 persistence modules $V_1 \xrightarrow{f} V_2$ and $W_1 \xrightarrow{g} W_2$ induces isomorphisms of kernels, images, and cokernels.

6. Complete the proof of Theorem **??** by showing that the six-term sequence is exact at $\operatorname{coker}(f_1)$.

7. a) Prove that an isomorphism of length-2 persistence modules $V_1 \xrightarrow{f} V_2$ and $W_1 \xrightarrow{g} W_2$ induces isomorphisms of kernels, images, and cokernels of the maps f and g (i.e., $\ker(f) \approx \ker(g)$ as vector spaces).

   b) Prove the corresponding statement for general length-n persistence modules (i.e., Proposition **??**).

8. Prove Lemma **??**: if $0 \to U \xrightarrow{\iota} V \xrightarrow{p} W \to 0$ is a short exact sequence of vector spaces and $s : W \to V$ is a splitting of p, then the maps ι and s induce an isomorphism $V \approx U \oplus W$.

9. Prove Lemma **??** (the persistence version of the previous exercise).

10. Prove Lemma 4.5.6.

11. Corollary 4.5.4 states that a persistence vector space is determined by its list of ranks $\{\operatorname{rank}(f_{i,j}) \mid i \leqslant j\}$. Given a direct sum of birth-death persistence vector spaces $\bigoplus_{j=1}^{N} I(b_j, d_j)$, find a closed-form expression for its list of ranks. (Hint: the expression should depend on the multiplicity of each $I(i, j)$ in the decomposition).

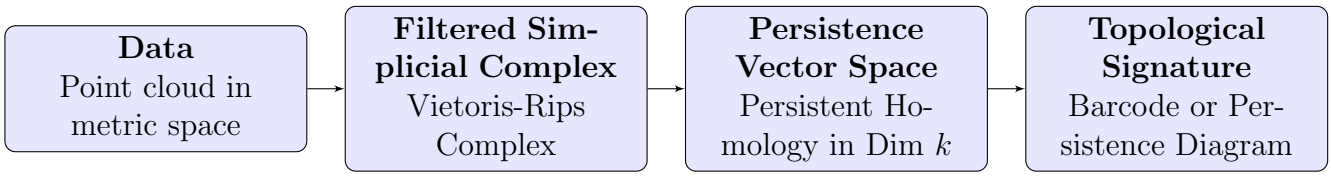
12. Complete the proof of Proposition **??** by showing that the remaining operation preserves the property of being (ρ, σ)-adapted.

13. Prove Proposition **??**.

5 Metrics on the Space of Barcodes

5.1 Reviewing the TDA Pipeline

The basic pipeline that we have developed so far for the topological data analysis framework is summarized below.



The obvious question remains: how do we compare topological signatures produced by a pair of datasets? The answer to this question is to consider the space of all topological signatures as a topological space itself! To do so we define a metric which measures distance between the signatures. The metric should satisfy two requirements:

1. it should be mathematically natural;
2. it should be easily computable.

In this chapter, we study a pair of metrics on persistence vector spaces. The first, the so-called *bottleneck distance*, is defined in terms of matching points in persistence diagrams. The second is called *interleaving distance* and is defined in a more direct way on persistence vector spaces. Finally we show that these two metrics are actually equivalent. This has several theoretical consequences, which we outline after proving the main theorem.

5.2 Bottleneck Distance for Persistence Diagrams

5.2.1 Notation

Let us lay out the basic notation that we will follow for the rest of this section. Throughout, let V and W be finitely-presented persistence vector spaces with interval module decompositions

$$\mathsf{V} \approx \bigoplus_{j=1}^{M} \mathsf{I}[b_j, d_j) \quad \text{and} \quad \mathsf{W} \approx \bigoplus_{j=1}^{M'} \mathsf{I}[b'_j, d'_j).$$

Let

$$\mathcal{D}(\mathsf{V}) = \{(b_j, d_j)\}_{j=1}^{M}$$

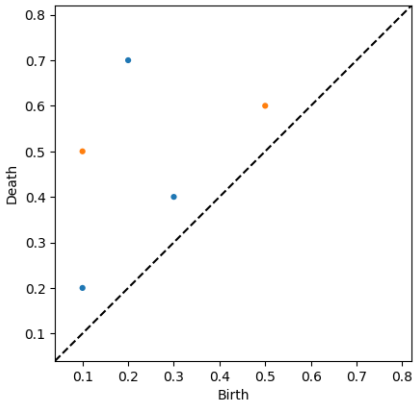


Figure 5.1: Persistence diagrams for V (blue) and W (orange).

denote the *persistence diagram of* V. This is the multiset of points in $\mathbb{R}^2$ given by the endpoints of the intervals in the interval module decomposition of V. We note that this is an abuse of notation—the persistence diagram is really a multiset because several copies of the same interval module can appear in the decomposition of V. The points in $\mathcal{D}(\mathsf{V})$ must therefore be considered with multiplicity. It will be convenient to introduce the notation $\alpha_j = (b_j, d_j)$ and $\beta_j = (b'_j, d'_j)$, so that

$$\mathcal{D}(\mathsf{V}) = \{\alpha_j\}_{j=1}^{M} \quad \text{and} \quad \mathcal{D}(\mathsf{W}) = \{\beta_j\}_{j=1}^{M'}.$$

5.2.2 Defining Bottleneck Distance

Motivation

Recall from Section 4.6.1 that the persistence diagram $\mathcal{D}(\mathsf{V})$ gives a visual summary of V by plotting the points in the plane. Also recall that, since $b_j < d_j$, all points in this plot lie above the line $x = y$ in the plane. Also note that an interval module decomposition can contain interval modules of the form $\mathsf{I}[b, \infty)$, so we must allow $\mathcal{D}(\mathsf{V})$ to contain points in the *extended* plane whose second coordinates are allowed to be infinity.

