Lecture notes Math I (PhD500)

Mark Voorneveld

August 25, 2023

Much of the following is relatively newly written, so I would be very, very appreciative if you would let me know of any typos, unclarities, and so on!

I concentrated on writing the main text and have not yet added end-of-section exercises everywhere. In the final roughly fifty pages of this file you find a solutions manual with short solutions for most exercises.

Contents

1 Binary relations

A binary relation is a relation between pairs of objects. For instance:

- ☒ city a is the capital of country b,
- ☒ person c is taller than person d,
- ☒ student e is registered for course f,
- ☒ number g is greater than or equal to number h,
- ☒ set i is a subset of set j.

How can you describe such a relation mathematically? The relation 'is the capital of' is about pairs $(\texttt{city}, \texttt{country})$ where `city` is an element of a set X of cities, `country` is an element of a set Y of countries, and `city` indeed happens to be the capital of `country`. So you can just specify the set of pairs for which this relation is true: it would include pairs like $(\texttt{Paris}, \texttt{France})$, since `Paris` is the capital of `France`, but not $(\texttt{Oslo}, \texttt{Italy})$, since `Oslo` is not the capital of `Italy`. Formally:

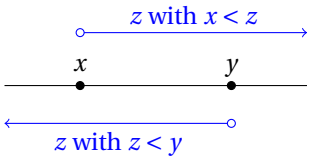
> **Definition 1.1** A ***binary relation*** on two sets X and Y is a subset of the Cartesian product $X \times Y$, that is, a set of ordered pairs (x, y) with $x \in X$ and $y \in Y$. If the pairs come from the same set ($X = Y$), we call it a ***binary relation on*** X.

If R is a binary relation on X and Y, i.e., if $R \subseteq X \times Y$, we often write xRy instead of $(x, y) \in R$ or use some other symbol than R to indicate that x and y satisfy the relation. For example, symbol $\leq$ is used for the binary relation 'is less than or equal to' on the real numbers and symbol $\subseteq$ is used for the binary relation 'is a subset of' on a collection of sets.

The next definition formulates some properties that a binary relation on a set may or may not possess. To help your intuition along, practise with them by letting binary relation R be one of the usual binary relations $\leq$, $<$, or $=$ on the set $X = \mathbb{R}$ of real numbers and ask yourself whether they satisfy each of these properties. The answer is in Figure 1.

> **Definition 1.2 (Properties for a binary relation R on a set X)**
>
> - ☒ ***Reflexivity:*** for all $x \in X$, xRx.
> - ☒ ***Irreflexivity:*** for all $x \in X$, not xRx.
> - ☒ ***Symmetry:*** for all $x, y \in X$, if xRy, then yRx.
> - ☒ ***Asymmetry:*** for all $x, y \in X$, if xRy, then not yRx.
> - ☒ ***Antisymmetry:*** for all $x, y \in X$, if xRy and yRx, then $x = y$.
> - ☒ ***Completeness:*** for all $x, y \in X$, xRy or yRx.
> - ☒ ***Transitivity:*** for all $x, y, z \in X$, if xRy and yRz, then xRz.
> - ☒ ***Negative transitivity:*** for all $x, y, z \in X$, if xRy, then xRz or zRy.
>
> Also corresponding adjective forms (reflexive, symmetric, etc.) are used.

Most properties have a straightforward interpretation. Completeness says that each pair of alternatives is comparable (xRy or yRx). Mathematicians tend to use an 'inclusive or' in mathematical formal statements, so it allows that both xRy and yRx are true. Negative transitivity is arguably the most opaque property. The picture illustrates it for the binary relation $<$ on the real number line: if $x < y$, then wherever you put a third number z, it will be to the right of x (i.e., $x < z$) or to the left of y (i.e., $z < y$).



Let's look at four important types of binary relations along with some canonical examples.

	$\leq$	$<$	$=$
reflexive	+	−	+
irreflexive	−	+	−
symmetric	−	−	+
asymmetric	−	+	−
antisymmetric	+	+	+
complete	+	−	−
transitive	+	+	+
negatively transitive	+	+	−

Figure 1: Binary relations $\leq$, $<$, and $=$ on $\mathbb{R}$ and properties they do (+) or do not (−) satisfy.

Definition 1.3 A binary relation on a set X is

- ☒ a ***partial order*** if it is reflexive, transitive, and antisymmetric;
- ☒ a ***linear order*** if it is complete, transitive, and antisymmetric;
- ☒ a ***weak order*** if it is complete and transitive;
- ☒ an ***equivalence relation*** if it is reflexive, transitive, and symmetric.

A partial order R is called 'partial' because not all elements need to be comparable: there may be elements x and y for which neither xRy nor yRx is true. A linear order is a partial order without incomparable elements, i.e., a partial order that is also complete. A complete binary relation is automatically reflexive (take $x = y$ in the definition of completeness). A set with a partial order is called a ***partially ordered set*** (or *poset*); a set with a linear order is called a ***linearly ordered set*** (or a *chain* or *totally ordered set*).

Example 1.1 It follows from Figure 1 that the binary $\leq$ on $\mathbb{R}$ is a partial order, a linear order, and a weak order. And that $=$ is an equivalence relation. ◁

Example 1.2 (Partial orders)

- ☒ A set $X \subseteq \mathbb{R}^n$ of n-dimensional real vectors is partially ordered by $\leq$ defined, for all $x, y \in \mathbb{R}^n$, as

$$x \leq y \quad \Longleftrightarrow \quad x_i \leq y_i \quad \text{for all coordinates } i = 1, \ldots, n.$$

  In $\mathbb{R}^2$ we have $(1,3) \leq (2,4)$, but vectors like $(1,0)$ and $(0,2)$ are incomparable.
- ☒ Let X be a set. Its power set $P(X) = \{Y : Y \subseteq X\}$ consists of all subsets of X. Set inclusion $\subseteq$ is a partial order on $P(X)$. But as soon as X contains two distinct elements, say a and b, it is not a linear order: $\{a\}$ and $\{b\}$ are incomparable as neither is a subset of the other.
- ☒ The set of real-valued functions on some domain X is partially ordered by $\leq$ defined, for all $f, g : X \to \mathbb{R}$, as

$$f \leq g \quad \Longleftrightarrow \quad f(x) \leq g(x) \quad \text{for all } x \in X.$$

  Geometrically, $f \leq g$ means that the graph of f lies below the graph of g.
- ☒ The cumulative distribution function (cdf) of a real-valued random variable is the function $F : \mathbb{R} \to [0,1]$ which assigns to each x the probability $F(x)$ that the random variable achieves a value less than or equal to x. Let F and G be cdf's of two random variables. As in our previous example, write $F \leq G$ if $F(x) \leq G(x)$ for all x. This partial order on the collection of all cdf's has a special name: if $F \leq G$, the second random variable is more likely to give low values and we say that the first random variable 'first-order stochastically dominates' the second. ◁

Example 1.3 (Preferences) In economic theory it is common to model the likes and dislikes of an economic agent by means of a weak order, often denoted $\succsim$, over a set X of alternatives. If a and b are two alternatives, then $a \succsim b$ means that according to the agent, a is at least as good as/weakly preferred to/weakly better than b.

Such a preference relation is usually not a partial order, for two reasons. Weak orders are complete — the agent can rank each pair of alternatives — but partial orders need not be. And if the agent likes two distinct alternatives a and b equally much, i.e., if both $a \succsim b$ and $b \succsim a$ are true, but $a \neq b$, then this preference relation fails to be antisymmetric. ◁

Different orders give rise to new ones. For instance, if a weak order $\succsim$ models an agent's preferences over a set X of alternatives, we can define strict preference

$$x \succ y \text{ if } x \succsim y \text{ but not } y \succsim x \qquad (\text{'}x \text{ is better than/strictly preferred to } y\text{'}),$$

and indifference

$$x \sim y \text{ if } x \succsim y \text{ and } y \succsim x \qquad (\text{'}x \text{ and } y \text{ are equally good/equivalent'}).$$

These new binary relations inherit nice properties of the ones we started with. The proofs tend to require little else than keeping careful track of the definitions and are a good way to practice with the material in this section.

Theorem 1.1

If $\succsim$ is a weak order on a set X, then

(a) indifference $\sim$ is an equivalence relation;

(b) strict preference $\succ$ is asymmetric and negatively transitive.

Proof: You should establish (a) yourself. I will do (b):

To see that $\succ$ is asymmetric, let $x, y \in X$ satisfy $x \succ y$. By definition of $\succ$, $x \succsim y$ but not $y \succsim x$. The latter, again by definition of $\succ$, implies that not $y \succ x$. Conclude that $\succ$ is asymmetric.

For negative transitivity of $\succ$, let $x, y, z \in X$ satisfy $x \succ y$. We must show that $x \succ z$ or $z \succ y$. Suppose, to the contrary, that both are false: not $x \succ z$ and not $z \succ y$. Since $\succsim$ is complete, it follows that $z \succsim x$ and $y \succsim z$. And with transitivity of $\succsim$: $y \succsim x$. This contradicts our starting point, that $x \succ y$. Conclude that $\succ$ is negatively transitive. □

Other properties hold as well. For instance, you might check that $\succ$ is transitive.

Exercises section 1

1.1 Using the partial orders from Example 1.2:

(a) Find vectors x and y in $\mathbb{R}^3$ with $x \leq y$ but not $y \leq x$. And with both $x \leq y$ and $y \leq x$. Finally, with neither $x \leq y$ nor $y \leq x$.

(b) Draw the graphs of the functions f and g from and to $\mathbb{R}$ with $f(x) = x$ and $g(x) = x^2$. Is $f \leq g$? What about $g \leq f$? Find a function h with $g \leq h$.

1.2 Let X be a nonempty set of people. Look at three binary relations: 'has the same birthday as', 'is born at most 10 days before', and 'is at least as tall as'. Are these relations necessarily partial orders, linear orders, weak orders, equivalence relations?

1.3 Negative transitivity got its name because R being negatively transitive is equivalent with the binary relation 'not R' being transitive:

Show that a binary relation R on a set X is negatively transitive if and only if

$$\text{for all } x, y, z \in X, \text{ if (not } xRy) \text{ and (not } yRz), \text{ then (not } xRz).$$

2 Finite, countable, and uncountable sets

In this section we classify sets according to how many elements they have. Take a quick look at Appendices A.3 and A.4 for a reminder of common notation and terminology for functions. For convenience, definitions that are used frequently in this section are repeated here:

Definition 2.1 Let A and B be sets. A function $f : A \to B$ is:

- ☒ ***injective*** if distinct arguments give distinct function values: for all $a, a' \in A$, if $a \neq a'$, then $f(a) \neq f(a')$;
- ☒ ***surjective*** if each element of B is the image of some element of A: for each $b \in B$ there is an $a \in A$ with $f(a) = b$.
- ☒ a ***bijection*** if it is both injective and surjective.

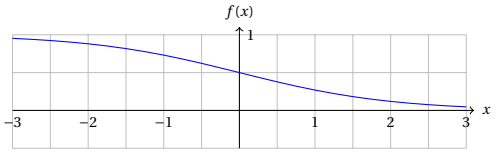
To explain to two children who have not yet learned to count that you are giving them equally many pieces of candy, you can line up these sweets in two rows in such a way that each piece of candy in the first row matches with a piece of candy in the second row, and vice versa: there is a bijection between the pieces of candy in the two rows. Hence we can express that two sets have the same number of elements (in mathematical terms, the same cardinality) without bringing numbers into play.

Definition 2.2 Set A ***has the same cardinality as*** set B if there is a bijection $f : A \to B$.

The sentence "A has the same cardinality as B" means that there is a bijection $f : A \to B$. But "B has the same cardinality as A" means that there is a bijection in the opposite direction, say a bijection $g : B \to A$. Of course these two things are equivalent: if you have a bijection from A to B, its inverse is a bijection from B to A, and vice versa. So there is no risk of confusion in saying that A and B have the same cardinality: there is a bijection in either direction.

What takes some getting used to is that if you start with an infinite set and throw away some of its elements, you may end up with a set that has the same cardinality:

Example 2.1 $\mathbb{R}$ and its subset $(0,1)$ have the same cardinality: the strictly decreasing function $f : \mathbb{R} \to (0,1)$ with $f(x) = 1/(1+e^x)$, whose graph is to the right, is a bijection. ◁



Example 2.2 The table below gives bijections from $\mathbb{N} = \{1,2,3,\ldots\}$ to the set $E = \{2,4,6,\ldots\}$ of even positive integers, to the set $S = \{1^2, 2^2, 3^2, \ldots\}$ of squared positive integers, and to the set $\mathbb{Z} = \{\ldots,-2,-1,0,1,2,\ldots\}$ of all integers:

$\mathbb{N}$	1	2	3	4	5	6	7	$\cdots$
E	2	4	6	8	10	12	14	$\cdots$
S	1	4	9	16	25	36	49	$\cdots$
$\mathbb{Z}$	0	-1	1	-2	2	-3	3	$\cdots$

So all these sets have the same cardinality as $\mathbb{N}$. Sometimes the bijections are obvious:

$$E = \{2n : n \in \mathbb{N}\} \qquad \text{and} \qquad S = \{n^2 : n \in \mathbb{N}\}$$

suggest bijections $n \mapsto 2n$ and $n \mapsto n^2$ from $\mathbb{N}$ to E and S, respectively. The bijection $f : \mathbb{N} \to \mathbb{Z}$ is given by $f(n) = (n-1)/2$ if n is odd and $f(n) = -n/2$ if n is even. ◁

Example 2.3 $\mathbb{N} \times \mathbb{N}$ has the same cardinality as $\mathbb{N}$. One bijection $f : \mathbb{N} \times \mathbb{N} \to \mathbb{N}$ is shown below.

$f(m,n)$	$n=1$	$n=2$	$n=3$	$n=4$	…
$m=1$	1	2	4	7	…
$m=2$	3	5	8	…	…
$m=3$	6	9	…	…	…
$m=4$	10	…	…	…	…
$\vdots$					

If you draw a line from function value 1 to 2 to 3 to 4 and so on, you see that it zigzags through all possible pairs of positive integers (m,n). ◁

We can now make a number of intuitive properties mathematically precise:

Definition 2.3 A set A is:

- ☒ ***finite*** if it is empty or has the same cardinality as $\{1,\ldots,n\}$ for some positive integer n;
- ☒ ***infinite*** if it is not finite;
- ☒ ***countably infinite*** if it has the same cardinality as $\mathbb{N}$;
- ☒ ***countable*** if it is finite or countably infinite;
- ☒ ***uncountable*** if it is not countable.
- ☒ If a set A is nonempty and finite, there is a bijection $f:\{1,\ldots,n\}\to A$ for some positive integer n. So we can speak of its first element $f(1)\in A$ and denote it by a_1, its second element $f(2)\in A$, denoted a_2, and so on: we can ***enumerate*** its elements

$$A=\{a_1,\ldots,a_n\}.$$

Likewise, for a countably infinite set A we can use a bijection $f:\mathbb{N}\to A$ to define, for each $n\in\mathbb{N}$, its n-th element $a_n=f(n)$. This gives an enumeration of the elements of A:

$$A=\{a_1,a_2,\ldots\}=\{a_n : n\in\mathbb{N}\}$$

Some authors reserve the word 'countable' for sets we call 'countably infinite'. Enumerations suggest a memory aid for what it means to be countable. In sports, members of a team are often identified by a number on their clothes. Imagine that for each of the numbers 1, 2, 3, … we print a unique T-shirt. For countable sets, there are enough T-shirts to dress all members (a surjective function from $\mathbb{N}$ to A); or the other way around, distinct members can get distinct shirts (an injective function from A to $\mathbb{N}$). To summarize:

Theorem 2.1

For a nonempty set A, the following are equivalent:

(a) A is countable;

(b) there is a surjective function $g:\mathbb{N}\to A$;

(c) there is an injective function $h:A\to\mathbb{N}$.

The next theorem states the most commonly used properties of countable sets.

Theorem 2.2 (Properties of countable sets)

(a) Each subset of a finite set is finite.

(b) Each subset of a countable set is countable.

(c) Each infinite set has a countably infinite subset.

(d) The union of countably many countable sets is countable.

(e) The Cartesian product of finitely many countable sets is countable.

Proof: **(a)** should be clear. For **(b)**, let B be a subset of a countable set A. If B is empty, it is countable. What if it is nonempty? Use Theorem 2.1: there is an injective function $h : A \to \mathbb{N}$. Its restriction to the smaller domain B remains injective, so B is countable as well.
(c) If set A is infinite, it is nonempty and we can pick an element a_1 in A. Now proceed recursively: if $a_1, \ldots, a_n$ have been chosen, then $A \setminus \{a_1, \ldots, a_n\}$ is nonempty, so we can again select an element $a_{n+1} \in A \setminus \{a_1, \ldots, a_n\}$. The set $\{a_k : k \in \mathbb{N}\}$ is a countably infinite subset.
(d) Look at countably many sets $A_1, A_2, A_3, \ldots$. If there are finitely many, this ends after, say, k steps. If there are infinitely many, we have one such set A_i for each $i \in \mathbb{N}$. Since the claim is about the elements in their union, we may assume without loss of generality that all these sets are nonempty. Enumerate the elements of each set A_i:

$$A_i = \{a_{i1}, a_{i2}, a_{i3}, \ldots\},$$

i.e., a_{i1} is its first element, a_{i2} its second, and so on. Each element a of the union $\cup_i A_i$ lies in some set A_m. If it lies in several, choose the one with the smallest index m. Say a is the n-th element of A_m: $a = a_{mn}$. In this way we associate with each a in the union a unique pair (m, n) of positive integers: we have a bijection between $\cup_i A_i$ and a subset of $\mathbb{N} \times \mathbb{N}$. Since $\mathbb{N} \times \mathbb{N}$ is countable (Example 2.3), so are this subset and $\cup_i A_i$.
(e) If any of the sets is empty, so is their product. So assume they're nonempty.

Let's start with the product of two nonempty countable sets A_1 and A_2. For each fixed $a_2 \in A_2$, the set $A_1 \times \{a_2\}$ has the same cardinality as A_1: since a_2 was fixed, the function f with $f(a_1, a_2) = a_1$ for each $a_1 \in A_1$ is a bijection between them. This makes

$$A_1 \times A_2 = \cup_{a_2 \in A_2} A_1 \times \{a_2\}$$

the union of countably many countable sets, hence countable.

The rest is induction: assume the product of $n \in \mathbb{N}$ nonempty countable sets is countable and now consider $n+1$ such sets, A_1 to A_{n+1}. The product $A_1 \times \cdots \times A_n$ of the first n sets is countable and, as above, for each $a_{n+1} \in A_{n+1}$ the set $A_1 \times \cdots \times A_n \times \{a_{n+1}\}$ has the same cardinality, making it countable as well. Hence

$$A_1 \times \cdots \times A_n \times A_{n+1} = \cup_{a_{n+1} \in A_{n+1}} A_1 \times \cdots \times A_n \times \{a_{n+1}\}$$

is the union of countably many countable sets and therefore countable itself. □

Example 2.4 ($\mathbb{Q}$ and $\mathbb{Q}^n$ are countable) The set $\mathbb{Q}$ of rational numbers is countable: Each rational number is a fraction of two integers and the integers form a countable set (Example 2.2). So the fractions with a fixed denominator form a countable set. Taking the union over the countably many denominators, it follows that $\mathbb{Q}$ is a union of countably many countable sets.

And since the product of finitely many countable sets is countable as well, it follows that for each $n \in \mathbb{N}$ the set $\mathbb{Q}^n$ of n-dimensional vectors of rational numbers is countable. ◁

So sets like $\{a, b, c\}$ are finite and sets like $\mathbb{N}$, $\mathbb{Z}$, and $\mathbb{Q}$ are countably infinite. But are there uncountable sets? Oh yes! Pay close attention to the next proof. It introduces a common proof technique in set theory, called — for reasons that will shortly be clear — a ***diagonal argument***.

Theorem 2.3

The set of real numbers is uncountable.

Proof: Suppose to the contrary that $\mathbb{R}$ is countable. Then we can enumerate all its elements $a_1, a_2, a_3, \ldots$. But this list cannot possibly contain each real number. For instance, define the real number

$$b = 0.b_1 b_2 b_3 \cdots$$

where

- ⊠ b_1 is a digit in $\{0, \ldots, 9\}$ distinct from whatever digit is in the *first* place after the decimal point in any decimal expansion of a_1;
- ⊠ b_2 is a digit in $\{0, \ldots, 9\}$ distinct from whatever digit is in the *second* place after the decimal point in any decimal expansion of a_2;
- ⊠ and so on.

I wrote 'any' instead of 'the' decimal expansion, since some real numbers have two different decimal expansions (for instance, $1/2 = 0.50000\cdots = 0.49999\cdots$. See Appendix A.5.). That makes the construction feasible: for b_i there are ten digits to choose from and at most two are excluded by the decimal expansion(s) of a_i.

This number b is not on the list: it differs from a_1 in its first decimal place, from a_2 in its second decimal place, and so forth. □

Although both are infinite, this theorem tells that the set of real numbers is 'more infinite' than the set of positive integers. We can certainly find an injective function (like $f(n) = n$) from $\mathbb{N}$ to $\mathbb{R}$, but not a surjective one, since the diagonal argument establishes that no countable list of real numbers contains all of them. Formally, $\mathbb{R}$ has a larger cardinality than $\mathbb{N}$:

Definition 2.4 If there is an injective function from set A to set B, but not a surjective one, we say that A ***has a smaller cardinality than*** B or that B ***has a larger cardinality than*** A.

Most of the material in this section was developed by Georg Cantor in the late 19th century. He is considered to be the creator of set theory. But his work was perceived as counterintuitive and shocking. Some contemporary mathematicians were less than charmed with it. Leopold Kronecker, a teacher of Cantor, apparently[1] went so far as to call him a "corrupter of youth". This probably overestimates how interested the average young person is in the foundations of mathematics.

One of Cantor's remarkable insights was that there are infinitely many different kinds of infinity: the infinite set $\mathbb{R}$ has a larger cardinality than $\mathbb{N}$, but there is a third set with a larger cardinality than $\mathbb{R}$, a fourth set with an even larger cardinality, and so on. This follows from repeated application of the following theorem, which says that there are fewer elements in a set than in its power set. Taking power sets of power sets, we have a recipe for generating sets with ever larger cardinality.

The theorem is straightforward for finite sets: $\{a, b\}$ has two elements, its power set $\{\emptyset, \{a\}, \{b\}, \{a, b\}\}$ has four. Generally, if a set A has $n \in \mathbb{N}$ elements, its power set consists of $2^n > n$ subsets, because picking a subset of A means deciding for each of its n elements whether it lies in the subset or not. So there are n choices, each with two options, making a total of 2^n subsets. The result is more surprising for infinite sets.

[1] J. W. Dauben (1983): "Georg Cantor and the Origins of Transfinite Set Theory", *Scientific American*, 248: 122–131.

Theorem 2.4

Each set has a smaller cardinality than its power set.

Proof: Let X be a set. The function that maps each element $x \in X$ to the singleton set $\{x\}$ is an injective function from X to its power set. But there is no surjective function. For any function f from X to its power set $P(X)$, I will prove that the set

$$Y = \{x \in X : x \notin f(x)\}$$

is not in its range. Suppose there were a $y \in X$ with $f(y) = Y$. Now ask whether y belongs to Y:

$$y \in Y \underbrace{\iff}_{\text{by def. of } Y} y \notin f(y) \underbrace{\iff}_{\text{since } f(y)=Y} y \notin Y.$$

This is a contradiction. So no function from X to $P(X)$ is surjective. □

We conclude this section by sketching a fascinating result of Alan Turing using diagonalization to show that for some problems in computer science you cannot write a program to solve them:

Example 2.5 (The halting problem) Anyone who ever wrote a crappy computer program realizes that non-terminating programs are a nuisance. Couldn't you just program a stopping-detector that terminates, for each possible program and each possible input, with the word '`true`' if that program ends in finite time on the given input, and with the word '`false`' otherwise? This is Alan Turing's *halting problem* and he proved — using a diagonalization argument — that the answer is 'no'.

The programs you can write in a programming language consist of finite strings of finitely many symbols in that language and consequently constitute a countable set. Enumerating them in an arbitrary way, we refer to them as program 1, program 2, and so on. Likewise, input 1, input 2, ... is some enumeration of the possible inputs.

Suppose there were a program $\texttt{halt} : \mathbb{N}\times\mathbb{N} \to \{\texttt{true},\texttt{false}\}$ that terminates for each program-input pair $(i, j) \in \mathbb{N} \times \mathbb{N}$ with output $\texttt{halt}(i, j) = \texttt{true}$ if program i ends in finite time on input j and with output $\texttt{halt}(i, j) = \texttt{false}$ otherwise. Now write a new program `paradox` (see the diagram below) that does, for each input i, the opposite of what program i does:

- ☒ if $\texttt{halt}(i, i) = \texttt{true}$, then $\texttt{paradox}(i)$ does not terminate but goes into an infinite loop;
- ☒ if $\texttt{halt}(i, i) = \texttt{false}$, then $\texttt{paradox}(i)$ terminates in finite time.

halt(program,input)	input 1	input 2	input 3	...
program 1	true	false	true	...
program 2	false	false	false	...
program 3	false	false	true	...
⋮	⋮	⋮	⋮	⋮
paradox(input)	loop	terminate	loop	...

So for each input i, program `paradox` terminates if and only if program i doesn't. And there lies the problem: since we enumerated all programs, program `paradox` is somewhere on our list. Say it is program k. But remember, `paradox` does the opposite of program k (= `paradox`) on input k, i.e., the opposite *of itself*, a clear contradiction. Hence the assumption that there is a program solving the halting problem is false! ◁

Exercises section 2

2.1 Is the function $x \mapsto x^2$ injective, surjective, bijective:

(a) From $\mathbb{R}$ to $\mathbb{R}$?

(b) From $\mathbb{R}$ to $\mathbb{R}_+ = [0, \infty)$?

(c) From $\mathbb{R}_+$ to $\mathbb{R}$?

(d) From $\mathbb{R}_+$ to $\mathbb{R}_+$?

(e) From $\mathbb{R}_- = (-\infty, 0]$ to $\mathbb{R}_+$?

If it is bijective, what is its inverse?

2.2 Write $A \approx B$ if set A has the same cardinality as set B. Show that this binary relation 'has the same cardinality as' ($\approx$) on an arbitrary collection of sets is an equivalence relation.

3 Vector spaces

In $\mathbb{R}^2$ — the set of ordered pairs (x_1, x_2) of real numbers x_1 and x_2 — you know how to add vectors and how to multiply them with a given number (or 'scalar'). For instance:

$$(1,-4)+(-7,12)=(1+(-7),-4+12)=(-6,8)$$

and

$$5(2,-1)=(5\cdot 2,5\cdot(-1))=(10,-5).$$

But this is not the only set where you know how to add elements or multiply them with a scalar:

Example 3.1 In the set of 2×3 matrices — those with two rows and three columns — we have

$$\begin{bmatrix}0&1&2\\3&4&5\end{bmatrix}+\begin{bmatrix}6&7&8\\9&10&11\end{bmatrix}=\begin{bmatrix}0+6&1+7&2+8\\3+9&4+10&5+11\end{bmatrix}=\begin{bmatrix}6&8&10\\12&14&16\end{bmatrix}$$

and

$$-2\begin{bmatrix}-2&3&0\\5&-1&3\end{bmatrix}=\begin{bmatrix}(-2)\cdot(-2)&(-2)\cdot 3&(-2)\cdot 0\\(-2)\cdot 5&(-2)\cdot(-1)&(-2)\cdot 3\end{bmatrix}=\begin{bmatrix}4&-6&0\\-10&2&-6\end{bmatrix}$$ ◁

Example 3.2 The sum of the functions $f:\mathbb{R}\to\mathbb{R}$ and $g:\mathbb{R}\to\mathbb{R}$ defined, for each $x\in\mathbb{R}$, by

$$f(x)=-3x^2+7x \quad\text{and}\quad g(x)=e^x-5$$

is the function $f+g$ assigning to $x\in\mathbb{R}$ the value

$$f(x)+g(x)=-3x^2+7x+e^x-5,$$

and three times the function f is the function $3f$ assigning to $x\in\mathbb{R}$ the value

$$3f(x)=-9x^2+21x.$$ ◁

Apparently, such diverse sets as $\mathbb{R}^2$, the set of 2×3 matrices, and the set of functions from and to the real numbers have something in common: there are well-behaved definitions of addition and scalar multiplication. In mathematical terms, they are vector spaces. The definition of a vector space is quite a lot to take in all at once. Don't worry, you can always look it up. The thing to remember is: *A vector space is a set V whose elements ('vectors') you can add together and multiply with numbers ('scalars'). Addition and scalar multiplication once again produce vectors in V and satisfy standard arithmetic properties.*

Definition 3.1 A (real) ***vector space*** is a set V on which two operations, addition and scalar multiplication, are defined such that

- ☒ V is closed under addition: for each pair of elements $x, y\in V$ there is a unique element $x+y$, the sum of x and y, in V,
- ☒ V is closed under scalar multiplication: for each $x\in V$ and each $\alpha\in\mathbb{R}$, there is a unique element αx, the (scalar) product of α and x, in V.

Moreover, addition and scalar multiplication are well-behaved in the following sense:

AXIOMS FOR ADDITION:

(V1) Commutativity: for all $x, y\in V: x+y=y+x$.

(V2) Associativity: for all $x, y, z\in V: (x+y)+z=x+(y+z)$.

(V3) Existence of a zero element: there is a $\mathbf{0} \in V$ such that for all $x \in V$: $x + \mathbf{0} = x$.

(V4) Existence of an additive inverse: for each $x \in V$ there is a $y \in V$ with $x + y = \mathbf{0}$.

AXIOMS FOR SCALAR MULTIPLICATION:

(V5) for all $x \in V$ and all $\alpha, \beta \in \mathbb{R}$: $(\alpha\beta)x = \alpha(\beta x)$.

(V6) for all $x \in V$: $1x = x$.

AXIOMS FOR DISTRIBUTIVITY:

(V7) for all $x, y \in V$ and all $\alpha \in \mathbb{R}$: $\alpha(x + y) = \alpha x + \alpha y$.

(V8) for all $x \in V$ and all $\alpha, \beta \in \mathbb{R}$: $(\alpha + \beta)x = \alpha x + \beta x$.

The numbers $\alpha, \beta \in \mathbb{R}$ are called ***scalars***, the elements of V are called ***vectors***.

Remark 3.1 Definition 3.1 introduces a *real* vector space, since scalar multiplication is defined for scalars in the set of real numbers $\mathbb{R}$. If the set $\mathbb{R}$ of scalars in this definition is replaced with any other field F of scalars — like the set $\mathbb{Q}$ of rational numbers or the set $\mathbb{C}$ of complex numbers — then we obtain the definition of a ***vector space*** V ***over a field*** F. Appendix A.1 contains the definition and examples of fields; Appendix B gives an informal introduction to complex numbers. We look at real vector spaces, unless explicitly stated otherwise. For instance, later on we discuss eigenvalues and eigenvectors and will need to consider complex vector spaces ($F = \mathbb{C}$). ◁

To check whether a set V with given definitions of addition and scalar multiplication is a vector space, there are ten things to prove: properties (V1) to (V8), but also that the set is closed under addition and scalar multiplication. Fortunately, the arithmetic rules (V1) to (V8) often follow easily from similar properties of the real numbers. And later, in Theorem 3.2 on page 18, you will see that you often won't need to verify them at all; a most convenient short-cut!

The following examples, partly generalizing earlier ones, introduce common vector spaces.

Example 3.3 For arbitrary $n \in \mathbb{N}$, the set $\mathbb{R}^n$ consists of n-tuples $x = (x_1, \ldots, x_n)$ of n real numbers. Two vectors $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$ in $\mathbb{R}^n$ are defined to be equal if $x_i = y_i$ for all coordinates $i = 1, \ldots, n$. $\mathbb{R}^n$ is a vector space under the operations of coordinatewise addition and scalar multiplication. Formally, for all $x = (x_1, \ldots, x_n)$, $y = (y_1, \ldots, y_n) \in \mathbb{R}^n$, and all scalars $\alpha \in \mathbb{R}$:

$$x + y = (x_1 + y_1, \ldots, x_n + y_n) \quad \text{and} \quad \alpha x = (\alpha x_1, \ldots, \alpha x_n).$$

Its zero element $\mathbf{0}$ is the vector $(0, \ldots, 0)$ with all n coordinates equal to zero. Vector space $\mathbb{R}^1$ consists of vectors of just one real number; we will simply write $\mathbb{R}$ instead of $\mathbb{R}^1$. ◁

Example 3.4 An $m \times n$ real matrix A is a rectangular array of the form

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

with m rows and n columns. The entry in row $i \in \{1, \ldots, m\}$ and column $j \in \{1, \ldots, n\}$ is denoted by A_{ij} or $a_{ij} \in \mathbb{R}$. Entries $a_{i1}, a_{i2}, \ldots, a_{in}$ form the i-th *row* of matrix A and entries $a_{1j}, a_{2j}, \ldots, a_{mj}$ form the j-th *column* of matrix A.

Two $m \times n$ matrices A and B are equal if all corresponding entries are equal: $a_{ij} = b_{ij}$ for all $i = 1, \ldots, m$ and $j = 1, \ldots, n$.

The set $\mathbb{R}^{m\times n}$ of $m\times n$ matrices with real entries is a vector space under entrywise addition and scalar multiplication: for all $A, B \in \mathbb{R}^{m\times n}$ and all $\alpha \in \mathbb{R}$,

$$(A+B)_{ij} = a_{ij} + b_{ij} \quad \text{and} \quad (\alpha A)_{ij} = \alpha a_{ij}.$$

Its zero element $\mathbf{0}$ is the $m\times n$ matrix with all entries equal to zero. ◁

Example 3.5 A ***sequence*** $(x_1, x_2, x_3, \ldots) = (x_n)_{n\in\mathbb{N}}$ in $\mathbb{R}$ assigns to each positive integer $n = 1, 2, 3, \ldots$ a real number $x_n \in \mathbb{R}$: it is a function from $\mathbb{N}$ to $\mathbb{R}$, although our notation is more common and convenient.

Two sequences $x = (x_1, x_2, x_3, \ldots)$ and $y = (y_1, y_2, y_3, \ldots)$ are equal if $x_n = y_n$ for all $n \in \mathbb{N}$. Denote the set of real sequences by $\mathbb{R}^{\mathbb{N}}$; it is a vector space under coordinatewise addition and scalar multiplication. Formally, for all $x = (x_1, x_2, x_3, \ldots), y = (y_1, y_2, y_3, \ldots) \in \mathbb{R}^{\mathbb{N}}$ and all $\alpha \in \mathbb{R}$:

$$x + y = (x_1, x_2, x_3, \ldots) + (y_1, y_2, y_3, \ldots) = (x_1 + y_1, x_2 + y_2, x_3 + y_3, \ldots)$$

and

$$\alpha x = \alpha(x_1, x_2, x_3, \ldots) = (\alpha x_1, \alpha x_2, \alpha x_3, \ldots).$$

Its zero element $\mathbf{0}$ is the sequence $(0, 0, 0, \ldots)$ with all coordinates equal to zero. ◁

Example 3.6 The set $C[a, b]$ of continuous functions $f : [a, b] \to \mathbb{R}$ is a vector space. Here, a and b are real numbers with $a \le b$. Addition and scalar multiplication are defined as in Example 3.2: for all $f, g \in C[a, b]$ and all $\alpha \in \mathbb{R}$,

$$(f + g)(x) = f(x) + g(x) \quad \text{and} \quad (\alpha f)(x) = \alpha(f(x)).$$

Its zero element $\mathbf{0}$ is the function from $[a, b]$ to $\mathbb{R}$ that is constant at zero. ◁

Example 3.7 A ***polynomial (function)*** with coefficients in $\mathbb{R}$ is a function of the form

$$p(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0,$$

where n is a nonnegative integer ($n \in \{0, 1, 2 \ldots\}$) and coefficients $a_n, a_{n-1}, \ldots, a_1, a_0$ are numbers in $\mathbb{R}$.

If $a_0 = a_1 = \cdots = a_n = 0$, then $p(x) = 0$ for all $x \in \mathbb{R}$. This p is called the ***zero polynomial***; we define its degree as -1. Otherwise, the degree $\deg(p)$ of polynomial p is the largest exponent n with $a_n \neq 0$: $x^3 + 17x - 2$ has degree three, $3x^4 + x^2$ has degree four, the constant polynomial $p(x) = 7$ has degree zero, and so on. Two polynomials

$$p(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0 \quad \text{and} \quad q(x) = b_m x^m + b_{m-1} x^{m-1} + \cdots + b_1 x + b_0,$$

are equal if they have the same degree and equal powers have equal coefficients ($a_i = b_i$ for all i).

The set $P(\mathbb{R})$ of real polynomials is a vector space with addition and scalar multiplication defined as follows: for polynomials

$$\begin{aligned} p(x) &= a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0, \\ q(x) &= b_n x^n + b_{n-1} x^{n-1} + \cdots + b_1 x + b_0, \end{aligned}$$

in $P(\mathbb{R})$ and scalar $\alpha \in \mathbb{R}$,

$$(p + q)(x) = (a_n + b_n)x^n + (a_{n-1} + b_{n-1})x^{n-1} + \cdots + (a_1 + b_1)x + (a_0 + b_0) \tag{1}$$

and

$$(\alpha p)(x) = (\alpha a_n)x^n + (\alpha a_{n-1})x^{n-1} + \cdots + (\alpha a_1)x + (\alpha a_0). \tag{2}$$

Its zero element $\mathbf{0}$ is the zero polynomial with $\mathbf{0}(x) = 0$ described above. ◁

In several examples we made new vector spaces from old ones by putting copies of them next to each other: $\mathbb{R}^2$ consists of two copies of $\mathbb{R}$, $\mathbb{R}^n$ consists of n copies of $\mathbb{R}$, and the set $\mathbb{R}^{\mathbb{N}}$ of real sequences consists of infinitely many copies of $\mathbb{R}$, one for each $n \in \mathbb{N}$. Also the set of real-valued functions on $[0,1]$ consists of infinitely many copies of $\mathbb{R}$, this time one for each $x \in [0,1]$. In all these cases, addition and scalar multiplication are defined coordinatewise. These are examples of product spaces:

Example 3.8 (Product spaces) If U and V are vector spaces, their ***product space*** $U \times V$ is the set of ordered pairs (u, v) with $u \in U$ and $v \in V$. Addition and scalar multiplication are defined coordinatewise. For all $(u, v), (u', v') \in U \times V$ and all scalars α:

$$(u, v) + (u', v') = (u + u', v + v') \qquad \text{and} \qquad \alpha(u, v) = (\alpha u, \alpha v).$$

The zero vector of $U \times V$ is $(\mathbf{0}_U, \mathbf{0}_V)$, where $\mathbf{0}_U$ and $\mathbf{0}_V$ are the zero vectors of U and V, respectively.

This definition easily generalizes to the product $V_1 \times \cdots \times V_n$ of any finite number of vector spaces. We can even do this for an arbitrary collection of vector spaces. Let I be any nonempty 'index' set. For each $i \in I$, let V_i be a vector space. The ***product space*** $\times_{i \in I} V_i$ is defined to be the set of all functions x on I such that $x(i) \in V_i$ for all $i \in I$. We often write x_i instead of $x(i)$ and denote x by $x = (x(i))_{i \in I}$ or $x = (x_i)_{i \in I}$. Addition and scalar multiplication are again defined coordinatewise: for all $x, y \in \times_{i \in I} V_i$ and all $\alpha \in \mathbb{R}$:

$$(x + y)(i) = x(i) + y(i) \qquad \text{and} \qquad (\alpha x)(i) = \alpha x(i) \qquad (i \in I).$$

The zero vector in $\times_{i \in I} V_i$ is $(\mathbf{0}_i)_{i \in I}$, where $\mathbf{0}_i \in V_i$ denotes the zero vector of V_i. ◁

The properties defining a vector space imply several others that are useful to know. Look, for instance, at property (V3) about the existence of a zero vector. It says that there is *at least one element* of V, conveniently referred to as $\mathbf{0}$, that you can add to any vector x and get the sum x as a result. But can there be more than one such vector? In specific examples of vector spaces, you know that this is not the case. But the axioms don't state it explicitly. Indeed, the uniqueness of the zero vector and some other elementary properties are *consequences* of the definition of a vector space. That is the main lesson behind the following theorem. These properties won't come as a terrible surprise and in special vector spaces like $\mathbb{R}^2$ you have probably been using them for years without further reflection. But using only the defining properties of a vector space, they hold in *each* vector space, not just $\mathbb{R}^2$.

Theorem 3.1

Let V be a vector space. The following properties hold:

(a) Cancellation law: for all $x, y, z \in V$, if $x + z = y + z$, then $x = y$.

(b) Unique zero vector: there is exactly one $\mathbf{0} \in V$ such that $x + \mathbf{0} = x$ for all $x \in V$.

(c) Unique additive inverse: for each $x \in V$ there is exactly one $y \in V$ with $x + y = \mathbf{0}$.

(d) for each $x \in V$: $0x = \mathbf{0}$.

(e) for each $\alpha \in \mathbb{R}$: $\alpha\mathbf{0} = \mathbf{0}$.

(f) for each $x \in V$ and each $\alpha \in \mathbb{R}$: $(-\alpha)x = -(\alpha x) = \alpha(-x)$.

One observation before moving to the proof. Since each vector $w \in V$ has a unique additive inverse, we can call it $-w$ and define ***subtraction*** in vector spaces via the equation $v - w = v + (-w)$: subtracting a vector means adding its additive inverse. And using property (f) with $\alpha = -1$ and $x = w$, we see that $-w$ is simply the vector w multiplied with scalar -1.

Proof: **(a)** Let $x, y, z \in V$ satisfy $x+z = y+z$. By (V4), z has an additive inverse $w \in V$ with $z+w = \mathbf{0}$. So

$$x \overset{(V3)}{=} x+\mathbf{0} = x+(z+w) \overset{(V2)}{=} (x+z)+w = (y+z)+w \overset{(V2)}{=} y+(z+w) = y+\mathbf{0} \overset{(V3)}{=} y.$$

(b) Suppose both $\mathbf{0}$ and $\mathbf{0}'$ are candidates for the zero vector: for all $x \in V$, $x+\mathbf{0} = x+\mathbf{0}' = x$. By (V1), $\mathbf{0}+x = \mathbf{0}'+x$, so $\mathbf{0} = \mathbf{0}'$ by the cancellation law.
(c) Let $x \in V$. Suppose both y and y' are candidates for its additive inverse: $x+y = x+y' = \mathbf{0}$. By (V1), $y+x = y'+x$. By the cancellation law: $y = y'$.
(d) Let $x \in V$. Then $0x+1x \overset{(V8)}{=} (0+1)x = 1x \overset{(V3)}{=} 1x+\mathbf{0} \overset{(V1)}{=} \mathbf{0}+1x$. By the cancellation law: $0x = \mathbf{0}$.
(e) Let $\alpha \in \mathbb{R}$. Then $\alpha\mathbf{0}+\alpha\mathbf{0} \overset{(V7)}{=} \alpha(\mathbf{0}+\mathbf{0}) \overset{(V3)}{=} \alpha\mathbf{0} \overset{(V3)}{=} \alpha\mathbf{0}+\mathbf{0} \overset{(V1)}{=} \mathbf{0}+\alpha\mathbf{0}$. By the cancellation law: $\alpha\mathbf{0} = \mathbf{0}$.
(f) Let $x \in V$ and $\alpha \in \mathbb{R}$. By (c), the element $-(\alpha x)$ is the unique element of V such that $\alpha x+(-(\alpha x)) = \mathbf{0}$. Hence, if $\alpha x+(-\alpha)x = \mathbf{0}$, it follows that $(-\alpha)x = -(\alpha x)$. Now

$$\alpha x+(-\alpha)x \overset{(V8)}{=} (\alpha+(-\alpha))x = 0x \overset{\text{(d)}}{=} \mathbf{0}.$$

Thus, $(-\alpha)x = -(\alpha x)$. Similarly, $\alpha(-x) = -(\alpha x)$. □

We often look at subsets of a vector space V with additional nice properties. If such a smaller set — with addition and scalar multiplication as in the larger set V — satisfies all properties of a vector space, it is called a (linear) subspace:

Definition 3.2 A subset W of vector space V is a ***(linear) subspace*** of V if W itself is a vector space (using the rules for addition and scalar multiplication on V).

It is not necessary to verify all conditions on a vector space to conclude that W is a subspace of V. Intuitively, for most of the properties (V1) to (V8), the fact that they hold on the larger set V imply that they automatically hold on the subset W:

Theorem 3.2

A subset W of a vector space V is a subspace if and only if it satisfies the following three properties:

(i) W contains the zero vector from V: $\mathbf{0} \in W$,

(ii) W is closed under addition: $x+y \in W$ whenever $x \in W$ and $y \in W$,

(iii) W is closed under scalar multiplication: $\alpha x \in W$ whenever $x \in W$ and $\alpha \in \mathbb{R}$.

You are asked for the easy proof in Exercise 3.7.

Example 3.9 Verifying the three properties in the theorem above, you see that

$$W_1 = \{x \in \mathbb{R}^2 : 3x_1 - 4x_2 = 0\}$$

is a subspace of $\mathbb{R}^2$. But the following sets are not:

$$W_2 = \emptyset, \qquad W_3 = \{x \in \mathbb{R}^2 : x_1 = 0 \text{ or } x_2 = 0\}, \qquad W_4 = \{x \in \mathbb{R}^2 : x_1 \text{ and } x_2 \text{ are integers}\}.$$

W_2 does not contain the zero vector. W_3 is not closed under addition: it contains $(1,0)$ and $(0,1)$, but not their sum $(1,1)$. And W_4 is not closed under scalar multiplication: it contains $x = (1,1)$, but not its scalar multiple $\frac{1}{2}x = (\frac{1}{2}, \frac{1}{2})$. ◁

Example 3.10 If V is a vector space, then both V and $\{\mathbf{0}\}$ are subspaces. Also the intersection of a collection of subspaces of a vector space V is a subspace of V: since each of the subspaces separately satisfies the conditions of Theorem 3.2, their intersection satisfies them as well. ◁

It is often irrelevant whether we treat a vector in $\mathbb{R}^n$ as a row vector $x = (x_1, \ldots, x_n)$ or a column vector

$$x = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}. \tag{3}$$

To save space, one typically writes x as a row vector. Whenever the distinction *is* important, it is tradition to treat x as a column vector and denote the row vector as its transpose:

Definition 3.3 The ***transpose*** of an $m \times n$ matrix A is the $n \times m$ matrix $A^\top$ with entries $(A^\top)_{ij} = a_{ji}$: the consecutive rows of A become the consecutive columns of $A^\top$. Treating vector x in (3) as an $n \times 1$ matrix (n rows, but only one column):

$$x = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \text{ has transpose } x^\top = (x_1, \ldots, x_n).$$

Example 3.11

$$A = \begin{bmatrix} 0 & 1 & 2 \\ 3 & 4 & 5 \end{bmatrix} \text{ has transpose } A^\top = \begin{bmatrix} 0 & 3 \\ 1 & 4 \\ 2 & 5 \end{bmatrix} \text{ and } x = \begin{bmatrix} 3 \\ 8 \\ -2 \end{bmatrix} \text{ has transpose } x^\top = (3, 8, -2).$$

◁

Exercises section 3

3.1 The definition of a vector space uses two operations. Are the following claims true or false?

(a) One of these operations is addition.

(b) One of these operations is subtraction.

(c) One of these operations tells how to multiply two vectors with each other.

(d) One of these operations tells how to multiply a vector with a number.

3.2 (a) Is the set $W = \{x \in \mathbb{R}^n : x_1, \ldots, x_n \in \mathbb{Z}\}$ of vectors in $\mathbb{R}^n$ with integer coordinates a subspace of $\mathbb{R}^n$?

(b) Is the set $W = \{A \in \mathbb{R}^{n \times n} : A = A^\top\}$ of symmetric matrices a subspace of $\mathbb{R}^{n \times n}$?

(c) Is the set $W = \{f \in C[0,1] : f(0) = f(1)\}$ a subspace of $C[0,1]$?

(d) Is the set W of real sequences $x = (x_n)_{n \in \mathbb{N}}$ with $x_k \neq 0$ for only finitely many terms $k \in \mathbb{N}$ a subspace of $\mathbb{R}^{\mathbb{N}}$?

3.3 (Systems of linear equations) Let A be an $m \times n$ matrix and let $b \in \mathbb{R}^m$ be a vector with at least one coordinate distinct from zero. Are the following sets subspaces of $\mathbb{R}^n$?

(a) $\{x \in \mathbb{R}^n : Ax = \mathbf{0}\}$

(b) $\{x \in \mathbb{R}^n : Ax = b\}$

3.4 (The zero space) Prove that for any vector space V, the set $\{\mathbf{0}\}$ consisting only of its zero vector is a subspace.

3.5 For each i in a nonempty index set I, let V_i be a vector space.

(a) Prove, by verifying the properties in Definition 3.1, that the product space $\times_{i \in I} V_i$ (see Example 3.8) is a vector space.

(b) Prove that Examples 3.1, 3.2, 3.3, 3.4, and 3.5 are vector spaces by showing that they are product spaces $\times_{i\in I} V_i$ for suitable choices of the index set I and vector spaces V_i.

3.6 Prove, by verifying the properties in Definition 3.1, that the set $P(\mathbb{R})$ of real polynomials is a vector space.

3.7 (a) Prove Theorem 3.2.

(b) Prove the following alternative characterization of subspaces: W is a subspace of V if and only if W is nonempty and $\alpha x + \beta y \in W$ whenever $x, y \in W$ and $\alpha, \beta \in \mathbb{R}$.

4 Linear combinations, basis and dimension

4.1 Linear combinations, span, and linear (in)dependence

If you use the two fundamental operations of addition and scalar multiplication repeatedly on the set of vectors $W = \{w_1, w_2, w_3\}$ in $\mathbb{R}^2$ with

$$w_1 = (1,0), \quad w_2 = (-2,3), \quad w_3 = (2,1), \tag{4}$$

you can construct vectors like

$$3w_1 - 2w_2 + 4w_3 = (15,-2).$$

Such expressions are called linear combinations:

Definition 4.1 Let W be a subset of vector space V.

- ⊠ Vector $x \in V$ is a ***linear combination*** of vectors in W if there is a finite number $n \in \mathbb{N}$ of elements $w_1, \ldots, w_n$ in W and scalars $\alpha_1, \ldots, \alpha_n$ in $\mathbb{R}$ such that

$$x = \alpha_1 w_1 + \cdots + \alpha_n w_n.$$

- ⊠ The set of all linear combinations of vectors in W is called the ***span*** of W, denoted $\text{span}(W)$. By convention, $\text{span}(\emptyset) = \{\mathbf{0}\}$.
- ⊠ If $W' = \text{span}(W)$, one says that W' is spanned by W or that W spans W'.
- ⊠ If a vector space is spanned by a finite set W, it is ***finite-dimensional***; otherwise it is infinite-dimensional.

Notice that we only speak of linear combinations of finitely many vectors. By Theorem 3.2, the span of a subset W of a vector space V is a subspace of V.

Example 4.1 In $\mathbb{R}^3$, the span of the single vector $v \neq \mathbf{0}$ consists of all points on the line through v and the origin $(0,0,0)$. ◁

Example 4.2 The set of polynomials over $\mathbb{R}$ is spanned by the 'monomials' $1, x, x^2, x^3, x^4, \ldots$. It cannot be spanned by finitely many polynomials, because among them you could pick the one with the highest degree, say n. But then all linear combinations of those finitely many polynomials have a degree less than or equal to n: it does not contain higher-degree polynomials! So the set of polynomials is infinite-dimensional. ◁

The three vectors in (4) span $\mathbb{R}^2$. You don't even need all three of them. Omitting for instance the vector w_2, it is still possible to write each vector $b = (b_1, b_2) \in \mathbb{R}^2$ as a linear combination of w_1 and w_3:

$$(b_1, b_2) = (b_1 - 2b_2)\underbrace{(1,0)}_{=w_1} + b_2\underbrace{(2,1)}_{=w_3}.$$

In particular, the omitted vector $w_2 = (-2,3)$ is a linear combination of w_1 and w_3:

$$(-2,3) = -8(1,0) + 3(2,1).$$

Equivalently, moving all vectors to one side of the equality sign,

$$-8w_1 - 1w_2 + 3w_3 = (0,0) = \mathbf{0}.$$

In such a case, where a vector in W can be written as a linear combination of other vectors in W or, equivalently, where we can express the zero vector as a linear combination of distinct vectors in W with at least one of the scalars different from zero, we call W linearly dependent.

Definition 4.2 Let V be a vector space.

- ☒ A finite number $n \in \mathbb{N}$ of vectors $v_1, \ldots, v_n$ in V are ***linearly dependent*** if there are scalars $\alpha_1, \ldots, \alpha_n$, not all equal to zero, such that

$$\alpha_1 v_1 + \cdots + \alpha_n v_n = \mathbf{0}.$$

  Otherwise, they are ***linearly independent***.
- ☒ A (possibly infinite) subset W of V is ***linearly dependent*** if it contains a finite number $n \in \mathbb{N}$ of distinct vectors that are linearly dependent. Otherwise, W is ***linearly independent***.

In particular:

A finite set $W = \{w_1, \ldots, w_n\}$ of different vectors is linearly independent if and only if

$$\alpha_1 w_1 + \cdots + \alpha_n w_n = \mathbf{0} \tag{5}$$

implies that $\alpha_1 = \cdots = \alpha_n = 0$.

So to check for linear independence of a finite set W, write down and solve the equation (5) for the alphas. If all of them are zero, your set is linearly independent; otherwise it is linearly dependent.

Example 4.3 We argued above that the set of vectors $\{(1,0), (-2,3), (2,1)\}$ is linearly dependent. What about $\{(1,0), (2,1)\}$? We solve

$$\alpha_1(1,0) + \alpha_2(2,1) = (0,0).$$

Rewritten as a system of linear equations

$$\begin{aligned} \alpha_1 + 2\alpha_2 &= 0, \\ \alpha_2 &= 0, \end{aligned}$$

it follows that $\alpha_1 = \alpha_2 = 0$ is the only solution: these vectors are linearly independent. ◁

Example 4.4 In the vector space $C[0,1]$ of continuous functions from $[0,1]$ to $\mathbb{R}$, the functions f and g with $f(x) = 3x^2 - x$ and $g(x) = 4e^x$ are linearly independent:

$$\alpha_1 f + \alpha_2 g = \mathbf{0} \quad \Longleftrightarrow \quad \text{for all } x \in [0,1]: \quad \alpha_1(3x^2 - x) + \alpha_2(4e^x) = 0. \tag{6}$$

Substituting $x = 0$ gives that $\alpha_2 = 0$. So (6) simplifies to

$$\text{for all } x \in [0,1]: \quad \alpha_1(3x^2 - x) = 0.$$

Substituting $x = 1$ gives that $\alpha_1 = 0$. Conclude that the only linear combination of f and g that gives the zero function has $\alpha_1 = \alpha_2 = 0$: they are linearly independent. ◁

Here are some other easy, but useful observations about linearly (in)dependent sets:

- ☒ Linearly dependent sets must be nonempty, so the empty set $\emptyset$ is linearly independent.
- ☒ A set $\{w\}$ consisting of a single vector $w \in V$ is linearly independent if and only if w is not the zero vector. Indeed, set $\{w\}$ is linearly dependent if and only if $\alpha w = \mathbf{0}$ for some nonzero scalar α. Multiplying both sides with $\frac{1}{\alpha}$ and using that $\frac{1}{\alpha}\mathbf{0} = \mathbf{0}$, this is equivalent with $w = \mathbf{0}$.
- ☒ A set $W \subseteq V$ is linearly dependent if and only if $W = \{\mathbf{0}\}$ or there exist distinct vectors $w, w_1, \ldots, w_n$ in W such that w is a linear combination of $w_1, \ldots, w_n$. See Exercise 4.3.

4.2 Basis and dimension

Bases are the building blocks of vector spaces. A basis for a vector space V is a subset that is so large that each element of V can be written as a linear combination of vectors in the basis, but so small that you cannot omit elements from the basis and still span the entire vector space V. Formally:

Definition 4.3 A ***basis*** for a vector space V is a linearly independent subset of V that spans V.

This subsection contains two crucial results:

1. Every vector space has a basis;
2. In a finite-dimensional vector space, all bases have the same number of elements.

The latter result allows us to unambiguously define the ***dimension*** of a finite-dimensional vector space V, denoted $\dim(V)$, to be the number of elements of a basis.

Example 4.5 In the recurrent example in this section we saw that vectors $(1,0)$ and $(2,1)$ span $\mathbb{R}^2$ and are linearly independent: $\{(1,0),(2,1)\}$ is a basis of $\mathbb{R}^2$, making it two-dimensional. ◁

Our next example introduces a more common basis for $\mathbb{R}^2$ and, more generally, for $\mathbb{R}^n$.