Our main question is: how do we quantitatively compare the persistence diagrams $\mathcal{D}(\mathsf{V})$ and $\mathcal{D}(\mathsf{W})$? Throughout our discussion, let us consider the following running example.

Example 5.2.1 (Running Example)**.** Let

$$\mathsf{V} = \mathsf{I}[0.1, 0.2) \oplus \mathsf{I}[0.2, 0.7) \oplus \mathsf{I}[0.3, 0.4)$$

and

$$\mathsf{W} = \mathsf{I}[0.1, 0.5) \oplus \mathsf{I}[0.5, 0.6).$$

The persistence diagrams are shown in Figure 5.1.

First Attempt

As a first idea, we might try to "optimally match" the points in the diagrams. That is, for each $\alpha \in \mathcal{D}(\mathsf{V})$ and $\beta \in \mathcal{D}(\mathsf{W})$, we define some cost function $c(\alpha, \beta)$. We then attempt to pair α's with β's so that the total cost is as low as possible. We easily see from our Running Example that this is not possible—persistence diagrams do not necessarily even have the same number of points!

Refining the Idea

The idea of matching can be placed on solid footing by allowing points in either diagram to be matched with "phantom points" lying on the diagonal line $x = y$. These points can be understood as interval modules of the form $\mathsf{I}[b, b)$ (this is really just a different notation for the zero module $\mathsf{0} = \{\{0\}\}_{r \geqslant 0}$). We could then refine the matching idea by introducing a pair of cost functions: for each $\alpha \in \mathcal{D}(\mathsf{V})$ and $\beta \in \mathcal{D}(\mathsf{W})$, define a cost function $c_1(\alpha, \beta)$—the cost of matching α to β—as well as a cost function $c_2(\alpha)$ (respectively, $c_2(\beta)$)—the cost of matching the point α (respectively, β) to a phantom point on the diagonal line. We can then define a distance between $\mathcal{D}(\mathsf{V})$ and $\mathcal{D}(\mathsf{W})$ as follows.

A *partial bijection* from $\mathcal{D}(\mathsf{V})$ to $\mathcal{D}(\mathsf{W})$ consists of:

- a subset $A \subset \mathcal{D}(\mathsf{V})$,
- a subset $B \subset \mathcal{D}(\mathsf{W})$ with the same number of points as A, and
- a bijection $\phi : A \to B$.

The *matching distance between* $\mathcal{D}(\mathsf{V})$ *and* $\mathcal{D}(\mathsf{W})$ relative to c_1 and c_2 is

$$d_{c_1,c_2}(\mathcal{D}(\mathsf{V}), \mathcal{D}(\mathsf{W})) = \min_{\phi : A \to B} \max \left\{ \max_{\alpha \in A} c_1(\alpha, \phi(\alpha)), \max_{\alpha \notin A} c_2(\alpha), \max_{\beta \notin B} c_2(\beta) \right\},$$

where the minimum is taken over all partial bijections.

The quantity

$$\max \left\{ \max_{\alpha \in A} c_1(\alpha, \phi(\alpha)), \max_{\alpha \notin A} c_2(\alpha), \max_{\beta \notin B} c_2(\beta) \right\},$$

which depends on a choice of partial bijection $\phi : A \to B$, is called the *objective function* for bottleneck distance. We take the convention that the maximum over an empty set is zero; e.g., if $A = \varnothing$, then

$$\max_{\alpha \in A} c_1(\alpha, \phi(\alpha)) := 0.$$

Bottleneck Distance

It turns out that the most natural choice of matching distance is given by taking the cost c_1 to be

$$c_1(\alpha, \beta) = d_\infty(\alpha, \beta) = \|\alpha - \beta\|_\infty$$

and c_2 to be the d_∞-distance to the nearest point on the diagonal. The reader can check the closest point to $\alpha = (b, d)$ is given by

$$c_2(\alpha) = \frac{d-b}{2}.$$

We now denote this quantity by $m(\alpha)$ (for the *midpoint* of α, considered as an interval). We must also define the costs for points in persistence diagrams with second coordinate infinity. For $\alpha = (b, d)$ and $\beta = (b', d')$, we extend our definitions by setting

$$d_\infty(\alpha, \beta) = \begin{cases} \infty & \text{if exactly one of } d \text{ or } d \text{ is } \infty \\ |b - b'| & \text{if } d = d' = \infty \end{cases}$$

and

$$m(\alpha) = \infty$$

if $d = \infty$.

The *bottleneck distance* is the matching distance with respect to these specific cost functions. We use the notation

$$d_b(\mathcal{D}(\mathsf{V}), \mathcal{D}(\mathsf{W})) = \min_{\phi: A \to B} \max \left\{ \max_{\alpha \in A} d_\infty(\alpha, \phi(\alpha)), \max_{\alpha \notin A} m(\alpha), \max_{\beta \notin B} m(\beta) \right\}.$$

We claim that d_b defines a generalized notion of a metric, but we require some definitions to do so precisely.

Let X be a set. An *extended metric* on X is a function

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

such that the usual axioms of a metric hold for all $x, x', x'' \in X$:

1. $d(x, x') \geqslant 0$, with equality if and only if $x = x'$,
2. $d(x, x') = d(x', x)$,
3. $d(x, x'') \leqslant d(x, x') + d(x', x'')$.

The only difference is that each point must be interpreted as allowing the value ∞ to appear. For example, the triangle inequality is satisfied any time that one of the quantities on the righthand side is ∞. If the quantity on the lefthand side is ∞, this forces at least one of the quantities on the righthand side to be ∞.

We use the notation PDiag for the collection of all persistence diagrams. That is,

$$\text{PDiag} := \{\mathcal{D}(\mathsf{V}) \mid \mathsf{V} \text{ is a finitely-presented PVS}\}.$$

We leave the proof of the following proposition as an exercise.

Proposition 5.2.1. *Bottleneck distance d_b defines an extended metric on the set* PVect.

Distance in the Running Example

To compute the distance between the persistence diagrams of V and W in the Running Example (Example 5.2.1), we label the points in $\mathcal{D}(\mathsf{V})$ as

$$\alpha_1 = (0.1, 0.2), \quad \alpha_2 = (0.2, 0.7), \quad \alpha_3 = (0.3, 0.4)$$

and in $\mathcal{D}(\mathsf{W})$ as

$$\beta_1 = (0.1, 0.5), \quad \beta_2 = (0.5, 0.6).$$

We then construct the following *cost matrix*, whose entries give the cost of matching an α_j with a β_k (in which case the entry is $d_\infty(\alpha_j, \beta_k)$), as well as matching an α_j (respectively, β_k) with the diagonal (in which case the entry is $m(\alpha_j)$ (respectively, $m(\beta_k)$)).

$$\partial_1 = \begin{array}{c} \\ \beta_1 \\ \beta_2 \\ diag \end{array} \begin{array}{c} \begin{array}{cccc} \alpha_1 & \alpha_2 & \alpha_3 & diag \end{array} \\ \begin{bmatrix} 0.3 & 0.2 & 0.2 & 0.2 \\ 0.4 & 0.2 & 0.3 & 0.05 \\ 0.05 & 0.05 & 0.25 & X \end{bmatrix} \end{array}.$$

The cost matrix summarizes the data relevant to picking an optimal matching between the diagrams.

Now we consider all possible partial bijections $\phi : A \to B$. There is always the option to match all points in both diagrams with the diagonal. In this case $A = B = \varnothing$, so the objective function for bottleneck distance gives the value

$$\max\left\{\max_{\alpha\in\varnothing} d_\infty(\alpha, \phi(\alpha)), \max_{\alpha\notin A} m(\alpha), \max_{\beta\notin B} m(\beta)\right\} = \{0, 0.2, 0.25\} = 0.25$$

A more economical matching is given by taking $A = \{\alpha_3\}$ and $B = \{\beta_1\}$ so that $\phi(\alpha_3) = \beta_1$. Then the objective function gives

$$\max\left\{\max_{\alpha\in A} d_\infty(\alpha, \phi(\alpha)), \max_{\alpha\notin A} m(\alpha), \max_{\beta\notin B} m(\beta)\right\} = \{0.2, 0.05, 0.2\} = 0.2.$$

The reader can check that this is in fact the optimal matching. We conclude that

$$d_b(\mathcal{D}(\mathsf{V}), \mathcal{D}(\mathsf{W})) = 0.2.$$

5.2.3 Variations on Bottleneck Distance

The choices $c_1 = d_\infty$ and $c_2 = m$ might not initially seem like the most natural choices for matching costs. We will see below that these choices arise because they connect bottleneck distance to a more intrinsically-defined metric on the space of persistence vector spaces called *interleaving distance*. It is natural to study other choices of cost functions, such as $c_1 = d_p$ and $c_2 = m_p$, with $m_p(\alpha)$ denoting the distance from the point α to the closest point on the diagonal with respect to p-distance. The resulting matching distances are called *Wasserstein p-distances* due to their relationship to distances arising

in the theory of optimal transport. When using these p-distance costs, the definition of d_{c_1,c_2} is modified to $d_{c_1,c_2}^{1/p}$ so that the resulting dissimilarity is a true distance metric. The Wasserstein p-distances are also useful in applications, but to keep the discussion focused we will continue to consider bottleneck distance specifically.

5.3 Interleaving Distance for Persistence Vector Spaces

5.3.1 Definition

Let $\mathsf{V} = \{V_r\}_{r\geqslant 0}$ and $\mathsf{W} = \{W_r\}_{r\geqslant 0}$ be finitely-presented persistence vector spaces. Our notion of a morphism between persistence vector spaces is rather rigid in that it demands parameters to be matched up exactly; that is, a morphism $\ell : \mathsf{V} \to \mathsf{W}$ consists of a map $\ell_r : V_r \to W_r$ for all $r \geqslant 0$ satisfying the natural commutativity property $\ell_{r'} \circ L^{\mathsf{V}}_{r,r'} = L^{\mathsf{W}}_{r,r'} \circ \ell_r$; i.e., the following diagram commutes:

$$\begin{array}{ccc} V_r & \xrightarrow{L^{\mathsf{V}}_{r,r'}} & V_{r'} \\ {\scriptstyle \ell_r}\downarrow & & \downarrow{\scriptstyle \ell_{r'}} \\ W_r & \xrightarrow{L^{\mathsf{W}}_{r,r'}} & W_{r'} \end{array}$$

This rigidity results in strange situations where there are persistence vector spaces are intuitively similar in structure, but for which the only morphism between them $\ell_r = 0$ for all r.