Example 4.6 (Standard basis for $\mathbb{R}^n$) The i-th ***standard basis vector*** in $\mathbb{R}^n$ is the vector $e_i \in \mathbb{R}^n$ whose i-th coordinate is 1 and all other coordinates are 0:

$$e_1 = (1,0,\ldots,0), e_2 = (0,1,0,\ldots,0), \ldots, e_n = (0,\ldots,0,1).$$

The set $\{e_1,\ldots,e_n\}$ is easily seen to be a basis for $\mathbb{R}^n$ and is called the ***standard basis*** for $\mathbb{R}^n$, making it n-dimensional. Notice that $x = \sum_{i=1}^n x_i e_i$ for each $x \in \mathbb{R}^n$. For instance, $\mathbb{R}^2$ has standard basis $\{e_1,e_2\} = \{(1,0),(0,1)\}$ and each $x \in \mathbb{R}^2$ can be written as $x = (x_1,x_2) = x_1(1,0) + x_2(0,1)$. ◁

Example 4.7 The set $\{1,x,x^2,x^3,\ldots\}$ is a basis for the set of polynomials over $\mathbb{R}$, making it infinite-dimensional; similarly, the set $\{1,x,x^2,\ldots,x^n\}$ is a basis for the set of polynomials with degree at most n, making that subspace $(n+1)$-dimensional. ◁

Example 4.8 Recalling that $\text{span}(\emptyset) = \{\mathbf{0}\}$ and that $\emptyset$ is linearly independent, we see that $\emptyset$ is a basis for the vector space $\{\mathbf{0}\}$. Since its basis has zero elements, it is zero-dimensional. ◁

We speak of an ***ordered basis*** if we give the elements of a basis a specific order. The standard basis $\{e_1,\ldots,e_n\}$ for $\mathbb{R}^n$ and the polynomials $\{1,x,x^2,x^3,\ldots\}$ with increasing powers in $P(\mathbb{R})$ are examples of ordered bases.

According to our next result, whenever you have a linearly independent subset of a vector space V and a larger set spanning V, you can always find a basis in-between:

Theorem 4.1 (Existence of bases, extension and reduction)

Let the following be given:

- ☒ a vector space V,
- ☒ a linearly independent subset I of V,
- ☒ a subset S of V that spans V and contains I.

Then there is a basis B for V with $I \subseteq B \subseteq S$. Consequently:

Existence: every vector space has a basis;

Extension: every linearly independent set I in V is contained in a basis;

Reduction: every set that spans V contains a basis.

The first part uses an advanced set-theoretic tool, Zorn's Lemma. It is treated in Section 5 and we skip it here. But once we have our basis B with $I \subseteq B \subseteq S$, the three consequences are easy. For existence, let $I = \emptyset, S = V$. For extension, let I be the given linearly independent subset and $S = V$. For reduction, let S be the given spanning set and $I = \emptyset$.

Our second main result concerns the size of bases of finite-dimensional vector spaces:

Theorem 4.2

Let V be a vector space that has

- ⊠ $m \in \mathbb{N}$ vectors $s_1, \ldots, s_m$ that span V,
- ⊠ $n \in \mathbb{N}$ vectors $v_1, \ldots, v_n$ that are linearly independent.

Then $n \leq m$. Hence, V has a finite basis and all bases of V have the same number of elements.

Proof: Since $s_1, \ldots, s_m$ span V, there are scalars $\alpha_1, \ldots, \alpha_m$ such that

$$v_1 = \alpha_1 s_1 + \cdots + \alpha_m s_m.$$

Since $v_1 \neq \mathbf{0}$ by linear independence of $v_1, \ldots, v_n$, at least one α_i is distinct from zero. Relabeling if necessary, we may assume $\alpha_1 \neq 0$. Solve for s_1:

$$s_1 = \tfrac{1}{\alpha_1} v_1 + \left(-\tfrac{\alpha_2}{\alpha_1}\right) s_2 + \cdots + \left(-\tfrac{\alpha_m}{\alpha_1}\right) s_m.$$

So $v_1, s_2, \ldots, s_m$ span V: we replaced s_1 by v_1, but still have a set that spans V.

Repeat the process with v_2: there are scalars $\alpha_1, \ldots, \alpha_m$ such that

$$v_2 = \alpha_1 v_1 + \alpha_2 s_2 + \cdots + \alpha_m s_m.$$

As before, $v_2 \neq \mathbf{0}$ and not all of $\alpha_2, \ldots, \alpha_m$ can be zero by linear independence of the v_i's. Relabeling if necessary, we may assume that $\alpha_2 \neq 0$ and solve for s_2 to show that $v_1, v_2, s_3, \ldots, s_m$ span the same set as $v_1, s_2, \ldots, s_m$. So $v_1, v_2, s_3, \ldots, s_m$ span V: we replaced s_1 and s_2 by v_1 and v_2, but still have a set that spans V.

If $m < n$, this process will eventually exhaust the s_i's and lead to the conclusion that $v_1, \ldots, v_m$ spans V. This is impossible by linear independence of $v_1, \ldots, v_m, v_{m+1}, \ldots, v_n$, since v_n is not in the span of $v_1, \ldots, v_m$. Hence, $n \leq m$.

For the final claim, V has a basis by Theorem 4.1. Since a basis is linearly independent, our previous step gives that the basis is finite. Let B_1 and B_2 be two bases for V with m and n elements, respectively. Since a basis is linearly independent and spans V, two-fold application of our previous step yields that $n \leq m$ and $m \leq n$, proving that $m = n$: the two bases have the same number of elements. □

This theorem can be used to show, without any computations, that certain sets cannot possibly be linearly independent or span a given vector space.

Example 4.9

- ⊠ Since $\mathbb{R}^3$ is spanned by the three standard basis vectors $(1,0,0)$, $(0,1,0)$, and $(0,0,1)$, the four vectors $(1,3,-2)$, $(-2,1,4)$, $(0,1,0)$, and $(0,-4,5)$ cannot be linearly independent.
- ⊠ In $\mathbb{R}^3$, the three vectors $(1,0,0)$, $(0,1,0)$, and $(0,0,1)$ are linearly independent. So the two vectors $(1,2,3)$ and $(0,-2,5)$ cannot span $\mathbb{R}^3$. ◁

And on the positive side:

Theorem 4.3

In an n-dimensional vector space V,

(a) each linearly independent subset with exactly n elements is a basis;

(b) each subset of exactly n elements that spans V is a basis.

Proof: I prove (a); the proof of (b) is similar (Exercise!). By the extension part of Theorem 4.1, our set I of n linearly independent vectors can be extended to a basis of V. But V is n-dimensional, so each basis has n elements. Since I already has n elements, it follows that I is a basis. □

Example 4.10 The three vectors $(1,1,1)$, $(0,1,1)$, and $(0,0,1)$ in $\mathbb{R}^3$ are linearly independent: if you solve

$$\alpha_1 \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} + \alpha_2 \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix} + \alpha_3 \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix},$$

looking at its first coordinate shows that $\alpha_1 = 0$. Then its second coordinate gives $\alpha_2 = 0$ and its third coordinate gives $\alpha_3 = 0$. And $\mathbb{R}^3$ is three-dimensional, so by Theorem 4.3 they are a basis of $\mathbb{R}^3$. ◁

Exercises section 4

4.1 Are the following sets of vectors linearly independent?

(a) $W = \{(1,0),(2,-1)\}$ in $\mathbb{R}^2$.

(b) $W = \{(1,2,3),(0,2,3),(-4,4,5)\}$ in $\mathbb{R}^3$.

(c) $W = \{(1,2,3),(0,2,3),(1,-2,-3)\}$ in $\mathbb{R}^3$.

(d) $W = \{3, x, 2x^2 + x - 2\}$ in the space $P(\mathbb{R})$ of polynomials.

(e) $W = \{f, g\}$ consisting of functions f with $f(x) = x$ and g with $g(x) = 1/(x+2)$ in $C[0,1]$.

4.2 Prove:

(a) $\text{span}(W)$ is the smallest subspace containing W, i.e., $\text{span}(W) \subseteq U$ for every subspace U containing W.

(b) $\text{span}(W)$ is the intersection of all subspaces containing W.

4.3 Prove: A subset W of vector space V is linearly dependent if and only if $W = \{\mathbf{0}\}$ or there exist distinct vectors $w, w_1, \ldots, w_n$ in W such that w is a linear combination of $w_1, \ldots, w_n$.

5 Why each vector space has a basis: Zorn's lemma

Several of the more advanced results in these notes require an intricate tool from set theory, Zorn's lemma. To state it, we need a few definitions.

Definition 5.1 (Maximal elements, chains, upper bounds) Let $\mathscr{A}$ be a collection of sets.

- ☒ A member M of $\mathscr{A}$ is ***maximal*** (with respect to set inclusion) if $\mathscr{A}$ contains no strictly larger set, i.e., there is no set N in $\mathscr{A}$ with $M \subseteq N$ and $M \neq N$.
- ☒ A subset $\mathscr{C}$ of $\mathscr{A}$ is a ***chain*** if for each pair of elements A and B in $\mathscr{C}$, either $A \subseteq B$ or $B \subseteq A$.
- ☒ A chain $\mathscr{C}$ in $\mathscr{A}$ has an ***upper bound*** if there is an element $U \in \mathscr{A}$ that contains all members of $\mathscr{C}$: $C \subseteq U$ for all $C \in \mathscr{C}$.

Example 5.1 If S is a set, then the ***power set*** of S is the set of all subsets of S. It is denoted by 2^S. For instance, if $S = \{1,2,3\}$, then

$$2^S = \{\emptyset, \{1\}, \{2\}, \{3\}, \{1,2\}, \{1,3\}, \{2,3\}, \{1,2,3\}\}.$$

The power set 2^S always contains a maximal element, namely S. But if S has two or more elements, 2^S is not a chain: if i and j are distinct elements of S, then $\{i\}$ and $\{j\}$ belong to 2^S, but neither $\{i\} \subseteq \{j\}$ nor $\{j\} \subseteq \{i\}$. Note that $\mathscr{C} = \{\emptyset, \{1\}, \{1,3\}, \{1,2,3\}\}$ is a chain. ◁

Example 5.2 Let $\mathscr{C} = \mathscr{A}$ be the collection of intervals $(-\infty, a]$ with $a \in \mathbb{R}$. Then $\mathscr{C}$ is a chain: for any pair of elements $(-\infty, a]$ and $(-\infty, b]$ of $\mathscr{C}$ either $(-\infty, a] \subseteq (-\infty, b]$ (if $a \leq b$) or $(-\infty, b] \subseteq (-\infty, a]$ (if $b \leq a$). This chain has no upper bound, because for each $(-\infty, a]$, you can find a larger set in $\mathscr{C}$, for instance $(-\infty, a+1]$. ◁

Exercise 5.1 gives more examples. One of the main things to get straight is that maximal elements of $\mathscr{A}$ are defined in terms of set inclusion: it has nothing to do with large numbers. And maximal elements of $\mathscr{A}$ do not have to contain many elements either. The only requirement on a maximal element is that it is not a subset of any other set in $\mathscr{A}$. For instance, in $\mathscr{A} = \{\{0\}, \{1\}, \{2\}, \{2,3\}\}$ there are three maximal elements: $\{0\}, \{1\}$, and $\{2,3\}$. These are the sets that are not contained in any other set of $\mathscr{A}$.

Zorn's lemma gives a condition for a collection of sets to have a maximal element:

Theorem 5.1 (Zorn's lemma)

If each chain in a collection of sets $\mathscr{A}$ has an upper bound, then $\mathscr{A}$ has a maximal element.

In most applications of Zorn's lemma within economic theory, the upper bound is simply the union of all sets in the chain.

Zorn's lemma may not be very intuitive, but turns out to be equivalent with more palatable assumptions in hard-core axiomatic set theory such as the Axiom of Choice, according to which the Cartesian product of any nonempty collection of nonempty sets is again nonempty. Like most applied mathematicians, we simply presume it to be true. In fact, it is one of the axioms/assumptions in the standard approach to axiomatic set theory that underlies applied mathematics, often referred to as the ZFC axiom system. Z and F are Zermelo and Fraenkel, who did fundamental work in this area, and C is an explicit reminder that it includes the Axiom of Choice — and by equivalence, Zorn's lemma.

Let us use Zorn's lemma to prove the part of Theorem 4.1 we hadn't established yet: suppose that in a vector space V we have a linearly independent subset I and a subset S that spans V, satisfying $I \subseteq B$. Then we can find a basis B with $I \subseteq B \subseteq S$.

Let $\mathscr{A}$ be the collection of all linearly independent subsets of V that contain I and are contained in S. Since $I \in \mathscr{A}$, this collection is nonempty.

We use Zorn's lemma to show that $\mathscr{A}$ has a maximal element B. This B is the desired basis: by construction, $I \subseteq B \subseteq S$ and B is linearly independent. It spans V because S spans V and each element s of S lies in $\operatorname{span}(B)$: if $s \notin \operatorname{span}(B)$, then $I \subseteq B \cup \{s\} \subseteq S$ is linearly independent, contradicting the maximality of B.

To use Zorn's lemma, let $\mathscr{C}$ be a chain in $\mathscr{A}$. I claim that the union $U = \cup_{C \in \mathscr{C}} C$ of its elements is an upper bound. Clearly, $I \subseteq U \subseteq S$ and $C \subseteq U$ for each $C \in \mathscr{C}$. To show that $U \in \mathscr{A}$, it remains to establish that U is linearly independent.

So let $u_1, \ldots, u_n$ be finitely many elements of U and suppose there are scalars $\alpha_1, \ldots, \alpha_n$ such that $\alpha_1 u_1 + \cdots + \alpha_n u_n = \mathbf{0}$. For each $i = 1, \ldots, n$, $u_i \in U$ implies that there is a set $C_i \in \mathscr{C}$ with $u_i \in C_i$. Since $\mathscr{C}$ is a chain, we may assume without loss of generality that $C_i \subseteq C_n$ for all i, i.e., that C_n is the largest of these n sets. Hence, $\{u_1, \ldots, u_n\} \subseteq C_n$. The linear independence of C_n implies that $\alpha_1 = \cdots = \alpha_n = 0$. Since each chain in $\mathscr{A}$ has an upper bound, $\mathscr{A}$ has a maximal element by Zorn's lemma!

Exercises section 5

5.1 Answer the following questions:

- ☒ give an example of a collection of sets in $\mathscr{A}$ that is not a chain;
- ☒ give an example of a collection of at least two sets in $\mathscr{A}$ that is a chain;
- ☒ does each chain in $\mathscr{A}$ have an upper bound?
- ☒ does $\mathscr{A}$ have a maximal element?

if $\mathscr{A}$ is:

(a) the collection of finite subsets of $\mathbb{N} = \{1, 2, 3, \ldots\}$;

(b) the collection consisting of $\{-37\}$ and all finite subsets of $\mathbb{N}$;

(c) the collection of subsets of $\mathbb{N}$ with at most two elements.

6 Linear functions

Definition 6.1 Let V and W be two vector spaces. Function $T : V \to W$ is ***linear*** or a ***linear transformation*** if it preserves the addition and scalar multiplication properties:

$$T(x+y) = T(x) + T(y) \quad \text{and} \quad T(\alpha x) = \alpha T(x) \qquad \text{for all } x, y \in V \text{ and scalars } \alpha. \tag{7}$$

The set of linear functions from V to W is denoted by $L(V, W)$.

Values of linear functions are sometimes written without parentheses: Tx instead of $T(x)$.

Example 6.1 By the rules of matrix multiplication, if A is an $m \times n$ matrix of real numbers, then

$$A(x+y) = Ax + Ay \qquad \text{and} \qquad A(\alpha x) = \alpha(Ax)$$

for all vectors x and y in $\mathbb{R}^n$ and all scalars α. So the function $T : \mathbb{R}^n \to \mathbb{R}^m$ with $T(x) = Ax$ is linear. For instance, $T : \mathbb{R}^3 \to \mathbb{R}^2$ with $T(x_1, x_2, x_3) = (3x_1 - x_3, 2x_1 + x_2 + 4x_3)$ is linear. It can be written as

$$T(x) = \begin{bmatrix} 3x_1 + 0x_2 - 1x_3 \\ 2x_1 + 1x_2 + 4x_3 \end{bmatrix} = x_1 \begin{bmatrix} 3 \\ 2 \end{bmatrix} + x_2 \begin{bmatrix} 0 \\ 1 \end{bmatrix} + x_3 \begin{bmatrix} -1 \\ 4 \end{bmatrix} = \begin{bmatrix} 3 & 0 & -1 \\ 2 & 1 & 4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = Ax$$

with

$$A = \begin{bmatrix} 3 & 0 & -1 \\ 2 & 1 & 4 \end{bmatrix}.$$

But $S : \mathbb{R}^2 \to \mathbb{R}$ with $S(x_1, x_2) = x_1 x_2$ is not linear: if we choose $x = (1, 0)$ and $y = (0, 1)$, then $S(x+y) \neq S(x) + S(y)$, because

$$S(x+y) = S(1,1) = 1, \qquad \text{but} \qquad S(x) + S(y) = S(1,0) + S(0,1) = 0 + 0 = 0. \qquad \triangleleft$$

Example 6.2 The function T that assigns to each continuous function $f : [0,1] \to \mathbb{R}$ its integral,

$$T(f) = \int_0^1 f(x)\,\mathrm{d}x$$

is a linear function from $C[0,1]$ to $\mathbb{R}$. This is due to familiar facts about integration: the integral of the sum of two functions is the sum of their integrals and the integral of a constant times a function is that constant times the integral of the function. That is,

$$\int_0^1 (f(x) + g(x))\,\mathrm{d}x = \int_0^1 f(x)\,\mathrm{d}x + \int_0^1 g(x)\,\mathrm{d}x \qquad \text{and} \qquad \int_0^1 \alpha f(x)\,\mathrm{d}x = \alpha \int_0^1 f(x)\,\mathrm{d}x. \qquad \triangleleft$$

A linear function is determined completely by its behavior on a basis: if $T : V \to W$ is linear and you know $T(x)$ for each x in a basis of vector space V, then you can compute the function value of any vector $v \in V$ by writing it as a linear combination $v = \alpha_1 x_1 + \cdots + \alpha_k x_k$ of basis vectors $x_1, \ldots, x_k$ and using linearity to obtain:

$$T(v) = T(\alpha_1 x_1 + \cdots + \alpha_k x_k) = \alpha_1 T(x_1) + \cdots + \alpha_k T(x_k).$$

Example 6.3 If $T : \mathbb{R}^2 \to \mathbb{R}$ is linear and $T(1,0) = 4$ and $T(0,1) = -3$, then for all $(x_1, x_2) \in \mathbb{R}^2$, we can write $(x_1, x_2) = x_1(1,0) + x_2(0,1)$. So by linearity,

$$T(x_1, x_2) = T(x_1(1,0) + x_2(0,1)) = T(x_1(1,0)) + T(x_2(0,1)) = x_1 T(1,0) + x_2 T(0,1) = 4x_1 - 3x_2. \qquad \triangleleft$$

Example 6.4 A function $T:\mathbb{R}^n \to \mathbb{R}^m$ is linear if and only if there is an $m \times n$ matrix A with $T(x) = Ax$. We already did half of the work in Example 6.1, showing that functions of the form $T(x) = Ax$ are linear. Conversely, if T is a linear function from $\mathbb{R}^n$ to $\mathbb{R}^m$, write $x \in \mathbb{R}^n$ in terms of the standard basis vectors $e_1, \ldots, e_n$ and use linearity to conclude that

$$T(x) = T\left(\sum_{j=1}^{n} x_j e_j\right) = \sum_{j=1}^{n} x_j T(e_j) = Ax,$$

where A is the matrix whose columns are $T(e_1)$ until $T(e_n)$, respectively. ◁

Before moving on to a more substantial theorem, I list a few properties related to linear functions that follow easily from the definitions (Exercise 6.5):

- ☒ The set $L(V,W)$ of linear functions from vector space V to vector space W is a subspace of the vector space of all functions from V to W.

And if $T: V \to W$ is linear, then:

- ☒ $T(\mathbf{0}) = \mathbf{0}$.
- ☒ The range of T, denoted $\text{range}(T) = \{T(v) : v \in V\}$, is a subspace of W.
- ☒ The set $\ker(T) = \{v \in V : T(v) = \mathbf{0}\}$ of vectors mapped to the zero vector is a subspace of V; it is called the ***kernel*** or ***null space*** of T.

Our next result relates the dimension of the range and kernel of a linear transformation. These dimensions are often called the rank and nullity, respectively, explaining the name of the theorem.

Theorem 6.1 (Rank-nullity theorem)

Let $T: V \to W$ be a linear transformation between vector spaces V and W, with V finite-dimensional. Then

$$\dim(\underbrace{\ker(T)}_{\subseteq V}) + \dim(\underbrace{\text{range}(T)}_{\subseteq W}) = \dim(V).$$

Proof: Write $\dim(V) = n$ and $\dim(\ker(T)) = m \leq n$. If $m = n$, then *all* $v \in V$ are mapped to $\mathbf{0}$ and $\text{range}(T) = \{\mathbf{0}\}$ has dimension zero. If $m < n$, let $v_1, \ldots, v_m$ be a basis of $\ker(T)$ and extend it with $n-m$ vectors $v_{m+1}, \ldots, v_n$ to a basis of V. I will prove that $Tv_{m+1}, \ldots, Tv_n$ is a basis of $\text{range}(T)$.

They span $\text{range}(T)$: if $w \in \text{range}(T)$, there is a $v \in V$ with $Tv = w$. Express v as a linear combination of basis vectors

$$v = \alpha_1 v_1 + \cdots + \alpha_m v_m + \alpha_{m+1} v_{m+1} + \cdots + \alpha_n v_n.$$

Now apply T. By linearity and the fact that $v_1, \ldots, v_m$ is a basis of its kernel:

$$w = Tv = \alpha_1 \underbrace{Tv_1}_{=\mathbf{0}} + \cdots + \alpha_m \underbrace{Tv_m}_{=\mathbf{0}} + \alpha_{m+1} Tv_{m+1} + \cdots + \alpha_n Tv_n = \alpha_{m+1} Tv_{m+1} + \cdots + \alpha_n Tv_n.$$

So $w \in \text{span}\{Tv_{m+1}, \ldots, Tv_n\}$. Since w was arbitrary, $\text{range}(T) = \text{span}\{Tv_{m+1}, \ldots, Tv_n\}$.

They are linearly independent: if scalars $\alpha_{m+1}, \ldots, \alpha_n$ are such that

$$\mathbf{0} = \alpha_{m+1} Tv_{m+1} + \cdots + \alpha_n Tv_n = T(\alpha_{m+1} v_{m+1} + \cdots + \alpha_n v_n),$$

then $\alpha_{m+1} v_{m+1} + \cdots + \alpha_n v_n \in \ker(T) = \text{span}\{v_1, \ldots, v_m\}$. Since $v_1, \ldots, v_m, v_{m+1}, \ldots, v_n$ is a basis of V, this implies that $\alpha_{m+1} = \cdots = \alpha_n = 0$: otherwise the vectors in this basis would be linearly dependent, a contradiction. □

Example 6.5 The columns of the 2×3 matrix

$$A = \begin{bmatrix} 3 & 0 & -1 \\ 2 & 1 & 4 \end{bmatrix}.$$

from Example 6.1 span $\mathbb{R}^2$, so the range of the linear function $T : \mathbb{R}^3 \to \mathbb{R}^2$ with $T(x) = Ax$ is all of $\mathbb{R}^2$. Hence $\dim(\text{range}(T)) = 2$. And $\dim(\mathbb{R}^3) = 3$, so according to the rank-nullity theorem, the null space of T has dimension one. That is easy to verify: solving $T(x) = Ax = \mathbf{0}$, for instance using Gaussian elimination, you see that the null space consist of all multiples of the vector $(1, -14, 3)$, i.e., it is the one-dimensional space $\text{span}\{(1, -14, 3)\}$. ◁

Exercises section 6

6.1 A linear function $T : \mathbb{R}^2 \to \mathbb{R}^3$ has $T(1,1) = (2,0,-3)$ and $T(0,2) = (-1,4,2)$. Find a matrix A such that $T(x) = Ax$.

6.2 Consider the linear function $T : \mathbb{R}^5 \to \mathbb{R}^4$ with $T(x) = Ax$, where

$$A = \begin{bmatrix} 1 & 4 & 5 & 6 & 9 \\ 3 & -2 & 1 & 4 & -1 \\ 1 & 0 & -1 & -2 & -1 \\ 2 & 3 & 5 & 7 & 8 \end{bmatrix}$$

(a) The null space of T is the set of all vectors x with $T(x) = \mathbf{0}$. Determine the null space of T.

(b) Using your previous answer, find a basis for the null space. What is its dimension?

Are the following sets of vectors also a basis of the null space? There are short answers with very few computations.

(c) $B_1 = \{(0,0,-3,1,1), (0,2,2,0,-2), (0,-1,2,-1,0)\}$.

(d) $B_2 = \{(0,-3,-3,0,3), (0,-1,-1,0,1)\}$.

(e) $B_3 = \{(0,1,1,0,-1), (0,3,0,1,-2)\}$.

6.3 (An alternative definition of linearity) Show that a function $T : V \to W$ between two vector spaces is linear if and only if $T(\alpha x + \beta y) = \alpha T(x) + \beta T(y)$ for all $x, y \in V$ and all scalars α, β.

6.4 Show that a linear function is injective if and only if its null space contains only the zero vector.

6.5 Prove the claims about linear functions preceding the rank-nullity theorem.

6.6 Recall from Example 3.5 that $\mathbb{R}^{\mathbb{N}}$ is the vector space of sequences $(x_1, x_2, x_3, x_4, \ldots)$ of real numbers. Show:

(a) The function $T : \mathbb{R}^{\mathbb{N}} \to \mathbb{R}^{\mathbb{N}}$ with $T(x_1, x_2, x_3, x_4, \ldots) = (x_2, x_3, x_4, \ldots)$ is linear, surjective, but not injective.

(b) The function $T : \mathbb{R}^{\mathbb{N}} \to \mathbb{R}^{\mathbb{N}}$ with $T(x_1, x_2, x_3, x_4, \ldots) = (0, x_1, x_2, x_3, \ldots)$ is linear, injective, but not surjective.

6.7 Let $T : V \to V$ be a linear function from and to the same finite-dimensional vector space V. Use the rank-nullity theorem to show that the following three claims are equivalent:

(a) T is bijective.

(b) T is injective.

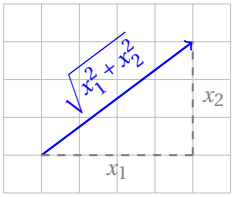
(c) T is surjective.

7 Normed vector spaces and inner product spaces

7.1 Normed vector spaces

Using Pythagoras' Law, the length $\|x\|$ of vector $x = (x_1, x_2) \in \mathbb{R}^2$ is defined as

$$\|x\| = \sqrt{x_1^2 + x_2^2}. \tag{8}$$



The goal of this section is to extend the notion of 'length' of a vector — or, at the very least, some intuitively desirable properties that such a notion should have — to arbitrary vector spaces. Let's start simple and just extend (8) to vectors of n real numbers:

Example 7.1 Define the length of vector $x = (x_1, \ldots, x_n)$ in $\mathbb{R}^n$ to be

$$\|x\|_2 = \sqrt{x_1^2 + \cdots + x_n^2}. \tag{9}$$

Does this definition conform with some of the properties you might expect from colloquial use of the word 'length'? It is immediate from (9) that:

> *Each vector has nonnegative length.*

So far so good: if you'd told me that a vector had length -7, I'd be a bit worried. Moreover, since the squared numbers x_i^2 are all nonnegative, the only way to get a vector of length zero is to set all coordinates equal to zero:

> *Only the zero vector has length zero.*

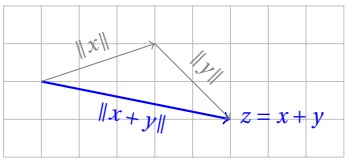
What is the length of 3 times the vector x (or -3 times, which has the same effect on the magnitude of the coordinates, but changes their sign)? Arguably, this is just three times the length of x. So:

> *Rescaling a vector by some number α rescales its length by a factor $|\alpha|$.*

Recall that $|\alpha| = \max\{\alpha, -\alpha\}$ is the absolute value of α: it equals α if $\alpha \geq 0$ and it equals $-\alpha$ if $\alpha < 0$. This property follows from substitution in (9):

$$\|\alpha x\|_2 = \sqrt{(\alpha x_1)^2 + \cdots + (\alpha x_n)^2} = \sqrt{(\alpha^2)(x_1^2 + \cdots + x_n^2)} = \sqrt{\alpha^2}\sqrt{x_1^2 + \cdots + x_n^2} = |\alpha|\,\|x\|_2.$$

A final property, the so-called triangle inequality, reflects the intuition that detours cannot decrease distance. Travelling the length of vector $z = x + y$ cannot be longer than first travelling the length of vector x and then the length of vector y. In other words:



> *The length of the sum of two vectors is at most the sum of their lengths.*

That lengths as defined in (9) have this property is established in Theorem 7.1. ◁

In arbitrary vector spaces, lengths are modelled by functions called norms that satisfy the emphasized properties from the example above:

Definition 7.1 Let V be a vector space. A function $x \mapsto \|x\| \in \mathbb{R}$ defined for all $x \in V$ is a ***norm*** on V if it satisfies:

(N1) for all $x \in V$: $\|x\| \geq 0$.

(N2) $\|x\| = 0$ if and only if $x = \mathbf{0}$.

(N3) for all $x \in V$ and all $\alpha \in \mathbb{R}$: $\|\alpha x\| = |\alpha| \, \|x\|$.

(N4) triangle inequality: for all $x, y \in V$: $\|x + y\| \leq \|x\| + \|y\|$.