Example 5.3.1. Consider the persistence vector spaces $\mathsf{V} = \mathsf{I}[0, 1)$ and $\mathsf{W} = \mathsf{I}[0.1, 1.1)$. The persistence vector spaces have essentially the same structure, but we claim that the only morphism $\ell : \mathsf{V} \to \mathsf{W}$ is the trivial one. Indeed, assume that ℓ_r is not the zero map for some $r \in [0, 1)$. Then $\ell_r \circ L^{\mathsf{V}}_{0,r}$ is nontrivial. On the other hand, $W_0 = \{0\}$, so $\ell_0 : V_0 \to W_0$ must be zero map. It follows that $L^{W}_{0,r} \circ \ell_0$ is the zero map, and this violates the commutativity condition for morphisms. Therefore ℓ is not a morphism.

We therefore wish to relax our notion of morphism. An *ϵ-morphism* $\phi : \mathsf{V} \to \mathsf{W}$ between $\mathbb{R}$-indexed persistence vector spaces consists of a map $\ell_r : V_r \to W_{r+\epsilon}$ for each r. Moreover, we require the maps to satisfy the commutativity property $\phi_{r'} \circ L^{\mathsf{V}}_{r,r'} = L^{\mathsf{W}}_{r+\epsilon,r'+\epsilon} \circ \phi_r$ for all $r \leqslant r'$.

Example 5.3.2. Consider the persistence vector spaces in Example 5.3.1. There is a nontrivial ϵ-morphism $\ell : \mathsf{V} \to \mathsf{W}$ for $\epsilon = 0.1$. There are also nontrivial δ-morphisms for each $0.1 < \epsilon < 1.1$. On the other hand, there is no ϵ-morphism for any $\epsilon < 0.1$, by the same reasoning as in the previous example.

A ϵ-interleaving of (finitely-presented) persistence vector spaces V and W is a pair of ϵ-morphisms $\phi : \mathsf{V} \to \mathsf{W}$ and $\psi : \mathsf{W} \to \mathsf{V}$ such that, for each r,

1. $\psi_{r+\epsilon} \circ \phi_r = L^{\mathsf{V}}_{r,r+2\epsilon}$ and
2. $\phi_{r+\epsilon} \circ \psi_r = L^{\mathsf{W}}_{r,r+2\epsilon}$.

When there exists an ϵ-interleaving of V and W, we say that V and W are *ϵ-interleaved*.

We now define a function d_i, called *interleaving distance*, whose domain consists of pairs of finitely-presented persistence diagrams and whose range is $\mathbb{R}_{\geqslant 0}$. It is defined by the formula

$$d_i(\mathsf{V}, \mathsf{W}) = \inf\{\epsilon \geqslant 0 \mid \mathsf{V} \text{ and } \mathsf{W} \text{ are } \epsilon\text{-interleaved}\}.$$

We make some elementary observations.

Lemma 5.3.1. *1. There exist finitely-presented persistence vector spaces such that no ϵ-interleaving exists for any ϵ. In this case, we declare the interleaving distance to be ∞.*

2. If there exists r_0 such that $V_s = \{0\}$ and $W_s = \{0\}$ for all $s \geqslant r_0$, then there is an ϵ-interleaving for $\epsilon = r_0/2$. It follows that, in this case, $d_i(\mathsf{V}, \mathsf{W})$ is well-defined and finite.

3. In the case that $d_i(\mathsf{V}, \mathsf{W})$ is finite, it is realized by an interleaving; that is, the infimum is actually a minimum.

4. The interleaving distance can be zero; in fact, $d_i(\mathsf{V}, \mathsf{W}) = 0$ if and only if $\mathsf{V} \approx \mathsf{W}$.

Proof. 1. Let $\mathsf{V} = \mathsf{I}[0, \infty)$ and $\mathsf{W} = \mathsf{I}[0, 1)$. Then for any ϵ-morphisms $\phi : \mathsf{V} \to \mathsf{W}$ and $\psi : \mathsf{W} \to \mathsf{V}$

$$\psi_{2+\epsilon} \circ \phi_2 : V_2 \to V_{2+2\epsilon}$$

is necessarily the zero map, whereas $L^{\mathsf{V}}_{2,2+2\epsilon}$ is the identity map on $\mathbb{F}$. In this case, we declare

$$d_i(\mathsf{V}, \mathsf{W}) = \infty.$$

2. In this case, define $\phi_r : V_r \to W_{r+r_0/2} = \{0\}$ and $\psi_r : W_r \to V_{r+r_0/2} = \{0\}$ to be zero maps. This defines an interleaving, since

$$L^{\mathsf{V}}_{r,r+r_0} : V_r \to V_{r+r_0}$$

and

$$L^{\mathsf{W}}_{r,r+r_0} : W_r \to W_{r+r_0}$$

are necessarily zero maps.

3. By the definition of a finitely-presented PVS, there exist finitely many values $r_1 < \cdots < r_N$ such that $L^{\mathsf{V}}_{r,s}$ and $L^{\mathsf{W}}_{r,s}$ are isomorphisms—and, in particular, identity maps—whenever $r, s \in [r_j, r_{j+1})$. Now suppose that $d_i(\mathsf{V}, \mathsf{W}) = \epsilon$ and take a sequence $\epsilon_1, \epsilon_2, \ldots \to \epsilon$ with ϵ_j-interleavings (ϕ^j, ψ^j). For the sake of convenience, assume that the ϵ_j converge monotonically in the sense that $\epsilon_k < \epsilon_j$ whenever $k > j$.