The pair $(V, \|\cdot\|)$, or just V if the norm is left implicit, is a ***normed vector space***.

In this terminology, Example 7.1 becomes:

Example 7.2 $(\mathbb{R}^n, \|\cdot\|_2)$ is a normed vector space, where

$$\|x\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2}.$$

The norm $\|\cdot\|_2$ is called the ***Euclidean norm*** or the ℓ_2***-norm***. If $n = 1$, the Euclidean norm on $\mathbb{R}$ is simply the absolute value, so $(\mathbb{R}, |\cdot|)$ is a normed vector space. ◁

The example above is arguably the most common normed vector space; here are a few more.

Example 7.3 $(\mathbb{R}^n, \|\cdot\|_1)$ is a normed vector space, where

$$\|x\|_1 = \sum_{i=1}^{n} |x_i|. \tag{10}$$

The norm $\|\cdot\|_1$ is sometimes called the ***Manhattan norm***, the ***taxicab norm***, or the ℓ_1***-norm***: on Manhattan's rectangular grid of streets moving north-south or east-west, and divided by city blocks, a trip of 4 blocks east and 3 blocks south (somewhat informally, the vector $(4, -3)$) is a trip with a length of $4 + |-3| = 7$ blocks. ◁

Example 7.4 $(\mathbb{R}^n, \|\cdot\|_\infty)$ is a normed vector space, where

$$\|x\|_\infty = \sup_i |x_i| = \max\{|x_1|, \ldots, |x_n|\}. \tag{11}$$

The norm $\|\cdot\|_\infty$ is called the ***supremum norm*** or the ℓ_∞***-norm***. ◁

Example 7.5 $(\mathbb{R}^n, \|\cdot\|_p)$ is a normed vector space, where $1 \leq p < \infty$, and

$$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p}. \tag{12}$$

The norm $\|\cdot\|_p$ is called the ***Hölder norm*** or ℓ_p***-norm***. The special cases $p = 1$ and $p = 2$ were encountered in Examples 7.3 and 7.2. The notation in Example 7.4 is explained by the fact that for each $x \in \mathbb{R}^n$:

$$\lim_{p \to \infty} \|x\|_p = \|x\|_\infty. \tag{13}$$

Exercise 7.6 establishes the triangle inequality. ◁

Example 7.6 Call a sequence $(x_1, x_2, x_3, \ldots)$ of real numbers ***bounded*** if we can find a number b such that $-b \leq x_i \leq b$ for each $i \in \mathbb{N}$. Let $B(\mathbb{N})$ denote the set of bounded real sequences. $(B(\mathbb{N}), \|\cdot\|_\infty)$ is a normed vector space, where

$$\|x\|_\infty = \sup_{i \in \mathbb{N}} |x_i| \qquad \text{for each } x = (x_1, x_2, x_3, \ldots) \in B(\mathbb{N}). \tag{14}$$

◁

Example 7.7 $(C[a,b], \|\cdot\|_\infty)$ is a normed vector space, where

$$\|f\|_\infty = \max\{|f(x)| : x \in [a,b]\} \qquad \text{for each } f \in C[a,b]. \tag{15}$$

◁

Example 7.8 $(C[a,b], \|\cdot\|_1)$ is a normed vector space, where

$$\|f\|_1 = \int_a^b |f(x)|\,\mathrm{d}x \qquad \text{for each } f \in C[a,b]. \tag{16}$$

◁

7.2 Inner product spaces

This seems to be the proper place to introduce some notation:

Definition 7.2 The ***inner product*** of two vectors $x, y \in \mathbb{R}^n$ is

$$\langle x, y\rangle = x_1y_1 + x_2y_2 + \cdots + x_ny_n = \sum_{i=1}^n x_iy_i.$$

This is also called the dot product of x and y and other common notation for the inner product of x and y includes $x \cdot y$, $(x \mid y)$, and $x^\top y$. The latter makes sense because the inner product of $x, y \in \mathbb{R}^n$ can be interpreted as the matrix product of $1 \times n$ row vector $x^\top$ and $n \times 1$ column vector y.

Example 7.9 Let $p = (p_1, \ldots, p_n)$ denote the vector of unit prices of $n \in \mathbb{N}$ distinct commodities. Let commodity vector $x = (x_1, \ldots, x_n)$ specify for each commodity i the quantity x_i you want to purchase. Since x_i units of commodity i at price p_i cost p_ix_i, this will cost you $\langle p, x\rangle = p_1x_1 + \cdots + p_nx_n$. ◁

The inner product of $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$ with itself is

$$\langle x, x\rangle = x_1x_1 + \cdots + x_nx_n = \sum_{i=1}^n x_i^2.$$

Comparing this with the definition of the Euclidean norm of a vector x in Example 7.2, we find:

$$\|x\|_2 = \sqrt{\langle x, x\rangle} \qquad \text{and} \qquad \langle x, x\rangle = \|x\|_2^2.$$

This link between the inner product and the norm is crucial for proving properties of the Euclidean norm; see Theorem 7.1.

The inner product on $\mathbb{R}^n$ has the following properties. Let $x, y, z \in \mathbb{R}^n$ and $\alpha \in \mathbb{R}$. Then:

- ☒ $\langle x, x\rangle = \sum_{i=1}^n x_i^2$ is the sum of squared real numbers, so $\langle x, x\rangle \geq 0$ with equality if and only if all coordinates equal zero ($x = \mathbf{0}$).
- ☒ $\langle x, y\rangle = \sum_{i=1}^n x_iy_i = \sum_{i=1}^n y_ix_i = \langle y, x\rangle$.
- ☒ $\langle x + y, z\rangle = \sum_{i=1}^n (x_i + y_i)z_i = \sum_{i=1}^n (x_iz_i + y_iz_i) = \sum_{i=1}^n x_iz_i + \sum_{i=1}^n y_iz_i = \langle x, z\rangle + \langle y, z\rangle$.
- ☒ $\langle \alpha x, y\rangle = \sum_{i=1}^n (\alpha x_i)y_i = \alpha \sum_{i=1}^n x_iy_i = \alpha\langle x, y\rangle$.

In general, we can define an inner product on a vector space V to be *any* real-valued function $\langle\cdot,\cdot\rangle$ of two vectors with these four properties:

Definition 7.3 Let V be a real vector space. A function $\langle\cdot,\cdot\rangle$ from $V \times V$ to $\mathbb{R}$ is an ***inner product*** on V if it satisfies

(I1) for all $x \in V$: $\quad \langle x, x\rangle \geq 0$.

(I2) $\langle x, x\rangle = 0$ if and only if $x = \mathbf{0}$.

(I3) symmetry: for all $x, y \in V$: $\quad \langle x, y\rangle = \langle y, x\rangle$.

(I4) linearity in first argument:

- ⊠ for all $x, y, z \in V$: $\quad \langle x + y, z\rangle = \langle x, z\rangle + \langle y, z\rangle$.
- ⊠ for all $x, y \in V$ and all $\alpha \in \mathbb{R}$: $\quad \langle \alpha x, y\rangle = \alpha\langle x, y\rangle$.

The pair $(V, \langle\cdot,\cdot\rangle)$, or just V if the inner product is left implicit, is an ***inner product space***.

Of course, using symmetry, the inner product is linear in its second argument as well.[2]

Example 7.10 On the space $C[a, b]$ of continuous real-valued functions on $[a, b]$ (with $a < b$), define

$$\text{for all } f, g \in C[a, b]: \qquad \langle f, g\rangle = \int_a^b f(x)g(x)\,\mathrm{d}x.$$

This defines an inner product: for (I1) and (I3), note that

$$\langle f, f\rangle = \int_a^b f(x)f(x)\,\mathrm{d}x = \int_a^b f(x)^2\,\mathrm{d}x \geq \int_a^b 0\,\mathrm{d}x = 0,$$

and

$$\langle f, g\rangle = \int_a^b f(x)g(x)\,\mathrm{d}x = \int_a^b g(x)f(x)\,\mathrm{d}x = \langle g, f\rangle.$$

Linearity of the integral implies (I4). For (I2): if $f = \mathbf{0}$, then $\langle f, f\rangle = \int_a^b 0\,\mathrm{d}x = 0$. And if $f \neq \mathbf{0}$, then f^2 is bounded away from zero on a subset of $[a, b]$ by continuity, so $\langle f, f\rangle = \int_a^b f(x)^2\,\mathrm{d}x > 0$. ◁

Our purpose is to prove that

> *any inner product space becomes a normed vector space if we define the norm by* $\|x\| = \sqrt{\langle x, x\rangle}$.

Properties (N1) and (N2) follow trivially from (I1) and (I2): $\|x\| = \sqrt{\langle x, x\rangle}$ is nonnegative by (I1) and by (I2) it is zero if and only if $x = \mathbf{0}$. For property (N3), we find that

$$\|\alpha x\| = \sqrt{\langle \alpha x, \alpha x\rangle} \overset{\text{(I4)}}{=} \sqrt{\alpha\langle x, \alpha x\rangle} \overset{\text{(I3)}}{=} \sqrt{\alpha\langle \alpha x, x\rangle} \overset{\text{(I4)}}{=} \sqrt{\alpha^2\langle x, x\rangle} = |\alpha|\sqrt{\langle x, x\rangle} = |\alpha|\|x\|.$$

The only challenge is to establish the triangle inequality, which we do in Theorem 7.1.

Two vectors $x, y \in V$ are ***orthogonal***, denoted $x \perp y$, if their inner product $\langle x, y\rangle$ is zero. For such x and y,

$$\|x + y\|^2 = \langle x + y, x + y\rangle = \langle x, x\rangle + \underbrace{\langle x, y\rangle + \langle y, x\rangle}_{=0 \text{ by orthogonality}} + \langle y, y\rangle = \|x\|^2 + \|y\|^2,$$

proving ***Pythagoras' Law***:

$$x \perp y \qquad \Longrightarrow \qquad \|x + y\|^2 = \|x\|^2 + \|y\|^2.$$

[2] In complex vector spaces, the inner product $\langle\cdot,\cdot\rangle$ is complex-valued and the symmetry requirement is $\langle x, y\rangle = \overline{\langle y, x\rangle}$, where $\overline{z}$ is the complex conjugate of $z \in \mathbb{C}$.

Theorem 7.1 (Important inequalities in inner product spaces)

In a vector space V with inner product $\langle\cdot,\cdot\rangle$ we have, for all $x, y \in V$:

Cauchy-Schwarz inequality: $|\langle x, y\rangle| \leq \|x\|\|y\|$.

Triangle inequality: $\|x+y\| \leq \|x\| + \|y\|$.

Proof: Both inequalities are true (with equality) if $y = \mathbf{0}$. Let's prove them if $y \neq \mathbf{0}$.

For Cauchy-Schwarz, choose scalar α such that $x - \alpha y$ is orthogonal to y:

$$\langle x - \alpha y, y\rangle = \langle x, y\rangle - \alpha\langle y, y\rangle = 0 \qquad \Longrightarrow \qquad \alpha = \langle x, y\rangle / \langle y, y\rangle.$$

Then

$$0 \leq \|x - \alpha y\|^2 = \langle x - \alpha y, x - \alpha y\rangle = \langle x - \alpha y, x\rangle - \alpha \underbrace{\langle x - \alpha y, y\rangle}_{=0}$$

$$= \langle x, x\rangle - \alpha\langle y, x\rangle = \langle x, x\rangle - \frac{\langle x, y\rangle^2}{\langle y, y\rangle} = \|x\|^2 - \frac{\langle x, y\rangle^2}{\|y\|^2}.$$

Rearranging terms and taking square roots gives $|\langle x, y\rangle| \leq \|x\|\|y\|$.

For the triangle inequality, use Cauchy-Schwarz to conclude that

$$\begin{aligned}\|x+y\|^2 &= \langle x+y, x+y\rangle = \langle x, x\rangle + \langle y, x\rangle + \langle x, y\rangle + \langle y, y\rangle \\ &\leq \langle x, x\rangle + 2|\langle x, y\rangle| + \langle y, y\rangle \leq \|x\|^2 + 2\|x\|\|y\| + \|y\|^2 = (\|x\| + \|y\|)^2.\end{aligned}$$

The triangle inequality follows by taking the square root. □

Exercises section 7

7.1 Show that for all $x \in \mathbb{R}^n$:

(a) $\|x\|_\infty \leq \|x\|_2 \leq \sqrt{n}\|x\|_\infty$,

(b) $\|x\|_\infty \leq \|x\|_1 \leq n\|x\|_\infty$,

(c) $\|x\|_2 \leq \|x\|_1 \leq \sqrt{n}\|x\|_2$.

7.2 (Reverse triangle inequality) Let $(V, \|\cdot\|)$ be a normed vector space. Prove that for all $x, y \in V$:

$$|\|x\| - \|y\|| \leq \|x - y\|.$$

7.3 Verify that all our examples are normed vector spaces; for Example 7.5, consult Exercise 7.6.

7.4 In a (real) vector space with inner product $\langle\cdot,\cdot\rangle$, prove that for all $x, y \in V$:

(a) ***Parallelogram law:*** $\|x+y\|^2 + \|x-y\|^2 = 2\|x\|^2 + 2\|y\|^2$.

(b) ***Polarization identity:*** $\langle x, y\rangle = \frac{1}{4}(\|x+y\|^2 - \|x-y\|^2)$.

HINT: Start with $\|x+y\|^2 = \langle x+y, x+y\rangle$ and expand. Do the same for $\|x-y\|^2$ and add or subtract the resulting equalities.

7.5 Show:

(a) The Cauchy-Schwarz inequality holds with equality if and only if one of x and y is a multiple of the other.

(b) The triangle inequality holds with equality if and only if one of x and y is a nonnegative multiple of the other.

7.6 (Inequalities of Hölder and Minkowski) Let $g : (0,\infty) \to \mathbb{R}$ be concave:

$$\text{for all } x, y \in (0,\infty) \text{ and all } \lambda \in [0,1]: \quad g(\lambda x + (1-\lambda)y) \geq \lambda g(x) + (1-\lambda)g(y)$$

and define $f : (0,\infty) \times (0,\infty) \to \mathbb{R}$ by $f(x,y) = y \cdot g\left(\frac{x}{y}\right)$.

(a) Prove by induction on n that for all $n \in \mathbb{N}$ and all positive real numbers $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$:

$$\sum_{i=1}^{n} f(x_i, y_i) \leq f\left(\sum_{i=1}^{n} x_i, \sum_{i=1}^{n} y_i\right).$$

In the remainder of this exercise, let $p > 1$ and let $q = p/(p-1)$. That is: $\frac{1}{p} + \frac{1}{q} = 1$.

(b) The function $g : (0,\infty) \to \mathbb{R}$ with $g(x) = x^{1/p}$ is concave. Use (a) to prove ***Hölder's inequality:***

$$\text{for all } x, y \in \mathbb{R}^n: \qquad \sum_{i=1}^{n} |x_i|\,|y_i| \leq \|x\|_p \|y\|_q.$$

(c) The function $g : (0,\infty) \to \mathbb{R}$ with $g(x) = (x^{1/p} + 1)^p$ is concave. Use (a) to prove ***Minkowski's inequality***:

$$\text{for all } x, y \in \mathbb{R}^n: \qquad \|x + y\|_p \leq \|x\|_p + \|y\|_p.$$

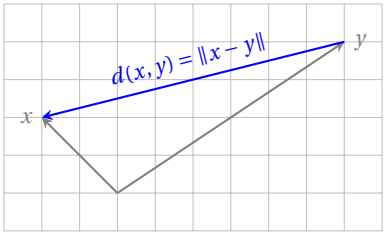
8 Metric spaces

Many crucial concepts in mathematics involve formalizations of intuitively pleasing, but imprecise statements like:

- ☒ a function $f : \mathbb{R}^n \to \mathbb{R}$ is continuous if it doesn't suddenly 'jump' up or down: small changes in its coordinates correspond with small changes in the function value;
- ☒ it is differentiable if, at least locally, it lies near its linear approximation;
- ☒ a sequence of numbers converges to limit L if these numbers eventually get arbitrarily close to L.

Phrases like 'small changes', 'locally', 'lies near', or 'arbitrarily close' can (and will) be made precise as soon as one has a way of measuring distances. In $\mathbb{R}^2$, you know how to relate length and distance:

> *The distance $d(x, y)$ between two points x and y is simply the length $\|x - y\|$ of their difference.*



Consequently, the distance function d inherits a number of desirable properties from the length. These are exactly the properties that characterize a distance function or metric on an arbitrary set:

Definition 8.1 Let X be a nonempty set. A function $d : X \times X \to \mathbb{R}$ is a ***distance function*** or ***metric*** if it satisfies:

(D1) for all $x, y \in X$: $\quad d(x, y) \geq 0$.

(D2) for all $x, y \in X$: $\quad d(x, y) = 0$ if and only if $x = y$.

(D3) symmetry: for all $x, y \in X$: $\quad d(x, y) = d(y, x)$.

(D4) triangle inequality: for all $x, y, z \in X$: $\quad d(x, z) \leq d(x, y) + d(y, z)$.

We call $d(x, y)$ the ***distance*** between x and y. The pair (X, d), or simply X if the metric is left implicit, is a ***metric space***.

The most important special case, which even motivated our definition, is:

Example 8.1 Each normed vector space $(V, \|\cdot\|)$ can be turned into a metric space (V, d) by defining

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

Metric d is the metric ***generated by*** norm $\|\cdot\|$. Hence, each of the normed vector spaces in the previous section generates a metric space. Properties (D1) and (D2) follow trivially from (N1) and (N2), whereas (D3) follows from

$$d(x, y) = \|x - y\| = \|-(y - x)\| \overset{\text{(N3)}}{=} |-1|\|y - x\| = \|y - x\| = d(y, x),$$

and the triangle inequality from

$$d(x, z) = \|x - z\| = \|(x - y) + (y - z)\| \overset{\text{(N4)}}{\leq} \|x - y\| + \|y - z\| = d(x, y) + d(y, z).$$

For instance, the standard Euclidean norm on $\mathbb{R}^n$ generates the metric d_2 with

$$d_2(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

for each pair of vectors $x, y \in \mathbb{R}^n$: its usual Euclidean distance. And the supremum norm on $C[a,b]$ generates the ***supremum metric*** d_∞ with

$$d_\infty(f,g) = \max\{|f(x) - g(x)| : x \in [a,b]\}$$

for each pair of functions $f, g \in C[a,b]$. ◁

So each normed vector space generates a metric space. But metric spaces are more general than that: a metric is defined over an arbitrary nonempty set. It doesn't have to be a vector space! Here is an example of a metric space that is not generated by a norm:

Example 8.2 Let X be an arbitrary nonempty set. Define d by

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

Then (X,d) is a metric space; d is called the ***discrete metric***. Properties (D1) to (D3) are evident. For the triangle inequality, let $x, y, z \in X$. To show:

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

This is true if $d(x,z) = 0$, so assume that $d(x,z) \neq 0$. Then $d(x,z) = 1$ and $x \neq z$. It follows that $x \neq y$ or $y \neq z$, or both. So at least one of the distances on the righthand side of the triangle inequality is one, finishing the proof. Even if X were a vector space, this metric cannot generated by a norm: it would violate the rescaling axiom (N3). ◁

Example 8.3 Let (X,d) be a metric space and Y a nonempty subset of X. If we restrict d to pairs of vectors in Y, then (Y,d) is again a metric space. ◁

Many notions from Euclidean geometry translate straightforwardly to arbitrary metric spaces:

Definition 8.2 Let (X,d) be a metric space, let $x \in X$ and $r > 0$. The ***open ball around*** x ***with radius*** r is the set

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

of points with a distance to x that is less than r.

Of course, the exact geometric shape of such balls depends on the given metric; see Exercises 8.1, 8.3, and 8.4. They play a crucial role in the formalizations of continuity, differentiability, and other notions involving limits, since they help us to define things like 'small changes' using balls with small radii.

Definition 8.3 A subset Y of a metric space (X,d) is ***bounded*** if it is contained in a sufficiently large ball, i.e. if there is an open ball $B(x,r)$ with $Y \subseteq B(x,r)$.

IMPORTANT CONVENTION: Unless explicitly stated otherwise, one commonly uses

- ☒ the Euclidean metric d_2 generated by the Euclidean norm $\|\cdot\|_2$ in $\mathbb{R}^n$,
- ☒ the supremum metric d_∞ generated by the supremum norm in function spaces like $C[a,b]$.

In line with the literature, I will often omit the subscripts in notations like $\|\cdot\|_2$ and d_∞ if this convention applies or if the exact norm or distance is irrelevant.

Exercises section 8

8.1 Draw the open ball $B(\mathbf{0},2)$ in $(\mathbb{R}^2,d)$ for the following metrics d:
(i) d_1, (ii) d_2, (iii) d_∞, (iv) the discrete metric.

8.2 Compute the distance between vectors $x = (1,0,4)$ and $y = (2,6,2)$ for the following metrics on $\mathbb{R}^3$:
(i) d_1, (ii) d_2, (iii) d_∞, (iv) the discrete metric.

8.3 In $C[0,2]$, sketch the open ball $B(f, \frac{1}{2})$ around the function f with $f(x) = x^2$.

8.4 The ***Hamming distance*** d_H on $\mathbb{R}^n$ assigns to each pair of vectors $x, y \in \mathbb{R}^n$ the number of coordinates in which they differ: $d_H(x,y)$ is the number of elements in the set $\{i \in \{1,\ldots,n\} : x_i \neq y_i\}$. This distance is an important measure of the accuracy of data transmission in computer science. Files can be seen as vectors of zeroes and ones. A transmission error means that a coordinate has changed. The Hamming distance measures the number of errors.

(a) What is the Hamming distance between the two vectors in Exercise 8.2?

(b) Prove that the pair $(\mathbb{R}^n, d_H)$ is a metric space.

(c) Prove that d_H is not generated by a norm.

(d) Draw the open ball $B(\mathbf{0}, 2)$ in $(\mathbb{R}^2, d_H)$.

8.5 Nonnegativity (D1) is traditionally included in the definition of a metric, but is redundant in the sense that it is implied by the other three properties. Show this.

8.6 Mathematicians have agreed on using properties (D1) to (D4) to define a distance function. In real-life applications of the word 'distance', these properties are not always satisfied. Can you think of scenarios where precisely one of the properties (D2) to (D4) is violated (but the other three still hold)?

8.7 Let (X,d) be a metric space. Define the functions $d' : X \times X \to \mathbb{R}$ and $d'' : X \times X \to \mathbb{R}$ as follows:

$$\text{for all } x,y \in X: \qquad d'(x,y) = \min\{d(x,y),1\} \quad \text{and} \quad d''(x,y) = \frac{d(x,y)}{d(x,y)+1}.$$

Show that d' and d'' are metrics as well.

8.8 (Reverse triangle inequality) Let (X,d) be a metric space. Prove that for all $x,y,z \in X$: $\quad |d(x,y) - d(y,z)| \leq d(x,z)$.

8.9 We know how to turn a normed vector space into a metric space. This exercise is about the opposite direction. Recall from Example 8.1 that a normed vector space $(V, \|\cdot\|)$ turns into a metric space (V,d) with $d(x,y) = \|x-y\|$. Show that this particular metric satisfies two additional properties:

(a) 'translation invariance': for all $x,y,z \in V$: $\qquad d(x+z, y+z) = d(x,y)$.

(b) 'homogeneity': for all $x,y \in V$ and all scalars $\alpha \in \mathbb{R}$: $\qquad d(\alpha x, \alpha y) = |\alpha| d(x,y)$.

Conversely, if a metric on a vector space satisfies these properties, it induces a norm! Formally, let d be a metric on vector space V that satisfies translation invariance and homogeneity. Define a norm $\|\cdot\|$ on V by $\|x\| = d(x,\mathbf{0})$ for each $x \in V$.

(c) Prove that $\|\cdot\|$ really is a norm.

(d) Prove that $\|x-y\| = d(x,y)$ for all $x,y \in V$.

8.10 (Personalized recommendations in big data analytics) When you buy products online, sites often give recommendations on other items you might like. They compare the set of items you bought/clicked on/searched for with those of other customers, then try to find the nearest such sets, and use this for personalized suggestions. A popular way to measure how similar/nearby two nonempty, finite sets A and B are, is the ***Jaccard distance***

$$d(A,B) = 1 - \frac{|A \cap B|}{|A \cup B|},$$

where notation like $|S|$ denotes the number of elements in a set S.

(a) Suppose your recent book purchases are

{Lila,The blazing world,Infinite jest,Ways of going home,Tourmaline},

and those of three other customers are:

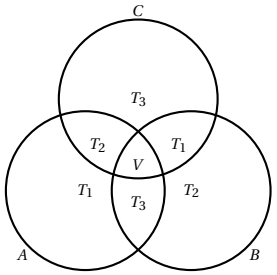
Customer 1:	{`Infinite jest`, `Tourmaline`, `Freshwater`, `Lila`},
Customer 2:	{`The blazing world`, `Ways of going home`, `Ghost wall`, `Elmet`},
Customer 3:	{`Lila`, `Infinite jest`, `The blazing world`, `Elmet`, `Freshwater`}.

Compute your Jaccard distance to each of these three customers. Which is/are nearest?

(b) Let X be a collection of nonempty, finite sets. Show that the Jaccard distance is a metric on X. Only the triangle inequality is tricky, but I'll walk you through it. It requires, for any three sets A, B, and C in X, that

$$d(A,C) \leq d(A,B) + d(B,C).$$

To prove this inequality, look at the three sets A, B, and C in the Venn diagram below, partitioned into four pieces called T_1, T_2, T_3, and V; notice that each T_i in its turn consists of two pieces.



Explain, one line at a time, why the following chain of (in)equalities, proving our result, holds true:

$$\begin{aligned}
d(A,B) + d(B,C) &= \frac{|A \cup B| - |A \cap B|}{|A \cup B|} + \frac{|B \cup C| - |B \cap C|}{|B \cup C|} \\
&\geq \frac{|T_1| + |T_2|}{|A \cup B \cup C|} + \frac{|T_2| + |T_3|}{|A \cup B \cup C|} \\
&\geq \frac{|T_1| + |T_2| + |T_3|}{|A \cup B \cup C|} \\
&= 1 - \frac{|V|}{|A \cup B \cup C|} \\
&\geq 1 - \frac{|A \cap C|}{|A \cup C|} \\
&= d(A,C).
\end{aligned}$$

9 Topology of metric spaces

In this section we introduce notions that will allow us, among other things, to formally define limits in metric spaces.

Definition 9.1 Let (X,d) be a metric space and let $U \subseteq X$.

- ☒ A point $u \in U$ is an ***interior point*** of U if U contains all points sufficiently close to u, i.e., if there is an $\varepsilon > 0$ such that $B(u,\varepsilon) \subseteq U$.
- ☒ Set U is ***open*** if each element of U is an interior point of U.
- ☒ Set U is ***closed*** if its complement $U^c = X \setminus U = \{x \in X : x \notin U\}$ is open.

Example 9.1 In $\mathbb{R}$, the set $U = \{x : -1 < x \leq 2\} = (-1,2]$ is not open: 2 is not an interior point of U. U is not closed either: -1 is not an interior point of U^c. ◁

Example 9.2 If (X,d) is a metric space, the open ball $B(x,r)$ from Definition 8.2 really is an open set: if $y \in B(x,r)$, take $\varepsilon = r - d(x,y) > 0$. We use the triangle inequality to show that $B(y,\varepsilon) \subseteq B(x,r)$, so that y is an interior point of $B(x,r)$. Let $z \in B(y,\varepsilon)$. Then

$$d(z,x) \leq d(z,y) + d(y,x) < \varepsilon + d(y,x) = r.$$

The set $\{y \in X : d(x,y) \leq r\}$ is closed. We show that its complement $\{y \in X : d(x,y) > r\}$ is open. Let $y \in X$ have $d(x,y) > r$. Take $\varepsilon = d(x,y) - r > 0$. By the triangle inequality, each $z \in B(y,\varepsilon)$ has

$$d(x,z) \geq d(x,y) - d(z,y) > d(x,y) - \varepsilon = r.$$ ◁

Example 9.3 If (X,d) is a metric space, each set $\{x\}$ of a single element $x \in X$ is closed, because its complement $\{x\}^c = \{y \in X : y \neq x\}$ is open: if $y \in \{x\}^c$, then $y \neq x$, so $\varepsilon = d(x,y) > 0$. Hence, $B(y,\varepsilon/2) \subseteq \{x\}^c$. ◁

Theorem 9.1

Let (X,d) be a metric space. Its family $\mathcal{O}$ of open sets satisfies:

(a) The empty set $\emptyset$ and the entire space X are open: $\emptyset, X \in \mathcal{O}$.

(b) The union of (arbitrarily many) open sets is an open set.

(c) The intersection of finitely many open sets is an open set.

Proof: **(a)** $\emptyset$ is open: otherwise, it must contain an element that is not an interior point of $\emptyset$. But it contains no points at all. Also X is open: by Definition 8.2, *any* ball $B(x,\varepsilon)$ is a subset of X.
(b) Let I be an arbitrary index set and, for each $i \in I$, let $U_i \subseteq X$ be open. To show: $\cup_{i\in I} U_i$ is an open set.

Let $u \in \cup_{i\in I} U_i$: there is a $j \in I$ with $u \in U_j$. Since U_j is open, there is an $\varepsilon > 0$ with $B(u,\varepsilon) \subseteq U_j \subseteq \cup_{i\in I} U_i$, showing that u is an interior point of $\cup_{i\in I} U_i$.
(c) Let I be a finite index set and, for each $i \in I$, let $U_i \subseteq X$ be open. To show: $\cap_{i\in I} U_i$ is an open set.

Let $u \in \cap_{i\in I} U_i$, i.e., $u \in U_i$ for all $i \in I$. Since U_i is open, $B(u,\varepsilon_i) \subseteq U_i$ for some $\varepsilon_i > 0$. Let $\varepsilon = \min\{\varepsilon_i : i \in I\} > 0$, which is well-defined since I is finite. Then $B(u,\varepsilon) \subseteq B(u,\varepsilon_i) \subseteq U_i$ for all $i \in I$, so $B(u,\varepsilon) \subseteq \cap_{i\in I} U_i$, making u an interior point of $\cap_{i\in I} U_i$. □

Example 9.4 An *arbitrary* union of open sets is open, and a *finite* intersection of open sets is open. The latter cannot be generalized to arbitrary intersections of open sets: for each $k \in \mathbb{N}$, the interval $(-1/k, 1/k)$ is an open subset of $\mathbb{R}$, but their intersection $\cap_{k\in\mathbb{N}} (-1/k, 1/k) = \{0\}$ consists of a single element and is *not* open. ◁

For many mathematical definitions and results involving open sets, the metric is irrelevant: it only matters that open sets have the three properties in the theorem above. The intellectual leap that takes you from metric spaces to topological spaces is simply to insist that these are the properties that characterize open sets:

Definition 9.2 A ***topology*** on a set X is a collection $\mathcal{O}$ of subsets of X that are called ***open*** sets and that satisfy:

- ☒ the empty set and the entire set X are open: $\emptyset \in \mathcal{O}$ and $X \in \mathcal{O}$;
- ☒ the union of arbitrarily many open sets is an open set;
- ☒ the intersection of finitely many open sets is an open set.

The pair $(X, \mathcal{O})$ — or just X if no confusion can arise — is called a ***topological space***.

By Theorem 9.1, metric spaces are topological spaces. But there are many others. Given an arbitrary set X, the ***trivial topology*** $\mathcal{O} = \{\emptyset, X\}$ calls only the empty set and X itself open; in the ***discrete topology*** $\mathcal{O} = P(X)$, the power set X, *all* subsets of X are called open.

Theorem 9.2 (Each open set is the union of open balls)

A subset of a metric space is open if and only if it can be written as a union of open balls.

Proof: If U is an open subset of metric space (X, d), then each element $u \in U$ is an interior point. So there is an $\varepsilon(u) > 0$ with $B(u, \varepsilon(u)) \subseteq U$. Hence

$$U = \bigcup_{u \in U} B(u, \varepsilon(u))$$

is indeed a union of open balls. Conversely, the union of open sets is an open set, so a set that can be written as a union of open balls is open itself. □

Taking complements in Theorem 9.1 and using De Morgan's Laws (Appendix A.2), we find:

Theorem 9.3

Let (X, d) be a metric space. Its collection $\mathcal{C}$ of closed sets satisfies:

(a) The empty set $\emptyset$ and the entire space X are closed.

(b) The intersection of (arbitrarily many) closed sets is closed.

(c) The union of finitely many closed sets is closed.

Example 9.1 and the two theorems above together convey an important message: as opposed to doors, which are either open or closed, in metric spaces there may be sets which are neither open nor closed and sets (like $\emptyset$ and X) which are both open and closed. A very useful way of detecting open and closed sets is discussed in Theorem 11.2.

Call a set U a ***neighborhood*** — often abbreviated ***nbd*** — of a point x if there is an open set O with $x \in O \subseteq U$. This means that x is an interior point of U if and only if U is a neighborhood of x and that U is open if and only if it is a neighborhood of each of its elements. A useful property of metric spaces is the ***Hausdorff property***, which says that each pair of distinct points has 'segregated neighborhoods': if $x \neq y$, there are neighborhoods U_x of x and U_y of y with $U_x \cap U_y = \emptyset$. Indeed, the open balls $U_x = B(x, \varepsilon)$

and $U_y = B(y,\varepsilon)$ with radius $\varepsilon = d(x,y)/2$ have an empty intersection by the triangle inequality: if there were an element $z \in U_x \cap U_y$, then

$$d(x,y) \leq d(x,z) + d(z,y) < \varepsilon + \varepsilon = d(x,y),$$

an obvious contradiction!

Definition 9.3 Let (X,d) be a metric space and let $U \subseteq X$.

- ⊠ The ***interior*** of U, denoted $\text{int}(U)$, is the largest open set contained in U. It is the union of all open sets contained in U:
$$\text{int}(U) = \cup_{W \subseteq U:\, W \text{ is open}} W. \tag{17}$$
- ⊠ The ***closure*** of U, denoted $\text{cl}(U)$, is the smallest closed set containing U. It is the intersection of all closed sets containing U:
$$\text{cl}(U) = \cap_{W \supseteq U:\, W \text{ is closed}} W. \tag{18}$$
- ⊠ The ***boundary*** of U, denoted $\text{bd}(U)$, is the set
$$\text{bd}(U) = \text{cl}(U) \cap \text{cl}(U^c). \tag{19}$$
As the intersection of two closed sets, it is a closed set. Elements of $\text{bd}(U)$ are called ***boundary points***.
- ⊠ A point $x \in X$ is an ***accumulation point*** of U if each neighborhood of x contains an element of U *other than* x: for each $\varepsilon > 0$, there is a point $u \neq x$ in U with $d(x,u) < \varepsilon$. Equivalently,
$$\text{for each } \varepsilon > 0: \quad (B(x,\varepsilon) \setminus \{x\}) \cap U \neq \emptyset. \tag{20}$$
The ***set of accumulation points*** of U is denoted $\text{acc}(U)$.
- ⊠ A point $x \in X$ is an ***isolated point*** of U if it is the only element of U in a sufficiently small neighborhood: there is an $\varepsilon > 0$ such that $B(x,\varepsilon) \cap U = \{x\}$.

An accumulation point x of U need not be an element of U, but making ε smaller and smaller, at least we can approximate it by elements $u \neq x$ of U to arbitrary precision. This also motivates the terminology: points of U accumulate or gather around x.

Example 9.5 Let (X,d) be a metric space and $x \in X$ an accumulation point of a subset U of X. Then it gets pretty crowded around x: each open ball $B(x,\varepsilon)$ contains *infinitely many* distinct points of U. Suppose, to the contrary, that a ball $B(x,\varepsilon)$ contains only finitely many distinct points of U, say $\{u_1,\ldots,u_m\}$. By (20), all these points are different from x, so their distance $d(u_i,x)$ to x is larger than zero. Take $\varepsilon' = \min\{d(u_1,x),\ldots,d(u_m,x)\} > 0$ to be the smallest distance. Since x is an accumulation point of U, the ball $B(x,\varepsilon') \subseteq B(x,\varepsilon)$ contains an element $u \in U$. By construction, this cannot be a point in $\{u_1,\ldots,u_m\}$. ◁

Formally, the interior as the largest open set contained in U or the union of all open sets in U is perfectly well-defined. In practice, however, it would be nice to have a more manageable definition in terms of neighborhoods or balls. As the name suggests, the interior of U is simply the set of all interior points of U. We can also find more manageable characterizations of the closure $\text{cl}(U)$, the set of points around which each open ball contains at least one element of U, and the boundary $\text{bd}(U)$ of U, the set of points around which each open ball contains at least one element of U *and* at least one element not in U, i.e., in its complement U^c. Analogously, you're on the boundary between two countries if — even if you walk around only a small distance — you will encounter points in both countries. Many textbooks in mathematics for economists use this characterization of the boundary as the definition.

Theorem 9.4

Let (X,d) be a metric space and U a subset of X.

(a) $\mathrm{int}(U)$ is the set of interior points of U:

$$\mathrm{int}(U) = \{u \in U \mid \text{there is an } \varepsilon > 0 \text{ with } B(u,\varepsilon) \subseteq U\};$$

(b) $\mathrm{cl}(U) = \{x \in X \mid \text{for each } \varepsilon > 0 : B(x,\varepsilon) \cap U \neq \emptyset\}$;

(c) $\mathrm{cl}(U) = U \cup \mathrm{bd}(U)$;

(d) $\mathrm{cl}(U) = U \cup \mathrm{acc}(U)$;

(e) $\mathrm{bd}(U) = \{x \in X \mid \text{for each } \varepsilon > 0 : B(x,\varepsilon) \cap U \neq \emptyset \text{ and } B(x,\varepsilon) \cap U^c \neq \emptyset\}$.

Proof: **(a)** $\subseteq$: By (17), the interior of U is the union of open sets, hence open itself: each element of $\mathrm{int}(U)$ is an interior point of $\mathrm{int}(U)$ and consequently of U.
$\supseteq$: If u is an interior point of U, there is an $\varepsilon > 0$ with $B(u,\varepsilon) \subseteq U$. Since $B(u,\varepsilon)$ is *an* open set contained in U (see Example 9.2) and $\mathrm{int}(U)$ is the *largest* open set contained in U:

$$u \in B(u,\varepsilon) \subseteq \mathrm{int}(U).$$

(b) $\subseteq$: Let $x \in \mathrm{cl}(U)$ and $\varepsilon > 0$. Suppose that $B(x,\varepsilon) \cap U = \emptyset$. The set $W = X \setminus B(x,\varepsilon)$ is closed, $U \subseteq W$, and $x \notin W$. By (18), $\mathrm{cl}(U) \subseteq W$. Since $x \in \mathrm{cl}(U)$, but $x \notin W$, we have a contradiction.
$\supseteq$: Let $x \in X$ be such that

$$\text{for each } \varepsilon > 0 : \quad B(x,\varepsilon) \cap U \neq \emptyset. \tag{21}$$

Suppose that $x \notin \mathrm{cl}(U)$. Since $\mathrm{cl}(U)$ is a closed set, its complement is open. This complement contains x, so there is a $\varepsilon > 0$ with

$$B(x,\varepsilon) \subseteq X \setminus \mathrm{cl}(U) \overset{(18)}{\subseteq} X \setminus U,$$

contradicting (21).
(c) $\subseteq$: Let $x \in \mathrm{cl}(U)$. If $x \in U$, we are done. So suppose $x \notin U$. Then

$$x \in \mathrm{cl}(U) \cap U^c \overset{(18)}{\subseteq} \mathrm{cl}(U) \cap \mathrm{cl}(U^c) \overset{(19)}{=} \mathrm{bd}(U).$$

$\supseteq$: By definition, $U \overset{(18)}{\subseteq} \mathrm{cl}(U)$ and $\mathrm{bd}(U) \overset{(19)}{=} \mathrm{cl}(U) \cap \mathrm{cl}(U^c) \subseteq \mathrm{cl}(U)$, so $U \cup \mathrm{bd}(U) \subseteq \mathrm{cl}(U)$.
(d) $\subseteq$: Let $x \in \mathrm{cl}(U)$. If $x \in U$, we are done. So suppose $x \notin U$. By (b), for each $\varepsilon > 0$:

$$(B(x,\varepsilon) \setminus \{x\}) \cap U \overset{x \notin U}{=} B(x,\varepsilon) \cap U \overset{(b)}{\neq} \emptyset, \qquad \text{so } x \in \mathrm{acc}(U).$$

$\supseteq$: By definition, $\mathrm{acc}(U) \overset{(b)}{\subseteq} \mathrm{cl}(U)$ and $U \overset{(18)}{\subseteq} \mathrm{cl}(U)$, so $U \cup \mathrm{acc}(U) \subseteq \mathrm{cl}(U)$.
(e) Follows from (b) and (19). □

Theorem 9.4 implies a number of useful characterizations of closed sets; its proof is Exercise 9.10.

Theorem 9.5 (Characterizations of closed sets)

Let (X,d) be a metric space and U a subset of X. The following statements are equivalent:

(a) U is closed;

(b) $U = \mathrm{cl}(U)$;

(c) U contains all its boundary points: $\mathrm{bd}(U) \subseteq U$;

(d) U contains all its accumulation points: $\mathrm{acc}(U) \subseteq U$.

Example 9.6 The following is true for the indicated sets U in $\mathbb{R}$:

U	$\mathrm{int}(U)$	$\mathrm{cl}(U)$	$\mathrm{bd}(U)$	$\mathrm{acc}(U)$	isolated
$\{1/n : n \in \mathbb{N}\}$	$\emptyset$	$U \cup \{0\}$	$U \cup \{0\}$	$\{0\}$	U
$(0,1]$	$(0,1)$	$[0,1]$	$\{0,1\}$	$[0,1]$	$\emptyset$
$\mathbb{Q}$	$\emptyset$	$\mathbb{R}$	$\mathbb{R}$	$\mathbb{R}$	$\emptyset$
$\mathbb{N}$	$\emptyset$	$\mathbb{N}$	$\mathbb{N}$	$\emptyset$	$\mathbb{N}$
$\{0\} \cup (1,2]$	$(1,2)$	$\{0\} \cup [1,2]$	$\{0,1,2\}$	$[1,2]$	$\{0\}$

Definition 9.4 A subset Y of a metric space (X,d) is ***dense*** in X if each nonempty, open set contains an element of Y.

This concept is useful for approximating elements of X by those in a smaller set Y: if Y is dense, there is always an element of Y arbitrarily close to any point $x \in X$, because the open ball around x with radius $\varepsilon > 0$ must contain an element of Y, no matter how small ε is. Making ε ever smaller and picking corresponding elements from $Y \cap B(x,\varepsilon)$, we can construct a sequence[3] in Y that converges to x.

Example 9.7 The set $\mathbb{Z}$ of integers is not dense in $\mathbb{R}$: the interval $\left(\frac{1}{4}, \frac{3}{4}\right)$ is nonempty and open but does not contain an integer. In contrast, the set $\mathbb{Q}$ of rational numbers is dense in $\mathbb{R}$: each real number like

$$\pi = 3.1415926535\cdots$$

can be approximated arbitrarily well by a rational number by adding decimal places one at a time:

$$3.1 = \frac{31}{10}, \quad 3.14 = \frac{314}{100}, \quad 3.141 = \frac{3141}{1000}, \quad \ldots$$

So each neighborhood of a real number contains a rational number. Likewise, the set $\mathbb{Q}^n$ of vectors with rational coordinates is dense in $\mathbb{R}^n$. ◁

A major application of dense sets for nonparametric regression models in economics and statistics will be formulated later in Theorem 18.1, the Stone-Weierstrass approximation theorem. Roughly speaking, it provides conditions under which even rather capricious functions can be closely approximated by polynomial ones.

[3] If you don't know yet what a convergent sequence is, don't worry, we'll get to that in Section 13.

Theorem 9.6

Let Y be a subset of metric space (X,d). The following are equivalent:

(a) Y is dense in X.

(b) For each $x \in X$ and each $\varepsilon > 0$ there is a point $y \in Y$ with $d(x,y) < \varepsilon$.

(c) For each $x \in X$ there is a sequence $(y_k)_{k\in\mathbb{N}}$ of points in Y converging to x.

(d) $\mathrm{cl}(Y) = X$.

Proof: We discussed implications (a) $\Longrightarrow$ (b) $\Longrightarrow$ (c) already after the definition of a dense set.

(c) $\Longrightarrow$ (d): Assume (c) is true. If $\mathrm{cl}(Y)$ does not equal X, then there is an $x \in X$ that does not belong to the closed set $\mathrm{cl}(Y)$. Since x lies in the open set $X \setminus \mathrm{cl}(Y)$, there is an entire ball around x, say with radius $\varepsilon > 0$, that lies in $X \setminus \mathrm{cl}(Y)$: there are no points in $\mathrm{cl}(Y)$, let alone in the smaller set Y, within distance ε from x. But then we cannot find a sequence in Y converging to x, contradicting (c).

(d) $\Longrightarrow$ (a): Assume (d) is true. Let U be nonempty and open. We need to argue that $U \cap Y \neq \emptyset$.

By definition of a metric space, X is nonempty. Since $\mathrm{cl}(\emptyset) = \emptyset \neq X$, it follows that Y is nonempty as well. If it were the case that $U \cap Y = \emptyset$, then $Y \subseteq U^c \subset X$. Taking closures, it follows that $\mathrm{cl}(Y) \subseteq \mathrm{cl}(U^c) = U^c \subset X$, where the equality follows from U^c being closed. This contradicts that $\mathrm{cl}(Y) = X$. $\square$

Exercises section 9

9.1 Determine the interior, closure, boundary, and set of accumulation points of the following subsets of $(\mathbb{R}^2, d_2)$:

(a) $\{x : x_1 > 0\}$ (b) $\{x : x_1^2 + x_2^2 = 4\}$

(c) $\{x : x_1 \leq x_2\}$ (d) $\{x : x_1 > 0, x_2 = \sin \frac{1}{x_1}\}$

(e) $\{x : x_1 x_2 \in \mathbb{Q}\}$

9.2 In $(C[0,1], d_\infty)$, what are the isolated points and the accumulation points of the set $\{f_n : n \in \mathbb{N}\}$, where:

(a) $f_n(x) = x^n$ for all $x \in [0,1]$.

(b) $f_n(x) = nx$ for all $x \in [0,1]$.

(c) $f_n(x) = x/n$ for all $x \in [0,1]$.

9.3 Let (V,d) be a metric space and $U \subseteq V$. Prove:

(a) $\mathrm{bd}(U) = \mathrm{cl}(U) \setminus \mathrm{int}(U)$.

(b) $\mathrm{acc}(U)$ is closed.

9.4 Show that finite subsets of metric spaces are closed.

9.5 Prove Theorem 9.3.

9.6 Show that the closure satisfies the following properties. Properties (a), (b), (c), and (e) are called ***Kuratowski closure axioms***.

(a) $\mathrm{cl}(\emptyset) = \emptyset$,

(b) $U \subseteq \mathrm{cl}(U)$,

(c) $\mathrm{cl}(\mathrm{cl}(U)) = \mathrm{cl}(U)$,

(d) If $A \subseteq B$, then $\mathrm{cl}(A) \subseteq \mathrm{cl}(B)$.

(e) $\mathrm{cl}(U \cup V) = \mathrm{cl}(U) \cup \mathrm{cl}(V)$.

9.7 Show that the interior satisfies the following properties.

(a) $\mathrm{int}(\emptyset) = \emptyset$.

(b) $\mathrm{int}(U) \subseteq U$.

(c) $\mathrm{int}(\mathrm{int}(U)) = \mathrm{int}(U)$.

(d) If $A \subseteq B$, then $\mathrm{int}(A) \subseteq \mathrm{int}(B)$.

(e) $\mathrm{int}(U \cap V) = \mathrm{int}(U) \cap \mathrm{int}(V)$.

9.8 Give examples of sets in $\mathbb{R}$ for which the following equations are false:

(a) $\mathrm{cl}(U \cap V) = \mathrm{cl}(U) \cap \mathrm{cl}(V)$.

(b) $\mathrm{cl}(U^c) = \mathrm{cl}(U)^c$.

(c) $\mathrm{int}(U \cup V) = \mathrm{int}(U) \cup \mathrm{int}(V)$.

(d) $\mathrm{int}(U^c) = \mathrm{int}(U)^c$.

9.9 Show that the closure and interior are dual in the sense that in every topological space:

(a) 'the complement of the closure is the interior of the complement': $\mathrm{cl}(U)^c = \mathrm{int}(U^c)$.

(b) 'the complement of the interior is the closure of the complement': $\mathrm{int}(U)^c = \mathrm{cl}(U^c)$.

9.10 Prove Theorem 9.5.

9.11 In $(C[0,1], d_\infty)$, consider the sets of functions

$$U = \{f : f(0) = 0\} \quad \text{and} \quad V = \{f : f \text{ is constant}\} \quad \text{and} \quad W = \{f : f(x) < 1 \text{ for all } x \in [0,1]\}.$$

For each of these three sets, is it open? Is it closed?

10 Metric subspaces and their open sets

Example 10.1 In a metric space (X,d) the metric tells us the distance between any two elements of X. In particular, it tells us the distance between points in any subset Y of X: we can create a *smaller* metric space which we with a slight abuse of notation will denote by (Y,d) by simply restricting the metric d on the entire $X \times X$ to the smaller domain $Y \times Y$. This (Y,d) is called a ***(metric) subspace*** of (X,d) and the collection of sets that are open in Y is called the ***subspace*** (or ***relative***) ***topology***. ◁

Of course, a metric subspace is not to be confused with a linear subspace of a vector space. Mathematics, just like ordinary languages, sometimes uses the same word for different things. It tends to be clear from the context which one is meant.

Metric subspaces arise naturally in economics when some elements of a larger metric space are not relevant to the analysis: commodity vectors are often modeled as the subset $Y = \mathbb{R}^n_+$ of nonnegative vectors in $X = \mathbb{R}^n$ with its usual distance and probability vectors (like mixed strategies in a finite game or a chance move) correspond with the subset Y of nonnegative vectors whose coordinates sum to one.

Moving to a metric subspace subtly changes which sets are open because whether something is an interior point depends on what metric space you live in: in (X,d) the open ball around a point v with radius $\varepsilon > 0$ consists of all points in X whose distance to v is less than ε, but in (Y,d) it consists of all points in Y whose distance to v is less than ε.

So which sets are open in a metric subspace? Fortunately, this is easy (Exercise 10.1): if (Y,d) is a metric subspace of (X,d), a subset U of Y is open if and only if it is of the form $U = Y \cap O$, where O is an open subset of the larger metric space (X,d). In words, the open subsets of Y are the open subsets of X intersected with Y. For instance, in $X = \mathbb{R}^2$ with its usual distance, the sets

$$U = \mathbb{R}^2_+ = \{x \in \mathbb{R}^2 : x_1, x_2 \geq 0\} \qquad \text{and} \qquad U' = \{x \in \mathbb{R}^2_+ : x_1 > x_2\}$$

are *not* open because elements like $(1,0)$ are not interior points. (You should draw their pictures!) But they *are* open in the metric subspace with $Y = \mathbb{R}^2_+$, because they are of the required form $Y \cap O$ for a suitable open set O in X:

$$U = Y \cap \mathbb{R}^2 \qquad \text{and} \qquad U' = Y \cap \{x \in \mathbb{R}^2 : x_1 > x_2\}$$

and the sets $\mathbb{R}^2$ and $\{x \in \mathbb{R}^2 : x_1 > x_2\}$ are open in $\mathbb{R}^2$.

Exercises section 10

10.1 Let Y be a subset of metric space (X,d).

(a) Show that an open ball in the metric subspace (Y,d) is the intersection of Y and an open ball in (X,d).

(b) Show that a subset U of Y is open in the metric subspace (Y,d) if and only if $U = Y \cap O$, where O is an open subset of the larger metric space (X,d). HINT: Use Theorem 9.2 to write an open set as a union of open balls. Combine this with the insight in the first part of this exercise.

10.2 In this exercise we practice with some metric subspaces of $X = \mathbb{R}^2$ with its usual distance.

(a) Draw the open ball around $(1,1)$ with radius 2, around $(4,4)$ with radius 1, and around $(6,0)$ with radius 1.

(b) Draw the same balls, but now in the metric subspace $Y = \mathbb{R}^2_+$.

(c) Give an example of a set that is open in Y, but not in X. Give an example of a set that is open in both Y and X. Give an example of a set that is neither open in Y nor in X.

(d) You can't find a subset of Y that is open in X but not in Y. Why?

(e) Draw a straight line in $\mathbb{R}^2$. Can you find a metric subspace where this line is an open set?

11 Continuous functions

11.1 The definition of continuity

Continuous functions have no sudden jumps in their function values. Look at a point a in the domain with function value $b = f(a)$. Suppose I challenge you to find points near a whose function value jumps outside a neighborhood of $f(a)$. Continuity says that you will fail at this task: as long as you stay sufficiently close to a, the function values will lie in the desired neighborhood.

Definition 11.1 Let (X,d) and (Y,d') be metric spaces, U a subset of X, and $f : U \to Y$ a function.

- ☒ Let $a \in U$ be a point in its domain and $b = f(a)$ its function value. Function f is ***continuous at*** a if for each neighborhood B of b there is a neighborhood A of a with $f(x) \in B$ for all $x \in A \cap U$.
- ☒ Function f is ***continuous*** if it is continuous at each point in its domain.

This definition easily extends to more general topological spaces. For functions between metric spaces it is enough to look only at *some* neighborhoods: open balls. This is the so-called (ε,δ)***-definition*** of continuity that you may have seen in undergraduate courses:

Theorem 11.1 (Local continuity via the (ε,δ)-definition)

Let (X,d) and (Y,d') be metric spaces, $f : U \to Y$ a function on a subset U of X, and let $a \in U$. The following are equivalent:

(a) f is continuous at a;

(b) (ε,δ)-definition of continuity at a: for each $\varepsilon > 0$ there is a $\delta > 0$ such that

$$\text{each } x \in U \text{ with } d(x,a) < \delta \text{ has } d'(f(x),f(a)) < \varepsilon. \tag{22}$$

Proof: **(a)** $\Longrightarrow$ **(b)** Assume f is continuous at a. For any $\varepsilon > 0$, the open ball $B(f(a),\varepsilon)$ is a neighborhood of $f(a)$, so by continuity there is a neighborhood A of a with $f(x) \in B(f(a),\varepsilon)$ for all $x \in A \cap U$. Since a is an interior point of this neighborhood, there is a $\delta > 0$ with $B(a,\delta) \subseteq A$. Hence, (22) holds.
(b) $\Longrightarrow$ **(a)** Assume (b) holds. Let B be a neighborhood of $f(a)$. Since $f(a)$ is an interior point of B, there is an $\varepsilon > 0$ with $B(f(a),\varepsilon) \subseteq B$. For this $\varepsilon > 0$ there is a $\delta > 0$ for which (22) holds. Take $A = B(a,\delta)$. Then each $x \in A \cap U$ has $f(x) \in B(f(a),\varepsilon) \subseteq B$. □

Consider a function $f : X \to Y$. The ***pre-image*** $f^{-1}(V)$ of a set $V \subseteq Y$ consists of all points in the domain $x \in X$ with a function value $f(x)$ in V, i.e., all points that are mapped into V:

$$f^{-1}(V) = \{x \in X : f(x) \in V\}.$$

Example 11.1 Consider $f : \mathbb{R} \to \mathbb{R}$ with $f(x) = x^2$. Then

$$\begin{aligned} f^{-1}(\{4\}) &= \{x \in \mathbb{R} : f(x) \in \{4\}\} = \{x \in \mathbb{R} : x^2 = 4\} = \{-2,2\}, \\ f^{-1}((-\infty,9]) &= \{x \in \mathbb{R} : f(x) \in (-\infty,9]\} = \{x \in \mathbb{R} : x^2 \le 9\} = [-3,3], \\ f^{-1}((1,16]) &= \{x \in \mathbb{R} : f(x) \in (1,16]\} = \{x \in \mathbb{R} : 1 < x^2 \le 16\} = [-4,-1) \cup (1,4]. \end{aligned}$$ ◁

In terms of pre-images, the (ε,δ)-definition of continuity at a becomes:

$$\text{for each } \varepsilon > 0, \text{ there is a } \delta > 0 \text{ with } B(a,\delta) \subseteq f^{-1}\big(B(f(a),\varepsilon)\big). \tag{23}$$

A useful result is the following characterization of continuous functions in terms of pre-images:

Theorem 11.2 (Continuity and pre-images)

Let (X,d) and (Y,d') be metric spaces and $f: X \to Y$. The following three claims are equivalent:

(a) Function f is continuous;

(b) Pre-images of open sets are open sets: if $V \subseteq Y$ is open, then $f^{-1}(V)$ is open;

(c) Pre-images of closed sets are closed sets: if $V \subseteq Y$ is closed, then $f^{-1}(V)$ is closed.