Now fix an r and define $\phi_r : V_r \to W_{r+\epsilon}$ as follows. Take j sufficiently large, so that

$$L^{\mathsf{W}}_{r+\epsilon,r+\epsilon_j} : W_{r+\epsilon} \to W_{r+\epsilon_j}$$

is the identity map then define

$$\phi_r = \left(L^{\mathsf{W}}_{r+\epsilon,r+\epsilon_j}\right)^{-1} \circ \phi^j_r.$$

Note that if we choose any $k > j$, then our assumptions imply

$$\left(L^{\mathsf{W}}_{r+\epsilon,r+\epsilon_k}\right)^{-1} \circ \phi^k_r = \left(L^{\mathsf{W}}_{r+\epsilon,r+\epsilon_j}\right)^{-1} \circ \phi^j_r.$$

We claim that the family $\phi = \{\phi_r\}_{r\geqslant 0}$ is an ϵ-morphism. Each ϕ_r is a linear map, so it remains to show that the commutativity condition holds; i.e., that

$$L^{\mathsf{W}}_{r+\epsilon,s+\epsilon} \circ \phi_r = \phi_s \circ L^{\mathsf{V}}_{r,s}$$

for all $r \leqslant s$. Choose j sufficiently large so that both $L^{\mathsf{W}}_{r+\epsilon,r+\epsilon_j}$ and $L^{\mathsf{W}}_{s+\epsilon,s+\epsilon_j}$ are identity maps. For convenience, let us denote these maps by A and B, respectively. Then we have

$$\begin{aligned}
L^{\mathsf{W}}_{r+\epsilon,s+\epsilon} &= \left(B^{-1} \circ L^{\mathsf{W}}_{r+\epsilon_j,s+\epsilon_j} \circ A\right) \circ \left(A^{-1} \circ \phi^j_r\right) \\
&= B^{-1} \circ \left(L^{\mathsf{W}}_{r+\epsilon_j,s+\epsilon_j} \circ \phi^j_r\right) \\
&= B^{-1} \circ \left(\phi^j_s \circ L^{\mathsf{V}}_{r,s}\right) \\
&= \left(B^{-1} \circ \phi^j_s\right) \circ L^{\mathsf{V}}_{r,s} \\
&= \phi_s \circ L^{\mathsf{V}}_{r,s}.
\end{aligned}$$

We similarly define an ϵ-morphism $\psi = \{\psi_r\}_{r\geqslant 0}$ from W to V by

$$\psi_r = \left(L^{\mathsf{V}}_{r+\epsilon,r+\epsilon_j}\right)^{-1} \circ \psi^j_r,$$

where j is chosen to be sufficiently large. We then need to show that (ϕ, ψ) define an ϵ-interleaving. That is, we must show that $\psi_{r+\epsilon} \circ \phi_r = L^{\mathsf{V}}_{r,r+2\epsilon}$ and $\phi_{r+\epsilon} \circ \psi_r = L^{\mathsf{W}}_{r,r+2\epsilon}$. We will demonstrate the first of these equalities, with the second following similarly. Choose j sufficently large so that $A := L^{\mathsf{W}}_{r+\epsilon,r+\epsilon_j}$ and $C := L^{\mathsf{V}}_{r+2\epsilon,r+2\epsilon_j}$ are identity maps. Then

$$\begin{aligned}
\psi_{r+\epsilon} \circ \phi &= C^{-1} \circ \psi^j_{r+\epsilon_j} \circ A \circ \phi_r \\
&= C^{-1} \circ \psi^j_{r+\epsilon_j} \circ \phi^j_r \\
&= C^{-1} \circ L^{\mathsf{V}}_{r,r+2\epsilon_j} \\
&= L^{\mathsf{V}}_{r,r+2\epsilon}.
\end{aligned}$$

This completes the proof of this part of the lemma.

4. This follows from the previous observation and the fact that an isomorphism $\ell : \mathsf{V} \to \mathsf{W}$ is exactly the same concept as a 0-interleaving.

□

We denote the *set of isomorphism classes of finitely-presented persistence vector* spaces over $\mathbb{F}$ by PVect. Note that PVect is really a quotient set: it contains equivalence classes under the equivalence relation of PVS isomorphism. This means that if finitely-presented persistence vector spaces V and W are isomorphic, then they each represent the same point $[\mathsf{V}] = [\mathsf{W}]$ in PVect.

Theorem 5.3.2. *Interleaving distance d_i induces an extended metric on* PVect.

By abuse of notation, we continue to denote the induced metric by d_i.

Proof. We define

$$d_i([\mathsf{V}], [\mathsf{W}]) = d_i(\mathsf{V}, \mathsf{W}).$$

We first show that this quantity is well-defined. We need to show that if $\mathsf{V} \approx \mathsf{V}'$ and $\mathsf{W} \approx \mathsf{W}'$, then $d_i(\mathsf{V}, \mathsf{W}) = d_i(\mathsf{V}', \mathsf{W}')$. Fix isomorphisms $\ell : \mathsf{V} \to \mathsf{V}'$ and $p : \mathsf{W} \to \mathsf{W}'$. Then for any ϵ-interleaving (ϕ, ψ) of $\mathcal{V}$ and $\mathcal{W}$, we can construct an ϵ-interleaving (ϕ', ψ') of V' and W' via the formula

$$\phi'_r = p_{r+\epsilon} \circ \phi_r \circ \ell_r^{-1}.$$

We can similarly construct an ϵ-interleaving of V and W from any ϵ-interleaving of V' and W'. It follows that the sets we take infimums over to compute the interleaving distances are the same and therefore that $d_i(\mathsf{V}, \mathsf{W}) = d_i(\mathsf{V}', \mathsf{W}')$.