Proof: **(a) ⇒ (b)** Assume f is continuous. Let $V \subseteq Y$ be open. To show that $f^{-1}(V)$ is open, we show that each $a \in f^{-1}(V)$ is an interior point. So let $a \in f^{-1}(V)$. Then $b = f(a)$ lies in the open set V, so V is a neighborhood of b. By continuity there is a neighborhood A of a with $f(x) \in V$ for all $x \in A$. So $a \in A \subseteq f^{-1}(V)$. Since a is an interior point of A, it is also an interior point of the larger set $f^{-1}(V)$.
(b) ⇒ (a) Assume that pre-images of open sets are open sets. To show that f is continuous at each $a \in X$, let $a \in X$ and $\varepsilon > 0$. Since $B(f(a),\varepsilon)$ is open, so is its pre-image $f^{-1}(B(f(a),\varepsilon))$. Moreover, it contains a. Hence, there is a $\delta > 0$ such that $B(a,\delta) \subseteq f^{-1}(B(f(a),\varepsilon))$, finishing the proof.
(b) ⇒ (c) Let $V \subseteq Y$ be closed. Then V^c is open, so its pre-image $f^{-1}(V^c)$ is open as well. So

$$f^{-1}(V) = \{x \in X : f(x) \in V\} = \{x \in X : f(x) \notin V\}^c = \{x \in X : f(x) \in V^c\}^c = f^{-1}(V^c)^c,$$

its complement, is closed. The proof that (c) implies (b) is similar. □

The next result is about the composition of two functions. The composition is what you get from plugging one function $f: X \to Y$ into another function $g: Y \to Z$. For each $x \in X$, you can compute $f(x) \in Y$. And since g is defined on Y, you can plug $f(x)$ into g and compute $g(f(x)) \in Z$. Formally:

Definition 11.2 Consider two functions $f: X \to Y$ and $g: Y \to Z$. The ***composition*** $(g \circ f): X \to Z$ is the function defined by

$$(g \circ f)(x) = g(f(x)) \qquad \text{for each } x \in X.$$

The expression $(g \circ f)$ is often pronounced as 'g after f'.

Be aware that typically 'g after f' and 'f after g' are different functions.

Example 11.2 Given $f: \mathbb{R} \to \mathbb{R}$ defined by $f(x) = x^2$ and $g: \mathbb{R} \to \mathbb{R}$ defined by $g(x) = x+1$, we have

$$(g \circ f)(x) = g(f(x)) = f(x) + 1 = x^2 + 1,$$

but

$$(f \circ g)(x) = f(g(x)) = (g(x))^2 = (x+1)^2 = x^2 + 2x + 1.$$ ◁

The composition of continuous functions is continuous:

Theorem 11.3 (Continuity of composition)

Let X, Y, Z be metric spaces. If $f: X \to Y$ is continuous at x and $g: Y \to Z$ is continuous at $y = f(x)$, then $g \circ f: X \to Z$ is continuous at x.

Proof: Define $z = g(y) = g(f(x)) = (g \circ f)(x)$. Let W be a neighborhood of z. By continuity of g at y there is a neighborhood V of y with $g(y') \in W$ for all $y' \in V$. By continuity of f at x there is a neighborhood U of x with $f(x') \in V$ and consequently $g(f(x')) = (g \circ f)(x') \in W$ for all $x' \in U$. So $g \circ f$ is continuous at x. □

11.2 Examples of continuous functions: working with the (ε,δ)-definition

The (ε,δ)-definition of continuity in (22) can be used to establish the continuity of reasonably elementary functions. The examples below are *not* chosen because they are particularly easy, but because they are the most important building blocks of more general continuous functions: they are about some of the most crucial algebraic operations like linearity, multiplication, and division.

Example 11.3 An ***affine function*** from $\mathbb{R}^n$ to $\mathbb{R}^m$ (with the usual Euclidean norm) is a function $f:\mathbb{R}^n\to\mathbb{R}^m$ of the form $f(x)=Ax+b$ for some $m\times n$ matrix A and vector $b\in\mathbb{R}^m$. An affine function is ***linear*** if $b=\mathbf{0}$. Each affine function $f:\mathbb{R}^n\to\mathbb{R}^m$ is continuous.

- ⊠ If $A=\mathbf{0}$, the zero matrix, this is trivial: the constant function $f(x)=b$ is continuous: $\|f(x)-f(a)\|=\|b-b\|=0$ is always smaller than $\varepsilon>0$, regardless of x and a.
- ⊠ If $A\neq\mathbf{0}$, denote its columns by $a^1,\dots,a^n$. By the triangle inequality we have, for all $x,y\in\mathbb{R}^n$,

$$\begin{aligned}\|f(x)-f(y)\|=\|Ax-Ay\|=\|A(x-y)\|&=\left\|\sum_{i=1}^n(x_i-y_i)\,a^i\right\|\\&\le\sum_{i=1}^n|x_i-y_i|\,\|a^i\|\le\sum_{i=1}^n\|x-y\|\,\|a^i\|=\left(\sum_{i=1}^n\|a^i\|\right)\|x-y\|.\end{aligned}$$

So for each $\varepsilon>0$ we can choose $\delta=\varepsilon/\left(\sum_{i=1}^n\|a^i\|\right)$. Then $\|x-y\|<\delta$ implies $\|f(x)-f(y)\|<\varepsilon$. ◁

Example 11.4 The function $f:\mathbb{R}^2\to\mathbb{R}$ with $f(x_1,x_2)=x_1x_2$ is continuous. To see that it is continuous at $a=(a_1,a_2)$, let $\varepsilon>0$. Choose $\delta=\min\{1,\varepsilon/(|a_1|+|a_2|+1)\}$. For each $x\in\mathbb{R}^2$ with $\|x-a\|<\delta$ it follows, using the triangle inequality, that

$$|x_2|=|x_2-a_2+a_2|\le|x_2-a_2|+|a_2|\le\|x-a\|+|a_2|<1+|a_2|. \tag{24}$$

and consequently that

$$\begin{aligned}|f(x)-f(a)| &= |x_1x_2-a_1a_2|=|x_1x_2\underbrace{-a_1x_2+a_1x_2}_{=0}-a_1a_2|\\ &\le |x_1x_2-a_1x_2|+|a_1x_2-a_1a_2|=|x_2|\,|x_1-a_1|+|a_1|\,|x_2-a_2|\\ &\overset{(24)}{\le} (|a_2|+1)\underbrace{|x_1-a_1|}_{\le\|x-a\|}+|a_1|\underbrace{|x_2-a_2|}_{\le\|x-a\|}\le(|a_1|+|a_2|+1)\|x-a\|\\ &< (|a_1|+|a_2|+1)\delta\le\varepsilon.\end{aligned}$$

◁

Example 11.5 The function $f:\mathbb{R}\setminus\{0\}\to\mathbb{R}$ with $f(x)=1/x$ is continuous: let $a\neq0$ and $\varepsilon>0$. Choose $\delta=\min\{\frac{\varepsilon|a|^2}{2},\frac{|a|}{2}\}$. For each $x\neq0$ with $|x-a|<\delta$, the triangle inequality gives

$$|x|\ge|a|-|a-x|>|a|-\frac{|a|}{2}=\frac{|a|}{2} \tag{25}$$

and consequently that

$$\left|f(x)-f(a)\right|=\left|\frac{1}{x}-\frac{1}{a}\right|=\left|\frac{a-x}{ax}\right|=\frac{|a-x|}{|a|\,|x|}\overset{(25)}{<}\frac{\delta}{|a|\,\frac12|a|}=\frac{2\delta}{|a|^2}\le\varepsilon.$$

◁

Example 11.6 (The coordinate criterion for continuity) Let (X,d) be a metric space. A function $f: X \to \mathbb{R}^n$ with $f(x) = (f_1(x),\dots,f_n(x))$ is continuous if and only if each of its coordinate functions $f_i : X \to \mathbb{R}$ (with $i = 1,\dots,n$) is continuous. Indeed, let $a \in X$. For each $i = 1,\dots,n$:

$$|f_i(x) - f_i(a)| \le \|f(x) - f(a)\| \le \sum_{j=1}^{n} |f_j(x) - f_j(a)|.$$

If we can make $\|f(x) - f(a)\|$ arbitrarily small, the first inequality shows that we can do the same for the i-th coordinate function f_i. Conversely, if we can make each term $|f_j(x) - f_j(a)|$ arbitrarily small, the second inequality shows that we can do the same for $\|f(x) - f(a)\|$. ◁

Similarly, although the proofs are tedious, it can be show that functions on $\mathbb{R}$ like

$$x \mapsto e^x, \qquad x \mapsto \sin x,$$

and functions on $(0,\infty)$ like

$$x \mapsto x^{1/p}\ (p > 0), \qquad x \mapsto \ln x$$

are continuous. These proofs are omitted.

Having established the continuity of certain affine/linear functions and of addition and division (Examples 11.3, 11.4, and 11.5) and recalling that the composition of continuous functions is continuous (Theorem 11.3), it follows that continuity is remarkably robust to all kinds of algebraic manipulations:

> *Any function that can be constructed from continuous functions using addition, subtraction, multiplication, division, and composition, is continuous (whenever defined).*

Example 11.7 As the composition of continuous functions, we know that $f : \mathbb{R}^3 \to \mathbb{R}^2$ with

$$f(x_1,x_2,x_3) = \left(e^{x_1^4} + \sin(x_2 x_3), \frac{37}{\sqrt{x_3^2+6}} \right)$$

is continuous. I'm pretty sure you don't want to prove this using the (ε,δ)-definition! ◁

In particular:

Theorem 11.4 (Arithmetic rules for continuous functions)

Let $f : X \to \mathbb{R}$ and $g : X \to \mathbb{R}$ be real-valued functions on a metric space (X,d). Assume that f and g are continuous at some point a in their domain. Then:

(a) For each real number c, the rescaled function cf is continuous at a;

(b) The sum function $f+g$ is continuous at a;

(c) The product function $f \cdot g$ (often simply denoted fg) is continuous at a;

(d) The quotient function f/g is continuous at a (provided $g(x) \neq 0$ for each $x \in X$).

11.3 Uniform and Lipschitz continuity

There are two common, more restrictive notions of continuity:

Definition 11.3 Let (X,d) and (Y,d') be metric spaces, $U \subseteq X$, and let $f : U \to Y$. Function f is:

- ☒ ***uniformly continuous*** if for each $\varepsilon > 0$ there is a $\delta > 0$ such that

$$\text{all } x, y \in U \text{ with } d(x,y) < \delta \text{ have } d'(f(x), f(y)) < \varepsilon.$$

- ☒ ***Lipschitz continuous*** if there is a real number $M \geq 0$, called a ***Lipschitz constant***, such that

$$\text{for all } x, y \in U: \qquad d'(f(x), f(y)) \leq M d(x,y).$$

In the (ε,δ)-definition of continuity at a point a, the chosen δ is allowed to depend *both* on a and on ε. For uniform continuity, the chosen δ is allowed to depend *only* on ε, which is more restrictive. So each uniformly continuous function is continuous. And, choosing $M > 0$ and $\delta = \varepsilon/M$, every Lipschitz continuous function is uniformly continuous. The reverse implications do not hold:

Example 11.8 The function $f : \mathbb{R} \to \mathbb{R}$ with $f(x) = x^2$ is continuous but not uniformly continuous. To see the latter, take $\varepsilon = 1$. Suppose there were a $\delta > 0$ such that if $|x-y| < \delta$, then $|f(x) - f(y)| < 1$. Take $x = \frac{1}{\delta}$ and $y = \frac{1}{\delta} + \frac{1}{2}\delta$. Then $|x-y| = \frac{1}{2}\delta < \delta$, but $|f(x) - f(y)| = |-1 - \frac{1}{4}\delta^2| > 1$, a contradiction. ◁

Example 11.9 The function $f : \mathbb{R}_+ \to \mathbb{R}$ with $f(x) = \sqrt{x}$ is uniformly continuous, but not Lipschitz continuous. For uniform continuity, let $\varepsilon > 0$. Let $\delta = \varepsilon^2$. If $|x-y| < \delta$, then $|f(x) - f(y)| = |\sqrt{x} - \sqrt{y}| \leq \sqrt{|x-y|} < \sqrt{\delta} = \varepsilon$. And f is not Lipschitz continuous: if it were, there'd be an M such that

$$\frac{|\sqrt{x} - \sqrt{y}|}{|x-y|} \leq M.$$

for all $x, y \geq 0$ with $x \neq y$. But

$$\frac{|\sqrt{x} - \sqrt{y}|}{|x-y|} = \frac{|\sqrt{x} - \sqrt{y}|}{|(\sqrt{x} - \sqrt{y})(\sqrt{x} + \sqrt{y})|} = \frac{1}{\sqrt{x} + \sqrt{y}}$$

is larger than M for x, y sufficiently close to zero. ◁

Exercises section 11

11.1 Make the proof sketch in Example 11.6 precise.

11.2 Let $X = \{(x_1, x_2) \in \mathbb{R}^2 : x_2 \neq 0\}$ and define the function $f : X \to \mathbb{R}$ for each $(x_1, x_2) \in X$ by $f(x_1, x_2) = x_1/x_2$. Prove that f is continuous. HINT: Use that the composition of continuous functions is again continuous. *Don't* do an (ε,δ)-proof; that is unnecessarily difficult and inelegant.

11.3 Prove Theorem 11.4.

11.4 Show that the function $f : (0,1) \to \mathbb{R}$ with $f(x) = x^2$ is uniformly continuous. (Hm... didn't we show something entirely different in Example 11.8? Sure, but the function has a different domain now!)

12 The product topology

Suppose that each person i in a set I of agents makes a choice x_i from a set X_i. This profile of choices, one for each agent, is an element $x = (x_i)_{i\in I}$ of the product set $X = \times_{i\in I} X_i$. To talk about things like continuity of payoff/utility functions on such a product space X, we need to define which of its subsets are open. And it would be nice if the open sets of X follow in some easy way from the open sets of its respective coordinate sets X_i. So throughout this section we assume that each X_i is a metric/topological space. On their product $X = \times_{i\in I} X_i$, we will define the so-called *product topology* which establishes a link between the open sets in X and the open sets in the coordinate sets X_i by including as open sets in X precisely those sets that are needed to make the coordinate projections continuous.

For instance, we know which sets are open in $\mathbb{R}$ with its usual Euclidean distance. The product topology on $\mathbb{R}\times\mathbb{R}$ makes sure that the two coordinate projections

$$\text{proj}_1(x_1, x_2) = x_1 \qquad \text{and} \qquad \text{proj}_2(x_1, x_2) = x_2$$

are continuous.

In general, given the product set $X = \times_{i\in I} X_i$ and some j in I, the j-th ***coordinate projection*** $\text{proj}_j : X \to X_j$ assigns to each element $x = (x_i)_{i\in I}$ of the product set its j-th coordinate:

$$\text{proj}_j(x) = x_j.$$

Which subsets of our product set must be open to make each coordinate projection continuous?

Firstly, continuity says that pre-images of open sets are open sets. So for each $j \in I$ and each open set O in X_j, its pre-image

$$\text{proj}_j^{-1}(O) = \{x \in X : \text{proj}_j(x) \in O\} = \{x \in X : x_j \in O\} \tag{26}$$

must be open: sets where *one* coordinate is restricted to lie in some open set O and all other coordinates are unrestricted must be open. So in $\mathbb{R}\times\mathbb{R}$, sets $A = (0,1)\times\mathbb{R}$ and $B = \mathbb{R}\times(7,\infty)$ need to be open according to (26): they restrict one coordinate to an open set $(0,1)$ or $(7,\infty)$, leaving the other coordinate free.

Secondly, by definition of a topological space also the intersection of finitely many such open sets must be open. These are sets where *finitely many* coordinates must lie in some specific open set, whereas for all other indices $j \in I$, x_j can be chosen without restrictions from X_j. We call such sets 'cylinder sets'. Recalling that the topological space X_j itself is open, they can be defined as follows:

Definition 12.1 A ***cylinder set*** in $\times_{i\in I} X_i$ is a product set $\times_{i\in I} Y_i$ where

- ☒ each Y_i is open in X_i and
- ☒ there are only finitely many indices $j \in I$ with $Y_j \neq X_j$.

We often look at products of only finitely many sets (the strategy space of finitely many players, the consumption bundles of finitely many consumers, ...). In this case, there are only finitely many indices/coordinates, so the second requirement on a cylinder set is automatically true and a cylinder set is just a product of open sets. For instance, the intersection $A \cap B = (0,1)\times(7,\infty)$ of our earlier two sets is a cylinder set, just like A and B themselves.

Thirdly, and again by definition of a topological space, the union of (arbitrarily many) open sets must be open. Hence, the union of cylinder sets must be open.

To summarize: if we insist that coordinate projections are continuous, then all unions of cylinder sets must be open sets. In Exercise 12.1 you are asked to verify that these particular open sets indeed satisfy the properties we want of a topology (Definition 9.2). It is therefore the smallest collection of open sets that makes coordinate projections continuous: our desired product topology.

Definition 12.2 For each i in an index set I, let X_i be a metric or topological space. A subset of the product set $X = \times_{i\in I} X_i$ is ***open in the product topology*** if it can be written as a union of cylinder sets.

CONVENTION: This is the most common way of defining open sets in product spaces. In line with much of the mathematical literature, we assume that product sets are endowed with the product topology, unless explicitly stated otherwise.

The following result says that in an interior point of a set Z in the product $X \times Y$ of two sets, you can slightly change both of its coordinates and still remain inside Z.

Theorem 12.1

Consider two topological spaces X and Y and their product $X \times Y$ with the product topology. Given a subset Z of $X \times Y$ and an element (x, y) of Z, the following are equivalent:

(a) (x, y) is an interior point of Z;

(b) there are open neighborhoods V of x and W of y with $V \times W \subseteq Z$.

Proof: If (x, y) is an interior point of Z, it has an open neighborhood that lies entirely inside Z. By definition of the product topology, this open neighborhood is the union of cylinder sets. Hence (x, y) lies in one of these cylinder sets. In $X \times Y$, a cylinder set is the product $V \times W$ of an open set V in X and an open set W in Y. So $(x, y) \in V \times W \subseteq Z$.

Conversely, if there are open neighborhoods V of x and W of y with $V \times W \subseteq Z$, then $V \times W$ is a cylinder set in $X \times Y$ and consequently open: (x, y) has an open neighborhood $V \times W$ inside Z, making it an interior point of Z. □

For products of Euclidean spaces, this looks much harder than it is. Consider $\mathbb{R}^m$ and $\mathbb{R}^n$ with their usual Euclidean distance. The product set

$$\mathbb{R}^m \times \mathbb{R}^n = \{(x, y) : x \in \mathbb{R}^m, y \in \mathbb{R}^n\}$$

has elements of the form

$$(x, y) = ((x_1, \ldots, x_m), (y_1, \ldots, y_n)) \tag{27}$$

with two vectors $x = (x_1, \ldots, x_m)$ in $\mathbb{R}^m$ and $y = (y_1, \ldots, y_n)$ in $\mathbb{R}^n$ next to each other. Without spurious parentheses, this is just a vector

$$(x_1, \ldots, x_m, y_1, \ldots, y_n) \tag{28}$$

of $m + n$ coordinates, i.e., a vector in $\mathbb{R}^{m+n}$. Formally, we can identify $\mathbb{R}^m \times \mathbb{R}^n$ with $\mathbb{R}^{m+n}$ through the natural bijection that maps (27) to (28). More importantly, their open sets are the same:

Theorem 12.2

$\mathbb{R}^m \times \mathbb{R}^n$ with the product topology and $\mathbb{R}^{m+n}$ with its Euclidean distance have the same open sets.

So in practice we tend to treat $\mathbb{R}^m \times \mathbb{R}^n$ and $\mathbb{R}^{m+n}$ as the same space: here we won't need to think about the product topology at all. Applying this repeatedly, we can identify the open sets in $\mathbb{R}^m \times \mathbb{R}^n \times \mathbb{R}^k$ with those in $\mathbb{R}^{m+n+k}$, etc.

Proof (Thm. 12.2): First, let Z be an open subset of $\mathbb{R}^m \times \mathbb{R}^n$ with the product topology. To show: Z is open in $\mathbb{R}^{m+n}$.

Let $z \in Z$. We will argue that it is an interior point of Z in $\mathbb{R}^{m+n}$. Write $z = (x, y)$, where $x = (z_1, \ldots, z_m) \in \mathbb{R}^m$ gathers the first m coordinates of z and $y = (z_{m+1}, \ldots, z_{m+n}) \in \mathbb{R}^n$ its final n coordinates. Since Z is open in $\mathbb{R}^m \times \mathbb{R}^n$ with the product topology, (x, y) is an interior point. By Theorem 12.1, there are open neighborhoods V of x in $\mathbb{R}^m$ and W of y in $\mathbb{R}^n$ with $V \times W \subseteq Z$. In particular, there are $\varepsilon, \varepsilon' > 0$ such that $B_{\mathbb{R}^m}(x, \varepsilon) \subseteq V$ and $B_{\mathbb{R}^n}(y, \varepsilon') \subseteq W$; I use subscripts to remind you in which Euclidean spaces these balls lie.

Take $\delta = \min\{\varepsilon, \varepsilon'\} > 0$. I will show that $B_{\mathbb{R}^{m+n}}(z, \delta) \subseteq Z$, making z an interior point of Z in $\mathbb{R}^{m+n}$. Let $z' \in B_{\mathbb{R}^{m+n}}(z, \delta)$ and, like $z = (x, y)$, write it as $z' = (x', y')$ with $x' \in \mathbb{R}^m$ and $y' \in \mathbb{R}^n$. Then

$$d_{\mathbb{R}^m}(x, x') = \sqrt{\sum_{i=1}^{m} (x_i - x_i')^2} \leq \sqrt{\sum_{i=1}^{m} (x_i - x_i')^2 + \sum_{j=1}^{n} (y_j - y_j')^2} = d_{\mathbb{R}^{m+n}}(z, z') < \delta \leq \varepsilon,$$

so $x' \in B_{\mathbb{R}^m}(x, \varepsilon) \subseteq V$. Likewise, $y' \in B_{\mathbb{R}^n}(y, \varepsilon') \subseteq W$. Hence,

$$z' = (x', y') \in V \times W \subseteq Z,$$

as we needed to show.

Conversely, let Z be open in $\mathbb{R}^{m+n}$ with its Euclidean topology. To show: Z is open in $\mathbb{R}^m \times \mathbb{R}^n$ with the product topology.

Let $z \in Z$. We use Theorem 12.1 to show that z is also an interior point of Z in the product topology. Since Z is open in the Euclidean topology, there is an $\varepsilon > 0$ with $B_{\mathbb{R}^{m+n}}(z, \varepsilon) \subseteq Z$. As above, write $z = (x, y)$ with $x \in \mathbb{R}^m$ and $y \in \mathbb{R}^n$. Then

$$B_{\mathbb{R}^m}\big(x, \varepsilon/\sqrt{2}\big) \times B_{\mathbb{R}^n}\big(y, \varepsilon/\sqrt{2}\big) \subseteq B_{\mathbb{R}^{m+n}}((x, y), \varepsilon) = B_{\mathbb{R}^{m+n}}(z, \varepsilon) \subseteq Z.$$

Letting the two balls in $\mathbb{R}^m$ and $\mathbb{R}^n$ play the roles of V and W in Theorem 12.1, it follows that z is an interior point of Z in the product topology. □

The theorem above presents the most useful case where the open sets in the product topology (unions of cylinder sets) are simply the open sets in a familiar metric space. Here is a more general result along the same lines:

Theorem 12.3 (The product topology is metrizable)

If (X, d_X) and (Y, d_Y) are metric spaces, then the function d with

$$d\big((x, y), (x', y')\big) = \max\big\{d_X(x, x'), d_Y(y, y')\big\} \qquad \text{for all } (x, y) \text{ and } (x', y') \text{ in } X \times Y \tag{29}$$

is a metric on $X \times Y$. Moreover, the sets that are open in the metric space $(X \times Y, d)$ are exactly the ones that are open in the product topology on $X \times Y$.

The crucial insight for the proof is that the ball around a point $(x, y) \in X \times Y$ with radius $\varepsilon > 0$ is just the product of the corresponding balls in X and Y, which is a cylinder set:

$$B((x, y), \varepsilon) = B_X(x, \varepsilon) \times B_Y(y, \varepsilon).$$

Indeed, by definition of the ***product metric*** in (29), (x', y') lies in the ball around (x, y) with radius ε if and only if the maximal distance between x and x' and between y and y' is less than ε. Equivalently, both distances are less than ε: x' lies in the ε-ball around x and y' lies in the ε-ball around y.

This allows us to translate back and forth between cylinder sets in the product topology and balls in the metric spaces; the rest of the proof is bookkeeping.

Proof (Thm. 12.3): The proof that d is a metric on $X \times Y$ is, with minor changes in notation, the same as that for the supremum metric d_∞ on $\mathbb{R}^n$ and therefore omitted. It remains to show that an open set in the metric space $(X \times Y, d)$ is open in the product topology and vice versa.

Let U be open in $(X \times Y, d)$: for each $(x, y) \in U$ there is an $\varepsilon > 0$ with

$$B((x, y), \varepsilon) = B_X(x, \varepsilon) \times B_Y(y, \varepsilon) \subseteq U.$$

The two balls around x and y are open neighborhoods of these two points, so each $(x, y) \in U$ is an interior point of U by Theorem 12.1. Conclude: U is open in the product topology.

Next, let U be open in the product topology on $X \times Y$: it is the union of cylinder sets. It therefore suffices to prove that each cylinder set is open in the metric space $(X \times Y, d)$. Such a cylinder set is of the form $V \times W$ for some open subsets V of X and W of Y.

So for each $(x, y) \in V \times W$, x is an interior point of V and y is an interior point of W: there are $\varepsilon_1, \varepsilon_2 > 0$ with $B_X(x, \varepsilon_1) \subseteq V$ and $B_Y(y, \varepsilon_2) \subseteq W$. Take $\varepsilon = \min\{\varepsilon_1, \varepsilon_2\}$. It follows that

$$B((x, y), \varepsilon) = B_X(x, \varepsilon) \times B_Y(y, \varepsilon) \subseteq V \times W.$$

So each $(x, y) \in V \times W$ is an interior point of $V \times W$ in our metric space: the cylinder set $V \times W$ is open in $(X \times Y, d)$. □

Exercises section 12

12.1 Show, by verifying the properties in Definition 9.2, that the collection of sets that are unions of cylinder sets is a topology.

13 The limit of a sequence and the limit of a function

13.1 The limit of a sequence

Next, we consider the limit of sequences. We defined sequences of real numbers in Example 3.5. Let us extend this to sequences in an arbitrary metric space.

Definition 13.1 Let (X,d) be a metric space.

- ☒ A ***sequence*** $(x_k)_{k\in\mathbb{N}} = (x_1, x_2, x_3, \ldots)$ in X assigns to each $k \in \mathbb{N}$ an element $x_k \in X$: it is a function from $\mathbb{N}$ to X, just denoted in a slightly more convenient way.
- ☒ We call x_k the k-th ***term*** of the sequence. The integer k is sometimes referred to as its ***label***.
- ☒ Consider a strictly increasing sequence of positive integers $k(1) < k(2) < k(3) < \cdots$. The sequence $(x_{k(1)}, x_{k(2)}, x_{k(3)}, \ldots) = (x_{k(n)})_{n\in\mathbb{N}}$ obtained from sequence $(x_k)_{k\in\mathbb{N}}$ is a ***subsequence*** of $(x_k)_{k\in\mathbb{N}}$.

Example 13.1 If $(x_1, x_2, x_3, \ldots)$ is a sequence, then x_{37} is the 37-th term of the sequence; it has label 37. The following are examples of subsequences:

- ☒ the even-numbered terms $(x_2, x_4, x_6, \ldots) = (x_{2n})_{n\in\mathbb{N}}$;
- ☒ the terms after the 27-th: $(x_{28}, x_{29}, x_{30}, \ldots) = (x_{27+n})_{n\in\mathbb{N}}$;
- ☒ the terms with labels $2^1, 2^2, 2^3, \ldots$: $(x_2, x_4, x_8, x_{16}, \ldots) = (x_{2^n})_{n\in\mathbb{N}}$.

The following are *not* subsequences:

- ☒ (x_1, x_2, x_3): subsequences have infinitely many terms.
- ☒ $(x_2, x_1, x_3, x_4, x_5, \ldots)$: subsequences have terms with increasing labels. ◁

Definition 13.2 Let (X,d) be a metric space, $(x_k)_{k\in\mathbb{N}}$ a sequence in X, and $x \in X$. Sequence $(x_k)_{k\in\mathbb{N}}$ ***converges to*** x (or has ***limit*** x), denoted $\lim_{k\to\infty} x_k = x$ or $x_k \to x$, if

$$\text{for each } \varepsilon > 0 \text{ there is a number } N \text{ such that for each integer } k \geq N \text{ we have } d(x_k, x) < \varepsilon.$$

Such a sequence is called ***convergent***.

Theorem 13.1 (Uniqueness of the limit)

In a metric space, a sequence can have at most one limit.