Now we wish to show that the induced function d_i defines a metric. Let $[\mathsf{V}]$ and $[\mathsf{W}]$ be elements of PVect. By Lemma 5.3.1,

$$0 = d_i([\mathsf{V}], [\mathsf{W}]) = d_i(\mathsf{V}, \mathsf{W}) \Leftrightarrow \mathsf{V} \approx \mathsf{W} \Leftrightarrow [\mathsf{V}] = [\mathsf{W}].$$

Symmetry of d_i also follows immediately from the definition.

It remains to show that d_i satisfies the triangle inequality. This can be done at the level of persistence vector spaces, as the result then follows for the induced metric on isomorphism classes. Let U, V and W be finitely-presented persistence vector spaces. We wish to show that

$$d_i(\mathsf{U}, \mathsf{W}) \leqslant d_i(\mathsf{U}, \mathsf{V}) + d_i(\mathsf{V}, \mathsf{W}). \tag{5.1}$$

If either of the interleaving distances on the righthand side of (5.1) are ∞ then we are done, so suppose note. By our observations above, the interleaving distances are realized by interleavings. Denote these by (ϕ^1, ψ^1), an ϵ_1-interleaving of U with V, and (ϕ^2, ψ^2), an ϵ_2-interleaving of V with W. Define

$$\phi_r = \phi^2_{r+\epsilon_1} \circ \phi^1_r$$

and

$$\psi_r = \psi^2_{r+\epsilon_2} \circ \psi^1_r.$$

We claim that (ϕ, ψ) is an $(\epsilon_1 + \epsilon_2)$-interleaving of U with W. Indeed,

$$\phi_r : U_r \xrightarrow{\phi^1_r} V_{r+\epsilon_1} \xrightarrow{\phi^2_{r+\epsilon_1}} W_{r+(\epsilon_1+\epsilon_2)}$$

and

$$\psi_r : W_r \xrightarrow{\psi^2_r} V_{r+\epsilon_1} \xrightarrow{\psi^1_{r+\epsilon_1}} U_{r+(\epsilon_1+\epsilon_2)}$$

satisfy

$$\begin{aligned}
\psi_{r+(\epsilon_1+\epsilon_2)} \circ \phi_r &= \left(\psi^1_{r+(\epsilon_1+\epsilon_2)+\epsilon_1} \circ \psi^2_{r+(\epsilon_1+\epsilon_2)}\right) \circ \left(\phi^2_{r+\epsilon_1} \circ \phi^1_r\right) \\
&= \psi^1_{r+(\epsilon_1+\epsilon_2)+\epsilon_1} \circ \left(\psi^2_{r+(\epsilon_1+\epsilon_2)} \circ \phi^2_{r+\epsilon_1}\right) \circ \phi^1_r \\
&= \psi^1_{r+(\epsilon_1+\epsilon_2)+\epsilon_1} \circ L^{\mathsf{V}}_{r+\epsilon_1, r+\epsilon_1+\epsilon_2} \circ \phi^1_r \\
&= \psi^1_{r+(\epsilon_1+\epsilon_2)} \circ \phi^1_{r+\epsilon_2} \circ L^{\mathsf{U}}_{r, r+\epsilon_2} \\
&= L^{\mathsf{U}}_{r+\epsilon_2, r+(\epsilon_1+\epsilon_2)} \circ L^{\mathsf{U}}_{r, r+\epsilon_2} \\
&= L^{\mathsf{U}}_{r, r+(\epsilon_1+\epsilon_2)}.
\end{aligned}$$

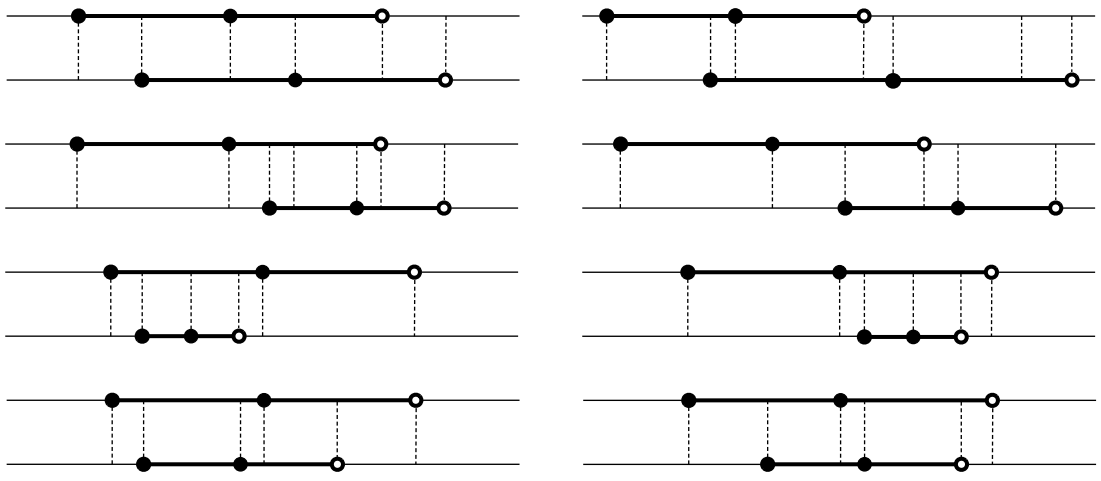
Similarly,

$$\phi_{r+(\epsilon_1+\epsilon_2)} \circ \psi_r = L^{\mathsf{W}}_{r, r+(\epsilon_1+\epsilon_2)}.$$

Since the lefthand side of (5.1) is an infimum over all interleavings, the inequality is satisfied. □

5.3.2 Interleaving Distance Between Interval Modules

In this subsection, we explicitly compute interleaving distance between the simplest types of persistence vector spaces: interval modules. It is useful to think of a interval module $I[a, b)$ as a copy of the real line with the points in the interval $[a, b)$ emphasized. It turns out that interleaving distance between interval modules $I[a, b)$ and $I[c, d)$ depends on relative ordering of endpoints and midpoints of the corresponding intervals. Without loss of generality, we may assume that $a \leqslant c$. Then there are 8 cases to consider:



First Example

We first examine the case in the upper left corner of the figure. That is, we consider interval modules $\mathcal{V} = I[a, b)$ and $\mathcal{W} = I[c, d)$ with midpoints $p = \frac{a+b}{2}$ and $q = \frac{c+d}{2}$ satisfying $a \leqslant c \leqslant p \leqslant q \leqslant b \leqslant d$.
Claim: In this case,

$$d_i(I[a, b), I[c, d)) = \max\{c - a, d - b\}.$$

To prove the claim, assume without loss of generality (the other case follows similarly) that $c-a$ is the larger of these values, and we set $\epsilon = c-a$. First note that there can be no ϵ'-interleaving for any $\epsilon' < \epsilon$. Indeed, suppose that there exists an ϵ'-interleaving (ϕ, ψ). Then $\phi_a : I[a, b)_a \to I[c, d)_{a+\epsilon'}$ would necessarily be the zero map, so that $\psi_{a+\epsilon'} \circ \phi_a$ is the zero map. On the other hand, $L^V_{a,a+2\epsilon'}$ is not the zero map, as the smallest value of r such that $L^V_{a,a+r}$ is the zero map is $b-a$. But $\epsilon' < \epsilon = c-a < \frac{b-a}{2}$ implies that $2\epsilon < b-a$.

We then wish to show that an ϵ-interleaving does exist. Take $\phi : \mathcal{V} \to \mathcal{W}$ and $\psi : \mathcal{W} \to \mathcal{V}$ to be the obvious ϵ-homomorphisms (unique up to rescaling). Then each map $\psi_{r+\epsilon} \circ \phi_r$ is clearly equal to $L^V_{r,r+2\epsilon}$. The issue which might arise is that the reverse map $\phi_{r+\epsilon} \circ \psi_r$ factors through zero—that is ψ_r is the zero map, but $L^W_{r,r+2\epsilon}$ is not. This is not the case: for this to happen, we would need ϕ_b to map to W_r with $r < d$. However, ϕ_b maps to $W_{b+\epsilon}$, and $b + \epsilon = b + c - a$. We claim that $b + c - a \geqslant d$. Indeed, this is equivalent to $c - a \geqslant d - b$, which was our initial assumption.

Second Example

Next consider the case pictured in the top right of the figure; that is, the case $a \leqslant c \leqslant p \leqslant b \leqslant q \leqslant d$, with the notation as above.
Claim: $d_i(\mathcal{V}, \mathcal{W}) = \frac{d-c}{2} = q - c$.

First note that there is no ϵ-interleaving with $\epsilon < q - c$. Indeed, for such a proposed interleaving, there would be $r \geqslant c$ such that ψ_r maps to $V_{r'}$ with $b < r' < q$. Then $\phi_{r'} \circ \psi_r$ would be the zero map, while the corresponding horizontal map would be nonzero. On the other hand, the obvious q-homomorphisms give an interleaving. In general, there will always be an ϵ-interleaving of interval modules when ϵ is taken to be the larger of the two interval radii.

Interleaving Distance for Interval Modules

We leave the reader to check the remaining cases. Once that is done, we have proved the following characterization of interleaving distance between interval modules. For points $(a, b), (c, d) \in \mathbb{R}^2$, recall that

$$d_\infty((a, b), (c, d)) = \max\{|c - a|, |d - b|\}.$$

We define

$$m(a, b) = \frac{|b - a|}{2}.$$

Proposition 5.3.3. *For interval modules $\mathcal{V} = I[a,b)$ and $\mathcal{W} = I[c,d)$, the interleaving distance is given by*

$$d_i(\mathcal{V},\mathcal{W}) = \min\{d_\infty((a,b),(c,d)), \max\{m(a,b),m(c,d)\}\}.$$

5.4 The Isometry Theorem

To each finitely-presented persistence vector space V, we associate a persistence diagram $\mathcal{D}(\mathsf{V})$. We can consider the assignment $\mathsf{V} \mapsto \mathcal{D}(\mathsf{V})$ as a map

$$\mathcal{D} : \mathrm{PVect} \to \mathrm{PDiag}.$$

Indeed, this map is well-defined since $[\mathsf{V}] = [\mathsf{W}]$ in PVect if and only if $\mathsf{V} \approx \mathsf{W}$, and this condition is equivalent to the condition that the diagrams $\mathcal{D}(\mathsf{V})$ and $\mathcal{D}(\mathsf{W})$ are equal. Each of the sets PVect and PDiag has been endowed with an extended metric (interleaving distance and bottleneck distance, respectively). The goal of this section is to show that these extended metrics are equivalent:

Theorem 5.4.1 (The Isometry Theorem)**.** *The map*

$$\mathcal{D} : \mathrm{PVect} \to \mathrm{PDiag}.$$

given by $[\mathsf{V}] \mapsto \mathcal{D}(\mathsf{V})$ is an isometry with respect to d_i and d_b.