Proof: Suppose sequence $(x_k)_{k\in\mathbb{N}}$ in metric space (X,d) has distinct limits x and x'. Let $\varepsilon = d(x,x')/2 > 0$. By convergence, there are N and N' such that

$$\text{for each integer } k \geq N\text{:}\quad d(x_k, x) < \varepsilon \qquad \text{and} \qquad \text{for each integer } k \geq N'\text{:}\quad d(x_k, x') < \varepsilon.$$

So if $k \geq \max\{N, N'\}$, then $d(x_k, x) < \varepsilon$ and $d(x_k, x') < \varepsilon$. The triangle inequality gives a contradiction:

$$d(x,x') = 2\varepsilon > d(x,x_k) + d(x_k,x') \geq d(x,x').$$ □

Some relatively routine examples of convergent sequences of real numbers can be found in Exercise 13.1; its solution in the appendix is very elaborate and explains possible strategies for solving this type of problems. Here are a few more:

Example 13.2 In $(\mathbb{R}^2, d_2)$, the sequence $(x_k)_{k\in\mathbb{N}}$ with $x_k = (2^{-k}, 1 - 1/k)$ converges to $(0,1)$: let $\varepsilon > 0$. Choose $N \in \mathbb{N}$ with $N > \frac{2}{\varepsilon}$. Then each $k \in \mathbb{N}$ with $k \geq N$ has

$$\|x_k - (0,1)\| = \|(2^{-k}, -1/k)\| \leq \frac{1}{2^k} + \frac{1}{k} < \frac{2}{k} < \varepsilon.$$

The first inequality uses the triangle inequality and the second inequality uses that $k < 2^k$. ◁

Example 13.3 In $(C[0,1], d_\infty)$, the sequence $(f_k)_{k\in\mathbb{N}}$ with $f_k(x) = x/k$ converges to the zero function: let $\varepsilon > 0$. Choose $N \in \mathbb{N}$ with $N > \frac{1}{\varepsilon}$. Then each $k \in \mathbb{N}$ with $k \geq N$ has:

$$d_\infty(f_k, \mathbf{0}) = \sup_{x\in[0,1]} |x/k - 0| = \frac{1}{k} < \varepsilon.$$

◁

Example 13.4 In $(C[0,1], d_\infty)$, the sequence $(f_k)_{k\in\mathbb{N}}$ with $f_k(x) = x^k$ does *not* converge. If $x \in [0,1)$, then $x^k \to 0$ as $k \to \infty$, so the candidate limit function f must satisfy $f(x) = 0$ for all $x \in [0,1)$. If $x = 1$, however, then $x^k = 1$ for all k, so the candidate limit function f must satisfy $f(1) = 1$. Hence, the candidate limit function cannot be continuous at $x = 1$: the sequence does not converge in $C[0,1]$. ◁

For sequences of real numbers, the following turns out to be useful:

Definition 13.3 A sequence $(x_k)_{k\in\mathbb{N}}$ of real numbers is ***monotonic*** if it is:

(weakly) increasing: $x_1 \leq x_2 \leq x_3 \leq \cdots$ or (weakly) decreasing: $x_1 \geq x_2 \geq x_3 \geq \cdots$

Theorem 13.2 (Bolzano-Weierstrass)

(a) Each monotonic, bounded sequence in $\mathbb{R}$ is convergent.

(b) Each sequence in $\mathbb{R}$ has a monotonic subsequence.

(c) Each bounded sequence in $\mathbb{R}$ has a convergent subsequence.

(d) Each bounded sequence in $\mathbb{R}^n$ has a convergent subsequence.

Proof: **(a)** Let $(x_k)_{k\in\mathbb{N}}$ be a monotonic, bounded sequence in $\mathbb{R}$. Assume it is weakly increasing (if it is weakly decreasing, change sup to inf and reason accordingly).

The set $\{x_1, x_2, x_3, \ldots\}$ is nonempty and bounded from above, so it has a *smallest* upper bound L, its supremum. The sequence converges to L.

Formally, let $\varepsilon > 0$. Since L is the smallest upper bound of $\{x_1, x_2, x_3, \ldots\}$, $L - \varepsilon$ is not an upper bound: there is an $N \in \mathbb{N}$ with $x_N > L - \varepsilon$. And the sequence is weakly increasing, so $L - \varepsilon < x_k \leq L$ for all $k \geq N$. In particular, $|x_k - L| < \varepsilon$ for all $k \geq N$.

(b) Let $(x_k)_{k\in\mathbb{N}}$ be a sequence in $\mathbb{R}$. If it has a weakly increasing subsequence, we are done. Now suppose there is no such subsequence: if we start from any term and look for later ones that become (weakly) larger, this search eventually fails. It fails at a term that is larger than all later ones. Call such a term a 'cliff'; graphically, it is a point after which the sequence drops down and remains forever lower. This shows that after any term of the sequence there is a cliff. So there is a subsequence of cliffs, where the sequence drops down: a decreasing subsequence.

(c) Each bounded sequence in $\mathbb{R}$ has a monotonic subsequence by (b). This subsequence is monotonic and bounded, hence convergent by (a).

(d) If $(x_k)_{k\in\mathbb{N}}$ is a bounded sequence in $\mathbb{R}^n$ and $i \in \{1, \ldots, n\}$ one of the n coordinates, then the sequence of i-th coordinates is bounded as well. By (c), $(x_k)_{k\in\mathbb{N}}$ has a subsequence for which the first coordinates converge. From this subsequence, we can take a subsequence for which also the second coordinates converge. Repeating the process n times gives a subsequence all of whose coordinates converge. □

Definition 13.4 (Divergence to $+\infty$ or $-\infty$) A sequence $(x_n)_{n\in\mathbb{N}}$ is said to

- ⊠ ***diverge to*** $+\infty$, denoted $\lim_{n\to\infty} x_n = +\infty$, if for each $r \in \mathbb{R}$, there is an $N \in \mathbb{N}$ such that $n \geq N$ implies $x_n \geq r$;
- ⊠ ***diverge to*** $-\infty$, denoted $\lim_{n\to\infty} x_n = -\infty$, if for each $r \in \mathbb{R}$, there is an $N \in \mathbb{N}$ such that $n \geq N$ implies $x_n \leq r$.

Informally, if $(x_n)_{n\in\mathbb{N}}$ diverges to $+\infty$, then all terms of the sequence eventually exceed r, no matter how large r is.

Theorem 13.3 (Continuity and sequences)

Function $f : X \to Y$ between metric spaces X and Y is continuous at $x \in X$ if and only if for each sequence $(x_k)_{k\in\mathbb{N}}$ in X:

$$\text{if } x_k \to x, \text{ then } f(x_k) \to f(x).$$

Proof: Assume f is continuous at $x \in X$ and let $(x_k)_{k\in\mathbb{N}}$ converge to x. To show: $f(x_k) \to f(x)$.

Let $\varepsilon > 0$. By continuity of f at x, there is a $\delta > 0$ with $B(x,\delta) \subseteq f^{-1}\big(B(f(x),\varepsilon)\big)$. Since $x_k \to x$, there is an $N \in \mathbb{N}$ such that $x_k \in B(x,\delta)$ for all $k \geq N$. Consequently, $f(x_k) \in B(f(x),\varepsilon)$ for all such k, proving that $f(x_k) \to f(x)$.

Conversely, assume that $f(x_k) \to f(x)$ whenever $x_k \to x$. If f is not continuous at x, there is a $\varepsilon > 0$ such that $B(x,\delta) \not\subseteq f^{-1}\big(B(f(x),\varepsilon)\big)$ for all $\delta > 0$. Hence, we can construct a sequence $x_k \in B(x,1/k)$ such that $f(x_k) \notin B(f(x),\varepsilon)$. Thus, $x_k \to x$, but $f(x_k) \not\to f(x)$, a contradiction. □

The next result gives a connection between closed sets and convergent sequences within such sets:

Theorem 13.4

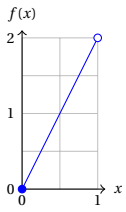
A subset U of a metric space X is closed if and only if for each convergent sequence in U, the limit belongs to U.

Proof: First, assume U is closed. Let $(x_n)_{n\in\mathbb{N}}$ be a convergent sequence in U with limit x. Suppose, to the contrary, that $x \in U^c$. Since U is closed, its complement U^c is open, so x is an interior point: there is an $\varepsilon > 0$ with $B(x,\varepsilon) \subseteq U^c$. But $x_n \to x$ means that terms x_n with n sufficiently large belong to this ball $B(x,\varepsilon)$ and consequently to U^c, contradicting that the sequence lies in U.

Conversely, assume that each convergent sequence in U has a limit in U. Suppose, however, that U is not closed: its complement U^c is not open. Hence, some $x \in U^c$ is not an interior point of U^c: for each $\varepsilon > 0$, $B(x,\varepsilon) \cap U \neq \emptyset$. In particular, for each $n \in \mathbb{N}$, there is an $x_n \in B(x,\frac{1}{n}) \cap U$. Sequence $(x_n)_{n\in\mathbb{N}}$ lies in U, but its limit x does not, a contradiction! □

13.2 The limit of a function

I only briefly discuss the notion of a limit of a function: it plays a very minor role in these notes. Before stating the precise definition, we look at an example that illustrates some of the desiderata. Define the function $f : [0,1) \to \mathbb{R}$ with $f(x) = 2x$. What we have in mind is this:



> *We want notation* $\lim_{x\to 1} f(x) = 2$ *to mean that* $f(x)$ *can be made arbitrarily close to* 2 *by selecting points* x *in the domain of* f *sufficiently close to, but distinct from* 1.

In particular, even though 1 is not in the domain of function f, there are points in the domain arbitrarily close to 1: we need the point in which we evaluate the limit to be an accumulation point of the domain. Moreover, we choose points really close to, but distinct from 1. We have to do this, because the function value is not defined in 1. Here is the official definition:

Definition 13.5 Let (X,d) and (Y,d') be metric spaces, let $U \subseteq X$, let $f : U \to Y$, let $a \in X$ be an accumulation point of U, and let $L \in Y$. We say that f ***has limit*** L ***in*** a or ***converges/goes to*** L ***as*** x ***goes to*** a, denoted $\lim_{x\to a} f(x) = L$, if for each $\varepsilon > 0$ there is a $\delta > 0$ such that

$$\text{each } x \in U \text{ with } 0 < d(x,a) < \delta \text{ has } d'(f(x),L) < \varepsilon. \tag{30}$$

Notice the parts of the definition: we require

- ☒ a to be an accumulation point of the domain to make sure that we can approach a via points in the domain of the function;
- ☒ $0 < d(x,a) < \delta$ to assure that we have points sufficiently close to but distinct from a;
- ☒ $d'(f(x),L) < \varepsilon$ to force function values to lie close to L;
- ☒ limit L to lie in metric space Y, into which the function f maps.

The final property seems obvious in our example: of course, the limit of a real-valued function must be a real number. But in other spaces it is less obvious: it might very well be that function values get close to some element L, but that L happens to lie *outside* Y, in which case the function does not converge!

As in Theorem 13.1, if a limit exists, it must be unique by the Hausdorff property of metric space Y.

If you compare Definition 13.5 with the (ε,δ)-definition of continuity at point a (Thm. 11.1), this continuity looks a lot like saying that $\lim_{x\to a} f(x) = f(a)$. This is made precise in the following theorem.

Theorem 13.5

Let (X,d) and (Y,d') be metric spaces, let $U \subseteq X$, $a \in U$, and let $f : U \to Y$. The following two claims are equivalent:

(a) f is continuous at a,

(b) a is an isolated point of U or a is an accumulation point of U and $\lim_{x\to a} f(x) = f(a)$.

Proof: **(a) ⇒ (b)** Assume that f is continuous in a and that a is not an isolated point of U. Then a is an accumulation point of U. Definition 13.5 immediately implies that $\lim_{x\to a} f(x) = f(a)$.
(b) ⇒ (a) If a is an isolated point of U, there is a $\delta > 0$ such that $B(a,\delta) \cap U = \{a\}$. But then, for each $\varepsilon > 0$: if $v \in U$ has $d(a,v) < \delta$, then $v = a$ and consequently $d'(f(v),f(a)) = 0 < \varepsilon$.

If a is an accumulation point of U and $\lim_{x\to a} f(x) = f(a)$, then (30) implies that (22) holds as long as $x \neq a$. It holds trivially for $x = a$. □

Exercises section 13

13.1 Show that the following sequences in $\mathbb{R}$ converge and determine their limit.

(a) $\left(\frac{2k-3}{k+1}\right)_{k\in\mathbb{N}}$

(b) $\left(\frac{2k}{3k^2+4}\right)_{k\in\mathbb{N}}$

(c) $\left(\frac{(-1)^k}{3k-1}\right)_{k\in\mathbb{N}}$

13.2 (Coordinate criterion for convergence) Prove that sequence $(x_k)_{k\in\mathbb{N}}$ in $(\mathbb{R}^n,d_2)$ converges to $x \in \mathbb{R}^n$ if and only if for each $i \in \{1,\ldots,n\}$ the sequence $(x_{ki})_{k\in\mathbb{N}}$ of its i-th coordinates converges to $x_i \in \mathbb{R}$.

13.3 Consider a set X with the discrete metric. Which sequences are convergent?

13.4 The following functions are defined on $\mathbb{R}^2 \setminus \{\mathbf{0}\}$. Do they have a limit $\lim_{x \to \mathbf{0}} f(x)$ as x goes to $\mathbf{0}$?

(a) $f(x_1, x_2) = \frac{x_1 x_2}{x_1^2 + x_2^2}$

(b) $f(x_1, x_2) = \frac{x_1^2 + x_2^2}{|x_1 + x_2| + |x_1 x_2|}$

(c) $f(x_1, x_2) = \frac{\sin(x_1 x_2)}{\sqrt{x_1^2 + x_2^2}}$ HINT: $|\sin(y)| \leq |y|$ for all $y \in \mathbb{R}$.

(d) $f(x_1, x_2) = \frac{x_1^4 + x_2^4}{x_1^2 + x_2^2}$

(e) $f(x_1, x_2) = \frac{x_1^2 + 3x_1^2 x_2 + x_2^2}{x_1^2 + x_2^2}$

(f) $f(x_1, x_2) = \frac{4x_1 x_2}{\sqrt{x_1^2 + x_2^2}}$

(g) $f(x_1, x_2) = \frac{x_1^2}{x_1^2 + x_2^2}$

14 Completeness

If a sequence $(x_n)_{n\in\mathbb{N}}$ in a metric space (X,d) converges to $x\in X$, a tail of the sequence eventually lies arbitrarily close to x. In particular, such elements will also lie close to each other: by convergence, for each $\varepsilon>0$ there in an $N\in\mathbb{N}$ such that if $n\geq N$, then $d(x_n,x)<\frac{1}{2}\varepsilon$. Using the triangle inequality, it follows that

$$\text{for all } m,n\geq N: \qquad d(x_m,x_n)\leq d(x_m,x)+d(x,x_n)<\frac{1}{2}\varepsilon+\frac{1}{2}\varepsilon=\varepsilon. \tag{31}$$

Sequences whose elements eventually cluster together will play an important role in our approach to dynamic optimization and a number of results that prepare us for that. We give them a special name:

Definition 14.1 Let (X,d) be a metric space. A sequence $(x_n)_{n\in\mathbb{N}}$ is a ***Cauchy sequence*** in X if

$$\text{for each } \varepsilon>0 \text{ there is an } N\in\mathbb{N} \text{ such that if } m,n\geq N, \text{ then } d(x_m,x_n)<\varepsilon.$$

By (31), every convergent sequence in a metric space is a Cauchy sequence. The converse is false:

Example 14.1 Let $X=(0,1)$. The sequence $\left(\frac{1}{2},\frac{1}{3},\frac{1}{4},\frac{1}{5},\ldots\right)$ is a Cauchy sequence in $(X,|\cdot|)$: let $\varepsilon>0$. Fix $N\in\mathbb{N}$ with $N>\frac{2}{\varepsilon}$. For all $m,n\geq N$, the triangle inequality gives:

$$d\left(\frac{1}{m},\frac{1}{n}\right)=\left|\frac{1}{m}-\frac{1}{n}\right|\leq\left|\frac{1}{m}\right|+\left|\frac{1}{n}\right|<\frac{\varepsilon}{2}+\frac{\varepsilon}{2}=\varepsilon.$$

But it has no limit in X, since $0\notin X$. ◁

Example 14.2 Let d_1 be the following metric on $C[0,1]$:

$$\text{for all } f,g\in C[0,1]: \qquad d_1(f,g)=\int_0^1|f(x)-g(x)|\,\mathrm{d}x.$$

This is the metric induced by the norm $\|\cdot\|_1$ from Example 7.8. For each $k\in\mathbb{N}$, define $f_k\in C[0,1]$ as follows: for each $x\in[0,1]$:

$$f_k(x)=\begin{cases}0 & \text{if } 0\leq x\leq\frac{1}{2},\\ 2^k\left(x-\frac{1}{2}\right) & \text{if } \frac{1}{2}<x\leq\frac{1}{2}+\frac{1}{2^k},\\ 1 & \text{if } \frac{1}{2}+\frac{1}{2^k}<x\leq 1.\end{cases}$$

The function f_k increases linearly from 0 to 1 on the interval $\left[\frac{1}{2},\frac{1}{2}+\frac{1}{2^k}\right]$; it is zero for lower values and one for higher values. The sequence $(f_k)_{k\in\mathbb{N}}$ is a Cauchy sequence (why?) in $C[0,1]$, but does not converge in $(C[0,1],d_1)$: the candidate limit f will have $f(x)=0$ if $x<\frac{1}{2}$ and $f(x)=1$ if $x>\frac{1}{2}$, so f is not continuous in $x=\frac{1}{2}$.

As an aside, notice that $x\mapsto x^k$ does converge in $(C[0,1],d_1)$, to the zero function. ◁

What goes wrong above is that sequences may be Cauchy sequences, but that the candidate limit is not part of the set under consideration. This motivates the following definition:

Definition 14.2 Let (X,d) be a metric space and $U\subseteq X$. The set U is ***complete*** if each Cauchy sequence in U has a limit in U. In particular, if X itself is complete, we call (X,d) a ***complete metric space***.

As an important special case, a complete normed vector space is called a ***Banach space***. The examples above show that $((0,1),|\cdot|)$ and $(C[0,1],d_1)$ are not complete. The space $(\mathbb{R}^n,d_2)$ *is* complete:

Theorem 14.1 (Completeness of $\mathbb{R}^n$ with the Euclidean distance)

$(\mathbb{R}^n, d_2)$ is complete.

Proof: Let $(x_k)_{k\in\mathbb{N}}$ be a Cauchy sequence in $(\mathbb{R}^n, d_2)$.
STEP 1: The sequence is bounded.

Take $\varepsilon = 1$. Since $(x_k)_{k\in\mathbb{N}}$ is a Cauchy sequence, there is an $N \in \mathbb{N}$ such that for all $k, m \geq N$: $d_2(x_k, x_m) < 1$. In particular, for all $k \geq N$: $d_2(x_k, x_N) < 1$. By the triangle inequality:

$$\text{for all } k \geq N: \qquad \|x_k\|_2 = d_2(x_k, \mathbf{0}) \leq d_2(x_k, x_N) + d_2(x_N, \mathbf{0}) < 1 + \|x_N\|_2.$$

It follows that the sequence is bounded by $\max\{\|x_1\|_2, \ldots, \|x_{N-1}\|_2, \|x_N\|_2 + 1\}$.
STEP 2: Since $(x_k)_{k\in\mathbb{N}}$ is bounded, it has a convergent subsequence $(x_{k(m)})_{m\in\mathbb{N}}$ with limit $x \in \mathbb{R}^n$ by the Bolzano Weierstrass Theorem 13.2.
STEP 3: The entire sequence $(x_k)_{k\in\mathbb{N}}$ converges to x.

By definition of convergence of the subsequence and of a Cauchy sequence, for each $\varepsilon > 0$, there are $M, N \in \mathbb{N}$ such that

$$d(x_{k(m)}, x) < \varepsilon/2 \text{ for all } m \geq M \qquad \text{and} \qquad d(x_k, x_m) < \varepsilon/2 \text{ for all } k, m \geq N.$$

Let $m \geq M$ be such that $k(m) \geq N$. Then for all $k \geq N$:

$$d(x_k, x) \leq d(x_k, x_{k(m)}) + d(x_{k(m)}, x) < \varepsilon/2 + \varepsilon/2 = \varepsilon.$$

So the sequence $(x_k)_{k\in\mathbb{N}}$ converges to x. □

This establishes that in $\mathbb{R}^n$ with its usual metric, a sequence is convergent *if and only if* it is a Cauchy sequence.

Theorem 14.2

Let (X, d) be a complete metric space and $U \subseteq X$. Set U is closed if and only if U is complete.

Proof: First assume U is closed. Consider a Cauchy sequence in U. Since $U \subseteq X$, it is a Cauchy sequence in X. X is complete, so the sequence has a limit in X. Since U is closed, the limit of the sequence in U lies in U as well (Theorem 13.4). This proves that each Cauchy sequence in U has a limit in U: U is complete.

Conversely, assume U is complete. We use Theorem 13.4 to show that U is closed. Consider a convergent sequence in U. Since it converges, it is a Cauchy sequence in U. Since U is complete, its limit belongs to U. This proves that for every convergent sequence in U, its limit belongs to U: U is closed. □

This, of course, is what goes wrong in the examples above: the candidate for a limit is not contained in the sets, that turn out not to be closed. Combining the previous two theorems, it follows that each closed subset in $(\mathbb{R}^n, d_2)$ is complete.

Definition 14.3 Let (Y, d) be a metric space.

☒ If X is an arbitrary nonempty set, define the ***space of bounded functions from*** X ***to*** Y as

$$B(X, Y) = \{f : X \to Y \mid f \text{ is bounded}\}.$$

Here, bounded means that $f(X)$ is a bounded subset of (Y, d).

☒ If X is a metric space, define the ***space of bounded, continuous functions from*** X ***to*** Y as

$$C(X,Y) = \{f : X \to Y \mid f \text{ is bounded and continuous}\}.$$

☒ Both spaces can be endowed with the ***supremum metric*** d_∞ that assigns to each pair of functions f, g the distance

$$d_\infty(f,g) = \sup\{d(f(x),g(x)) : x \in X\}. \tag{32}$$

☒ Both $(B(X,Y),d_\infty)$ and $(C(X,Y),d_\infty)$ are metric spaces.

Notice:

1. $C(X,Y) \subseteq B(X,Y)$;
2. The supremum in (32) is well-defined, since X is nonempty and f and g are bounded;
3. In $C(X,Y)$, the assumption that f is bounded can be dispensed with if X is compact; see Theorems 17.6 and 17.1(e);
4. Earlier, we wrote $B(\mathbb{N},\mathbb{R}) = B(\mathbb{N})$ and $C([a,b],\mathbb{R}) = C[a,b]$.

Theorem 14.3

If (Y,d) is complete, then so are $(B(X,Y),d_\infty)$ and $(C(X,Y),d_\infty)$.

Proof: For $(B(X,Y),d_\infty)$**:** Let $(f_k)_{k\in\mathbb{N}}$ be a Cauchy sequence in $(B(X,Y),d_\infty)$. In particular, for each $x \in X$, the sequence $(f_k(x))_{k\in\mathbb{N}}$ is a Cauchy sequence in Y. Since Y is complete, its limit

$$f(x) = \lim_{k\to\infty} f_k(x) \in Y \tag{33}$$

exists. For the function $f : X \to Y$ defined in (33) we establish two things:
(i) f is bounded, so that $f \in B(X,Y)$;
(ii) $(f_k)_{k\in\mathbb{N}}$ converges to f in $(B(X,Y),d_\infty)$.
Proof of *(i)*: $(f_k)_{k\in\mathbb{N}}$ is a Cauchy sequence, so for $\varepsilon = 1$, there is an $N \in \mathbb{N}$ such that

$$\text{for all } k,m \geq N: \qquad d_\infty(f_k,f_m) < 1. \tag{34}$$

By assumption, f_N is bounded:

$$\text{for some } y \in Y \text{ and some } \varepsilon > 0: \qquad f_N(X) \subseteq B(y,\varepsilon). \tag{35}$$

We will show that

$$f(X) \subseteq B(y,\varepsilon+2). \tag{36}$$

Let $z \in f(X)$: there is an $x \in X$ with $f(x) = z$. By (33) for $\varepsilon = 1$, there is an $M \in \mathbb{N}$ such that for all $k \geq M$:

$$d(f(x),f_k(x)) < 1. \tag{37}$$

Let $k \geq \max\{N,M\}$. By the triangle inequality and (37), (34), (35):

$$d(f(x),y) \leq d(f(x),f_k(x)) + d(f_k(x),f_N(x)) + d(f_N(x),y) < 1+1+\varepsilon,$$

proving (36).

Proof of *(ii)*: Let $\varepsilon > 0$. To show: there is an $N \in \mathbb{N}$ such that for all $k \geq N$: $d_\infty(f_k, f) < \varepsilon$.

Since $(f_k)_{k\in\mathbb{N}}$ is a Cauchy sequence, there is an $N \in \mathbb{N}$ such that for all $k, m \geq N$: $d_\infty(f_k, f_m) < \frac{1}{3}\varepsilon$.

So for all $x \in X$ and all $k, m \geq N$:

$$d(f_k(x), f(x)) \leq d(f_k(x), f_m(x)) + d(f_m(x), f(x)) < \frac{1}{3}\varepsilon + d(f_m(x), f(x)). \tag{38}$$

For each $x \in X$, $f(x) = \lim_m f_m(x)$, so we can choose m sufficiently large so also $d(f_m(x), f(x)) < \frac{1}{3}\varepsilon$. It follows that if $k \geq N$, then for all $x \in X$:

$$d(f_k(x), f(x)) < \frac{2}{3}\varepsilon < \varepsilon.$$

For $(C(X,Y), d_\infty)$: Let $(f_k)_{k\in\mathbb{N}}$ be a Cauchy sequence in $(C(X,Y), d_\infty)$. Since $C(X,Y) \subseteq B(X,Y)$, the sequence converges in $(B(X,Y), d_\infty)$ to a limit $f \in B(X,Y)$. It remains to show that f is continuous, i.e., lies in $C(X,Y)$.

Let $\varepsilon > 0$ and $a \in X$. Since $(f_k)_{k\in\mathbb{N}}$ converges to f, there is an $N \in \mathbb{N}$ such that

$$\text{if } k \geq N, \text{ then } d_\infty(f_k, f) < \frac{\varepsilon}{3},$$

or, in other words,

$$\text{for all } x \in X: \qquad d(f_k(x), f(x)) < \frac{\varepsilon}{3}.$$

Since f_m is continuous at a, there is a $\delta > 0$ such that

$$\text{if } x \in X \text{ and } d(x,a) < \delta, \text{ then } d(f_m(x), f_m(a)) < \frac{\varepsilon}{3}.$$

Combining these two things, it follows that if $x \in X$ and $d(x,a) < \delta$, then

$$d(f(x), f(a)) \leq d(f(x), f_m(x)) + d(f_m(x), f_m(a)) + d(f_m(a), f(a)) < \frac{\varepsilon}{3} + \frac{\varepsilon}{3} + \frac{\varepsilon}{3} = \varepsilon.$$

So f is indeed continuous! □

Example 14.3 ($C[a,b]$ with the supremum metric is complete) Taking $X = [a,b]$ for some real numbers a and b with $a < b$ and $Y = \mathbb{R}$, it follows that $(C[a,b], d_\infty)$ is complete.

This example shows that completeness may depend on the metric you choose: the set of continuous functions on the unit interval with the metric in Example 14.2 is *not* complete, but we just argued that it *is* complete under the supremum metric. ◁

15 The Banach contraction theorem

In this section, we prove the Banach contraction theorem, a result that is used in economics, for instance, to numerically approximate solutions to systems of equations, determine equilibria, or solve dynamic optimization problems. Roughly speaking, it says that a function from and to the same complete metric space that moves each pair of points closer together by a factor at least $\alpha \in [0,1)$, has a fixed point. It gives conditions for the existence of a fixed point, but also tells how to compute it by successive approximations. The method of successive approximation applies to equations of the form $T(x) = x$. A solution of such an equation is a fixed point of T, since T leaves x unaffected. To find a fixed point by successive approximation, start with an initial guess x_0 and compute $x_1 = T(x_0)$. Proceed in this way, computing successive points $x_{n+1} = T(x_n)$. Under appropriate assumptions, the sequence $(x_n)_{n=0,1,2,\ldots}$ converges to a fixed point of T.

Recall the notation $T^n(x) = T(T^{n-1}(x))$ for the n-fold composition of T with itself and $T^0(x) = x$, the sequence of successive approximations is often written as $(T^n(x_0))_{n=0,1,2,\ldots}$. Be careful not to confuse $T^n(x)$ with $T(x)^n$! The first involves the n-fold composition of T with itself, whereas the second is simply the value $T(x)$ raised to the power n.

Example 15.1 Consider $T : \mathbb{R} \to \mathbb{R}$ with $T(x) = x+1$. Then

$$\begin{aligned} T(x) &= x+1, \\ T^2(x) &= T(T(x)) = T(x+1) = x+2, \\ T^3(x) &= T(T^2(x)) = T(x+2) = x+3, \\ &\vdots \\ T^n(x) &= x+n, \end{aligned}$$

whereas $T(x)^n = (x+1)^n$. ◁

Definition 15.1 Let (X,d) be a metric space. A function $T : X \to X$ is a ***contraction*** if there exists a number $\alpha \in [0,1)$, the ***modulus*** of T, such that for all $x, y \in X$:

$$d(T(x), T(y)) \leq \alpha d(x,y). \tag{39}$$

Each contraction is Lipschitz continuous and consequently (uniformly) continuous.