An *isometry* of extended metric spaces (X, d_X) and (Y, d_Y) is a map $f : X \to Y$ satisfying

$$d_X(x,x') = d_Y(f(x),f(x'))$$

for all $x, x' \in X$. This is the same definition as in the case of (non-extended) metric spaces, with the only distinction being that the quantities are allowed to be ∞.

The proof of the theorem is done in two steps, which are the subjects of the following subsections.

5.4.1 Converse Stability

5.4.2 Algebraic Stability

6 Applications (Under Construction)

7 Appendix

7.1 Basic Background Material

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.

7.1.1 Basic Set Theory

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 7.1.1.

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 $\{b, c, a\}$.

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 7.1.2. 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 7.1.3. 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 7.1.4. Let $A = \{a, b\}$, $B = \{a, b, c, d\}$ and $C = \{b\}$. Then $A \subset B$ but $A \not\subset C$ (this should be read as "A is not a subset of C") because $a \in A$ but $a \notin C$, so the defining implication fails.

Sets of Sets

The elements of a set can themselves be sets!

Example 7.1.5. 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 7.1.6. 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 7.1.1. *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

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

□

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 7.1.7.

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 7.1.2. *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 $f(a) = f(f^{-1}(b)) = b$. □

7.1.2 Infinite Sets

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 7.1.8. 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 7.1.3. *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 7.1.4. *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 7.1.5. *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 7.1.3 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. □

An essentially straightforward corollary is left to the reader:

Corollary 7.1.6. *The set $\mathbb{R}$ of real numbers is uncountable.*

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 7.1.9. 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

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

7.1.3 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 7.1.10. 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 7.1.11. 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 $y = x - 2k$.

7.1.4 Supremum and Infimum

Let S denote some subset of $\mathbb{R}$. A real number $M \in \mathbb{R}$ is called an *upper bound* of S if for each $s \in S$, $s \leqslant M$. The number M is called a *least upper bound* of S if it has the property that whenever M' is an upper bound on S, it must be that $M' \leqslant M$.

Example 7.1.12. The set $[0,1]$ is upper bounded by any $M \in [1,\infty)$. Its least upper bound is 1. The set $[0,1)$ is upper bounded by any $M \in [1,\infty)$ and its least upper bound is also 1. The set $\{1 - 1/n \mid n \in \mathbb{Z}_{>0}\}$ has least upper bound 1. The sets $[0,\infty)$ and $\mathbb{Z}$ have no upper bounds.

We take the following statement as an axiom of the real numbers, meaning that we assume it to be true without proof:

Axiom 7.1.7 (The Completeness Axiom)**.** *If $S \subset \mathbb{R}$ is a nonempty set with at least one upper bound, then S has a least upper bound.*

Note that if a set S has a least upper bound, then it is unique. The *supremum* of a set $S \subset \mathbb{R}$ is defined to be

$$\sup(S) = \begin{cases} \text{the least upper bound of } S, & \text{provided } S \text{ has any upper bound} \\ \infty, & \text{otherwise.} \end{cases}$$

We similarly define a *lower bound* on S to be a real number m such that $m \leqslant a$ for all $a \in S$ and a *greatest lower bound* on S to be a lower bound m such that whenever m' is a lower bound on S, $m' \leqslant m$. The next proposition follows easily from our assumption of the Completeness Axiom.

Proposition 7.1.8. *If $S \subset \mathbb{R}$ is nonempty and has at least one lower bound, then it has a greatest lower bound.*

Proof. Suppose there exists a lower bound m' for S. Then $-m'$ is an upper bound for the set

$$-S = \{-s \mid s \in S\},$$

and the Completeness Axiom implies that $-S$ has a least upper bound m. Then $-m$ is a greatest lower bound for S. □

The *infimum* of a set $S \subset \mathbb{R}$ is defined to be

$$\inf(S) = \begin{cases} \text{the greatest lower bound of } S, & \text{provided } S \text{ has any lower bound} \\ -\infty, & \text{otherwise.} \end{cases}$$

7.2 Every Non-zero Vector Space Has a Basis

In this section we provide a proof of Theorem 1.2.1, 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 7.2.1. 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 7.2.1 (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 1.2.1.

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 $v \in \text{span}(B)$ or $v \notin B$, so it must be that $B \cup \{v\}$ is not linearly independent and we once again conclude that $v \in \text{span}(B)$.

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. □

7.3 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 7.1.5) to the interval $(0, 1)$ and to then apply Theorem 7.1.5 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.

Bibliography

[1] Awodey, Steve. “Category theory.” Oxford University Press, 2010.

[2] Bauer, Ulrich. ”Ripser: a lean C++ code for the computation of VietorisRips persistence barcodes.” Software available at https://github. com/Ripser/ripser (2017).

[3] Carlsson, Gunnar. “Topological pattern recognition for point cloud data.” Acta Numerica 23 (2014): 289-368.

[4] Hatcher, Allen. “Algebraic topology. 2002.” Cambridge UP, Cambridge 606, no. 9 (2002).

[5] Hoffman, K. and Kunze, R., Linear algebra. 1971. Englewood Cliffs, New Jersey.

[6] Johnson, Lee W., Ronald Dean Riess, and Jimmy Thomas Arnold. Introduction to linear algebra. Addison-Wesley Longman, 1993.

[7] 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.

[8] Munkres, James R. Topology. Prentice Hall, 2000.

Index