Example 15.2 $f : [0,\infty) \to [0,\infty)$ with $f(x) = \frac{1}{x+2}$ is a contraction with modulus $\frac{1}{4}$:

$$|f(x) - f(y)| = \left| \frac{1}{x+2} - \frac{1}{y+2} \right| = \frac{|y-x|}{|x+2|\,|y+2|} \leq \frac{|y-x|}{2 \cdot 2} = \frac{1}{4}|x-y|.$$ ◁

Example 15.3 $f : [0,\varepsilon] \to [0,\varepsilon]$ with $f(x) = x^2$ is a contraction as long as $0 < \varepsilon < \frac{1}{2}$:

$$|f(x) - f(y)| = \left|x^2 - y^2\right| = |(x+y)(x-y)| = |x+y|\,|x-y| \leq 2\varepsilon|x-y|.$$

It is not a contraction if $\varepsilon = \frac{1}{2}$, because $|x+y|$ can then be chosen arbitrarily close to 1. ◁

Example 15.4 (Functions with small derivatives) If $U \subseteq \mathbb{R}$ is a nonempty interval and there is a number $\alpha \in [0,1)$ such that the function $f : U \to U$ is differentiable on the interior of U with $|f'(x)| \leq \alpha$ for all $x \in \operatorname{int}(U)$, then f is a contraction: for any two distinct $x, y \in U$, the Mean Value Theorem (see Theorem A.2 in the appendix) implies that there is a $z \in U$ between x and y with

$$|f(x) - f(y)| = |f'(z)(x-y)| = |f'(z)|\,|x-y| \leq \alpha|x-y|.$$

For instance, $f:[1,\infty)\to[1,\infty)$ with $f(x)=\sqrt{x}$ is a contraction: $|f'(x)|=\frac{1}{2\sqrt{x}}\le\frac{1}{2}$ if $x\ge 1$.

It is important to keep in mind the requirement that the derivative is bounded in absolute value by some number α less than one. Simply having a derivative that is less than one is not enough (see Exercise 15.3). ◁

Definition 15.2 Let $f:S\to S$ be a function from and to a nonempty set S. A ***fixed point*** of f is an element $s\in S$ with $f(s)=s$.

Clearly, not every function has a fixed point. The Banach contraction theorem assures their existence for contractions on complete metric spaces:

Theorem 15.1 (Banach contraction theorem)

Let $T:X\to X$ be a contraction on a complete metric space (X,d). Then:

(a) Existence of a unique fixed point: there is a unique $x\in X$ with $T(x)=x$.

Let $x_0\in X$ and define the sequence $(x_k)_{k\in\mathbb{N}}$ of successive function values as follows:

$$\text{for each } k\in\mathbb{N}: \qquad x_k=T(x_{k-1})=T^k(x_0). \tag{40}$$

(b) Convergence: the sequence $(x_k)_{k\in\mathbb{N}}$ converges to x.

(c) Speed of convergence: if T has modulus $\alpha\in[0,1)$, then for each $k\in\mathbb{N}$:

$$d(T^k(x_0),x)\le\frac{\alpha}{1-\alpha}d(T^{k-1}(x_0),T^k(x_0)) \quad\text{and}\quad d(T^k(x_0),x)\le\frac{\alpha^k}{1-\alpha}d(T(x_0),x_0). \tag{41}$$

Proof: (a) and (b): We show that the sequence of successive function values in (40) is a Cauchy sequence. For each $k\in\mathbb{N}$, definition (40) and the fact that T is a contraction give:

$$d(x_{k+1},x_k)=d(T(x_k),T(x_{k-1}))\le\alpha d(x_k,x_{k-1}).$$

Repeating this step k times gives

$$d(x_{k+1},x_k)\le\alpha^k d(x_1,x_0). \tag{42}$$

By the triangle inequality, it follows that for each $p\in\mathbb{N}$:

$$\begin{aligned}
d(x_{k+p},x_k)&\le d(x_{k+p},x_{k+p-1})+d(x_{k+p-1},x_{k+p-2})+\cdots+d(x_{k+1},x_k)\\
&\le\left(\alpha^{k+p-1}+\alpha^{k+p-2}+\cdots+\alpha^k\right)d(x_1,x_0)\\
&\le\left(\alpha^k\sum_{n=0}^{\infty}\alpha^n\right)d(x_1,x_0)\\
&=\frac{\alpha^k}{1-\alpha}d(x_1,x_0),
\end{aligned}$$

so that $(x_k)_{k\in\mathbb{N}}$ is a Cauchy sequence. By completeness of X, $(x_k)_{k\in\mathbb{N}}$ has a limit $x\in X$. By continuity of T, the limit x is a fixed point:

$$T(x)=T(\lim_{k\to\infty}x_k)=\lim_{k\to\infty}T(x_k)=\lim_{k\to\infty}x_{k+1}=x.$$

To establish that x is the *unique* fixed point, suppose there are distinct $x,y\in X$ with $T(x)=x$ and $T(y)=y$. This gives a contradiction:

$$d(x,y)=d(T(x),T(y))\le\alpha d(x,y)<d(x,y).$$

(c): Consecutively using that x is a fixed point of T, T is a contraction, and the triangle inequality gives, for each $k \in \mathbb{N}$:

$$d(T^k(x_0), x) = d(T^k(x_0), T(x)) \leq \alpha d(T^{k-1}(x_0), x) \leq \alpha \left[d(T^{k-1}(x_0), T^k(x_0)) + d(T^k(x_0), x) \right].$$

Rearranging terms gives the first part of (c). The second part of (c) follows from substituting (42). □

The inequalities in (41) are important for several reasons:

- ⊠ They indicate that the sequence of successive approximations converges exponentially fast to the fixed point x.
- ⊠ They help to derive bounds on the number of steps that are required to obtain such a fixed point within a given level of precision.
- ⊠ They provide a motivation for the so-called "method of undetermined coefficients", which is just a fancy name for making educated guesses. If you start with an easy point x_0, successive function values of the contraction may be easy to compute. Given their convergence to the fixed point, these may help you to make an educated guess about what the fixed point has to be.

Example 15.5 (Numerically solving equations) Suppose you numerically solve for the solution(s) of the equation

$$\cos x = 2x.$$

Dividing both sides by 2, we are interested in the fixed points of the function $T : \mathbb{R} \to \mathbb{R}$ with $T(x) = \frac{1}{2}\cos x$. Since $|T'(x)| = \left|-\frac{1}{2}\sin x\right| \leq \frac{1}{2}$, T is a contraction with modulus $\frac{1}{2}$. So there is a unique solution x. If you start with $x_0 = 0$, then $T(x_0) = \frac{1}{2}\cos 0 = \frac{1}{2}$, so by (41):

$$d(x, T^n(0)) \leq \frac{(1/2)^n}{1-(\frac{1}{2})} \cdot \frac{1}{2} = \frac{1}{2^n}.$$

Approximating the solution to within precision $\varepsilon = 0.001$ requires at most $n = 10$ iterations:

$$d(x, T^{10}(0)) \leq \frac{1}{2^{10}} < 0.001$$

◁

If the whole idea is that function values $x, T(x), T^2(x), T^3(x), \ldots$ get closer and closer to one another, isn't it sufficient, instead of (39), to assume that

$$\text{for all } x, y \in X, \text{ if } x \neq y, \text{ then:} \qquad d(T(x), T(y)) < d(x, y), \tag{43}$$

to find a fixed point? Functions with this property are called ***nonexpansive***. Our next example shows that, in general, the answer is negative. In Exercise 15.2, however, we will see that nonexpansiveness does suffice if the metric space (X, d) is assumed to be compact.

Example 15.6 Function $f : [1, \infty) \to [1, \infty)$ with $f(x) = x + \frac{1}{x}$ is nonexpansive: it has derivative $f'(x) = 1 - \frac{1}{x^2} > 0$ on $(1, \infty)$, so f is strictly increasing. Let $x, y \in [1, \infty)$ have $x \neq y$. Without loss of generality, $x > y$. Then

$$|f(x) - f(y)| = f(x) - f(y) = x + \frac{1}{x} - y - \frac{1}{y} = x - y + \underbrace{\frac{1}{x} - \frac{1}{y}}_{<0} < x - y = |x - y|.$$

But f has no fixed point: $f(x) > x$ for all x in its domain. ◁

Sometimes the function T itself is not a contraction, but its n-fold replica is. Even then, T has a unique fixed point.

Theorem 15.2

Let (X,d) be a complete metric space and $T : X \to X$ a function. Suppose there is an $n \in \mathbb{N}$ such that $T^n = T \circ \cdots \circ T : X \to X$ is a contraction. Then there is a *unique* $x \in X$ with $T(x) = x$.

Proof: By Theorem 15.1, there is a unique $x \in X$ with $T^n(x) = x$. Also $T(x)$ is a fixed point of T^n:

$$T^n(T(x)) = T(T^n(x)) = T(x).$$

So $T(x) = x$. Moreover, each fixed point of T is a fixed point of T^n, which is unique. □

The following result helps to show that a fixed point may be contained in a smaller set than the one on which the contraction is originally defined.

Theorem 15.3

Let (X,d) be a complete metric space, $T : X \to X$ a contraction with fixed point x, and $X_1 \subseteq X$ nonempty and closed.

(a) If $T(X_1) \subseteq X_1$, the fixed point x lies in X_1;

(b) If $X_2 \subseteq X$ is such that $T(X_1) \subseteq X_2 \subseteq X_1$, the fixed point x lies in X_2.

Proof: **(a):** By the Banach fixed point theorem, $T : X_1 \to X_1$ has a unique fixed point. Since $X_1 \subseteq X$, this is a fixed point on X as well. But $T : X \to X$ has only one fixed point, namely x.
(b): By (a), $x \in X_1$, so $x = T(x) \in T(X_1) \subseteq X_2$. □

15.1 Application: ranking the relevance of websites

Google's PageRank algorithm uses Banach's contraction theorem to assign importance to webpages. Its main idea is that (1) the importance of a page is the importance bestowed upon it by other pages that link to it and (2) if a page j links to $n(j)$ others, it assigns $\frac{1}{n(j)}$ of its importance to each of them.

Formally, assume there are n pages. For each page i, let $n(i)$ be the number of other pages that i links to and let $L(i)$ be the set of other pages with a link to i. As a first attempt, the importance vector $x = (x_1, \ldots, x_n)$ we search for has nonnegative coordinates that are normalized to sum to one and that satisfy, for each page i,

$$x_i = \sum_{j \in L(i)} \frac{1}{n(j)} x_j.$$

In other words, x must be a probability vector, an element of the set

$$\Delta_n = \left\{ x \in \mathbb{R}^n_+ : \textstyle\sum_{i=1}^n x_i = 1 \right\},$$

with $x = Ax$, where A is an $n \times n$ matrix whose j-th column can be of two types. If page j is a 'dangling node' with no links to other pages, then all entries in column j are zero. And if j links to $n(j) > 0$ other pages, then the i-th entry in column j is $a_{ij} = \frac{1}{n(j)}$ if j links to i and $a_{ij} = 0$ otherwise.

Expression $x = Ax$ is a fixed-point equation. But this system of equations need not have a solution $x \in \Delta_n$, nor a unique one. Two adjustments are made to resolve this.

We first deal with dangling nodes. In A, these correspond to columns where all entries are zero. Replace each such column with $(1/n,\dots,1/n)$, giving equal weight to all pages. In the new matrix, which we call B, each column is a probability vector, i.e., an element of Δ_n. Secondly, we fix a number $\alpha \in (0,1)$, the 'damping factor' (usually around 0.85), and an arbitrary probability vector $p \in \Delta_n$ called the 'personalization vector'. How Google chooses p is secret, but it can help customize search results by assigning higher weight to sites often visited by a specific user. Let P be the $n \times n$ matrix where all columns equal p. We define a new matrix, where once again all columns are probability vectors:

$$M = \alpha B + (1-\alpha)P.$$

Banach's contraction theorem assures that there is a unique probability vector x^* solving

$$x^* = Mx^*,$$

and that it is the limit of the sequence $y, My, M^2y, M^3y, \dots$ for any probability vector y. This fixed point $x^* = (x_1^*,\dots,x_n^*)$ indicates the importance x_i^* that PageRank assigns to each webpage i.

To see that Banach's contraction theorem applies, we must show that the function $f : \Delta_n \to \Delta_n$ with $f(x) = Mx$ is a contraction and that Δ_n is complete. Instead of using the Euclidean distance, the clever move is to use the taxicab norm/distance. This leads to a major simplification. Each probability vector has length one: if $x \in \Delta_n$, then $\|x\|_1 = \sum_{i=1}^n |x_i| = \sum_{i=1}^n x_i = 1$.

We argue that function f is a contraction with modulus α. First note that $Px = p$ for all $x \in \Delta_n$, because each column of P equals p and therefore

$$Px = \sum_{i=1}^n x_i p = \underbrace{\Big(\sum_{i=1}^n x_i\Big)}_{=1} p.$$

Let $b_1,\dots,b_n$ be the columns of B, all with length one. For all $x, y \in \Delta_n$ the triangle inequality gives:

$$\begin{aligned} d_1(f(x),f(y)) &= \|Mx - My\|_1 = \|\alpha Bx - \alpha By\|_1 = \alpha \|B(x-y)\|_1 \\ &= \alpha \left\| \textstyle\sum_{i=1}^n (x_i - y_i) b_i \right\|_1 \le \alpha \sum_{i=1}^n |x_i - y_i| \|b_i\|_1 \\ &= \alpha \|x - y\|_1 = \alpha d_1(x,y). \end{aligned}$$

Finally, to see that Δ_n is complete, note that $\mathbb{R}^n$ with the taxicab metric d_1 is complete.[4] And Δ_n is a closed subset, hence complete as well. So we can indeed apply the contraction theorem.

Exercises section 15

15.1 (Blackwell's conditions) Let U be a nonempty set. Prove: If $T : B(U,\mathbb{R}) \to B(U,\mathbb{R})$ satisfies:

(a) monotonicity: if $f \le g$, then $T(f) \le T(g)$. [Here, $f \le g$ means $f(x) \le g(x)$ for all $x \in U$];

(b) discounting: there is a $\beta \in (0,1)$ such that for each $f \in B(U,\mathbb{R})$ and each nonnegative, constant function $c \in B(U,\mathbb{R})$:

$$T(f+c) \le T(f) + \beta c;$$

then T is a contraction with modulus β in $(B(U,\mathbb{R}), d_\infty)$.

15.2 (Nonexpansive maps on compact metric spaces) Let (X,d) be a compact metric space and $T : X \to X$ a nonexpansive function. We prove that T has a unique fixed point x and that $T^k(x_0) \to x$ for all $x_0 \in X$.

(a) Show that T has at most one fixed point.

[4] Do you see why? The easiest argument uses that $\mathbb{R}^n$ with its usual Euclidean distance is complete and that there is a close connection between the Euclidean and taxicab norm/distance: $\|x\|_2 \le \|x\|_1 \le \sqrt{n}\|x\|_2$ by Exc. 7.1.

(b) Show that T has a fixed point. HINT: show that $f : X \to [0,\infty)$ with $f(x) = d(x, T(x))$ is continuous, achieves a minimal value 0, and that the minimum location must be a fixed point.

(c) Show that if $x \in X$ is the fixed point of T and $x_0 \in X$, then $\lim_{k\to\infty} T^k(x_0) = x$. HINT: use f to show that distances $d(T^k(x_0), x)$ form a (weakly) decreasing sequence with limit 0.

It is important to realize that this is not just a special example of Banach's fixed point theorem or the variant that applies if T^k happens to be a contraction:

(d) Consider $T : [0,1] \to [0,1]$ with $T(x) = x/(1+x)$. Show that T is nonexpansive and, for each $k \in \mathbb{N}$, that $T^k(x) = x/(1+kx)$ and that T^k is not a contraction.

15.3 Consider $f : \mathbb{R} \to \mathbb{R}$ with $f(x) = \sqrt{x^2+1}$.

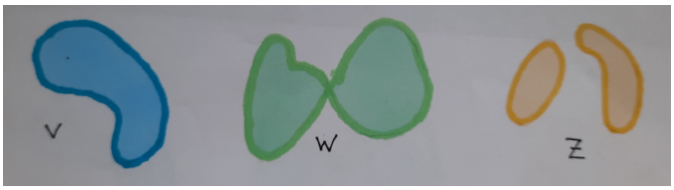
(a) Show that $|f'(x)| < 1$ for all $x \in \mathbb{R}$.

(b) Does f have a fixed point?

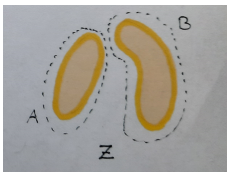
(c) Is f a contraction?

16 Connected sets

In the figure below it makes sense to call sets V and W connected, but to call Z disconnected.



The set Z is cut into pieces with some room in between: we can find open sets A and B such that some, but not all elements of Z belong to A, the remaining elements belong to B, and the sets A and B have nothing in common.



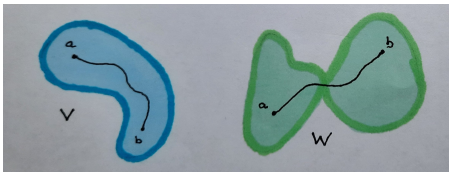
We will call the pair A and B a separation and call a set connected if it has no separation:

Definition 16.1 Let Y be a subset of a metric space. A ***separation*** of Y is a pair of open sets A and B such that

$$Y \subseteq A \cup B, \qquad Y \cap A \neq \emptyset, \qquad Y \cap B \neq \emptyset, \qquad A \cap B = \emptyset.$$

Set Y is ***connected*** if there is no separation of Y.

A second intuitive reason for calling sets like V and W connected is that between each pair of elements you can draw, without lifting your pen (continuity), a path that never leaves the set. For one pair of points in V and one pair of points in W, such a path is drawn below:



If we imagine that it takes one time unit to draw that path, the following definition is clear:

Definition 16.2 A subset Y of a metric space (X,d) is ***path-connected*** if for each pair of elements $a, b \in Y$ there is a continuous function $p : [0,1] \to X$ that

- ☒ starts at a: $p(0) = a$,
- ☒ ends at b: $p(1) = b$,
- ☒ and stays inside Y at all times: $p(t) \in Y$ for all $t \in [0,1]$.

We call p a ***path*** from a to b inside Y.

Example 16.1 Draw a plus sign (+) on a piece of paper. This is a path-connected set in $\mathbb{R}^2$: points that both lie on its vertical or both lie on its horizontal part can be connected by a straight path. And if one point lies on the vertical part and the other on the horizontal part, you can draw a kinked path between them that passes through the point where these horizontal and vertical parts intersect. ◁

If a set contains a path between each pair of elements, it is connected:

Theorem 16.1

Path-connected sets are connected.

Proof: Let Y be a path-connected set in a metric space and suppose, to the contrary, that A and B are a separation of Y. Since $Y \cap A$ and $Y \cap B$ are nonempty, we can pick an element a from the first set and an element b from the second. A and B are disjoint, so $a \neq b$. Since Y is path-connected we can pick a path p that starts in a and ends in b, i.e., $p(0) = a$ and $p(1) = b$.

Since A is an open neighborhood of $a = p(0)$, continuity of the path implies that points $p(t)$ on the path with t sufficiently close to zero lie in A; likewise, points $p(t)$ with t sufficiently close to one lie in B. So if we increase t from zero to one, we start out with points $p(t)$ in A and end with points in B: somewhere along the way we leave A forever. Let's look at this threshold point $p(t^*)$ with

$$t^* = \sup\left\{t \in [0,1] : p(t) \in A\right\} \in (0,1).$$

It cannot lie in the open set A, because then, by continuity of the path, so do $p(t)$'s with slightly larger $t > t^*$, contradicting our threshold t^*.

Likewise, $p(t^*)$ cannot lie in the open set B, because then so would all points $p(t)$ with slightly smaller $t < t^*$. But since t^* is the smallest upper bound on

$$\left\{t \in [0,1] : p(t) \in A\right\},$$

such a smaller t is not an upper bound: there is a t' between t and t^* with $p(t') \in A$, contradicting that all these points lie in B.

But this contradicts that $Y \subseteq A \cup B$: each element of Y and in particular the point $p(t^*)$ must belong to either A or B. □

The converse of Theorem 16.1 is false: there are sets that are connected, but not path-connected. Such examples are too complicated to treat in the space of these notes. The most common application of Theorem 16.1 is:

Example 16.2 (Convex sets are connected) In normed vector spaces, each convex set is connected. It is enough to show that convex sets are path-connected. The path between any two points a and b in our set is the obvious one: convexity says that the entire line piece

$$p(t) = (1-t)a + tb \quad \text{with} \quad t \in [0,1]$$

between a and b belongs to our set and this function p is continuous. ◁

In $\mathbb{R}$, we can say something stronger:

Example 16.3 A subset of $\mathbb{R}$ is connected if and only if it is convex. We already saw that convex sets are connected. Conversely, suppose $Y \subseteq \mathbb{R}$ is not convex: there are elements a and b of Y and a number $t \in (0,1)$ such that the point $c = (1-t)a + tb$ on the line piece between a and b does not belong to Y. But then the open intervals $A = (-\infty, c)$ and $B = (c, \infty)$ are a separation of Y, contradicting that Y is connected. ◁

The sets V and W in our first figure and the plus sign from Example 16.1 show that in metric spaces other than $\mathbb{R}$, connected sets need not be convex.

Earlier we saw that in every metric space (X,d) the empty set ($\emptyset$) and X itself are both open and closed. In most applications, those sets are unusual:

Example 16.4 The only sets in $\mathbb{R}^n$ that are both open and closed are $\emptyset$ and $\mathbb{R}^n$. Suppose $A \subseteq \mathbb{R}^n$ is both open and closed, but $\emptyset \neq A \neq \mathbb{R}^n$. Then also its complement $B = A^c$ is distinct from $\emptyset$ and $\mathbb{R}^n$. As the complement of a closed set, B is open. But then A and B are a separation of $\mathbb{R}^n$, contradicting that $\mathbb{R}^n$ is convex and consequently connected. ◁

If you plug the elements of a connected set into a continuous function, their images form a new connected set:

Theorem 16.2 (The continuous image of a connected set is connected)

If $f : X \to Y$ is a continuous function between two metric spaces and $C \subseteq X$ is connected, then $f(C) \subseteq Y$ is connected.

Proof: Suppose $f(C)$ is not connected, but has a separation by open sets A and B. Then (verify yourself!) their pre-images $A' = f^{-1}(A)$ and $B' = f^{-1}(B)$ are a separation of C. But C is connected: it has no separation. □

Example 16.5 Since a path p is a continuous function on the connected set $[0,1]$, the points $\{p(t) : t \in [0,1]\}$ on the path form a connected set. ◁

Example 16.6 (Continuous functions on a connected domain have a connected graph)
Let $f : X \to Y$ be a continuous function between two metric spaces. Its graph

$$\{(x, f(x)) : x \in X\}$$

is the image $g(X)$ of the continuous function $g : X \to X \times Y$ with $g(x) = (x, f(x))$. So if X is connected, the graph of f is connected.

For instance, the set $[-1,1]$ is convex, hence connected and the function $x \mapsto x^2$ is continuous, so the graph of this quadratic function, the set $\{(x, x^2) \in \mathbb{R}^2 : x \in [-1,1]\}$, is a connected subset of $\mathbb{R}^2$. ◁

Applying Theorem 16.2 to real-valued functions, we obtain the famous Intermediate Value Theorem: if a continuous function on a connected set achieves two distinct values, it also achieves all values in-between.

Theorem 16.3 (Intermediate Value Theorem)

Given:

- ☒ a continuous function $f : X \to \mathbb{R}$ on a metric space,
- ☒ a connected subset C of X,
- ☒ and two elements a and b of C,

the function achieves every function value between $f(a)$ and $f(b)$.

Proof: By our previous theorem, $f(C)$ is a connected subset of $\mathbb{R}$. By Example 16.3, it is convex. It contains $f(a)$ and $f(b)$, so it also contains each number on the line piece between them. □

In economics you encounter the use of connected sets for instance in expected utility theory where a variant of the Intermediate Value Theorem is used to show that for every lottery you can find an equivalent lottery that looks much simpler: it is a mixture of a really good and a really bad outcome. It is also a useful tool to see that some equations have solutions even if you cannot find the exact one:

Example 16.7 There is a real number x solving the equation $x^7 + x^2 - 1 = 0$, because the continuous function $f : \mathbb{R} \to \mathbb{R}$ with $f(x) = x^7 + x^2 - 1$ achieves a negative value $f(0) = -1$ at $x = 0$ and a positive value $f(1) = 1$ at $x = 1$, so by the Intermediate Value Theorem it achieves the value zero for some x in the connected interval $[0,1]$. ◁

Example 16.8 (A simple fixed-point result) Each continuous function $f : [0,1] \to [0,1]$ has a fixed point: $f(x^*) = x^*$ for some $x^* \in [0,1]$. Graphically, the graph of f and the diagonal line $y = x$ intersect.

If $f(0) = 0$, we are done: 0 is a fixed point. If $f(1) = 1$, we are done: 1 is a fixed point. If $f(0) > 0$ and $f(1) < 1$, then the continuous function g with $g(x) = f(x) - x$ is positive at 0 because $g(0) = f(0) - 0 > 0$, negative at 1 because $g(1) = f(1) - 1 < 0$. By the Intermediate Value Theorem, it achieves value 0 somewhere in $[0,1]$: there is an $x^* \in [0,1]$ with $g(x^*) = f(x^*) - x^* = 0$. In particular, $f(x^*) = x^*$, our desired fixed point. ◁

Brouwer's fixed-point theorem, Theorem 26.6, is a substantial generalization of this result.

The set Z in our first figure of this section consisted of two nonempty, closed parts with nothing in common and we could find a separation. This holds more generally:

Theorem 16.4 (Disjoint closed sets have disjoint open neighborhoods)

If C_1 and C_2 are nonempty, disjoint, closed subsets of a metric space, then there are disjoint open sets A and B with $C_1 \subseteq A$ and $C_2 \subseteq B$.

Proof: Each $c_1 \in C_1$ lies in the complement of C_2. This complement is open, so there is an $\varepsilon(c_1) > 0$ such that the entire open ball $B(c_1, \varepsilon(c_1))$ is disjoint from C_2. Likewise, for each $c_2 \in C_2$ there is an $\varepsilon(c_2) > 0$ making the open ball $B(c_2, \varepsilon(c_2))$ disjoint from C_1. As unions of open sets,

$$A = \bigcup_{c_1 \in C_1} B\left(c_1, \tfrac{\varepsilon(c_1)}{2}\right) \qquad \text{and} \qquad B = \bigcup_{c_2 \in C_2} B\left(c_2, \tfrac{\varepsilon(c_2)}{2}\right)$$

are open sets containing C_1 and C_2 respectively. They are disjoint: Suppose to the contrary that there were an element $x \in A \cap B$. So there are $c_1 \in C_1$ and $c_2 \in C_2$ with $x \in B(c_1, \varepsilon(c_1)/2)$ and $x \in B(c_2, \varepsilon(c_2)/2)$. Without loss of generality $\varepsilon(c_1) \leq \varepsilon(c_2)$. By the triangle inequality,

$$d(c_1, c_2) \leq d(c_1, x) + d(x, c_2) < \frac{\varepsilon(c_1)}{2} + \frac{\varepsilon(c_2)}{2} \leq \varepsilon(c_2),$$

which means that $c_1 \in C_1$ lies in the ball $B(c_2, \varepsilon(c_2))$. But by construction, that ball had no points in common with C_1: a contradiction. Conclude that A and B are indeed disjoint. □

Exercises section 16

16.1 Which of the following subsets of $\mathbb{R}$ are connected? A sketch can be helpful.

(a) $\{37\}$,

(b) $\{37, 38, 39\}$,

(c) $(37, 39]$,

(d) $[1,2) \cup (2,4]$,

(e) the set $\mathbb{Q}$ of rational numbers.

16.2 Which of the following subsets of $\mathbb{R}^2$ are connected? Again, sometimes a sketch is helpful.

(a) A circle,

(b) $\{x \in \mathbb{R}^2 : x_1 - 4x_2 \leq 8\}$,

(c) $\{x \in \mathbb{R}^2 : x_2 \in \{1,2\}\}$,

(d) $\{x \in \mathbb{R}^2 : x_2 = x_1^2\}$,

(e) $\{x \in \mathbb{R}^2 : x_2 \neq x_1^2\}$,

(f) $\{(a^2 - a, 3a) : a \in [0,1]\}$,

(g) $\mathbb{R}^2$ without its origin.

16.3 Use the Intermediate Value Theorem to show that there is an $x \in \mathbb{R}$ solving the equation $x^3 = 10 + \sqrt{x}$.

16.4 Show: a nonempty finite subset of a metric space is connected if and only if it has exactly one element.

17 Compactness in metric spaces

In optimization problems, it is common to impose restrictions on both the domain and the goal function to assure that the problem under consideration has a solution. The condition on the goal function usually involves some kind of continuity. In this section, we introduce a common constraint imposed on the domain, namely compactness. Compactness excludes all kinds of unruly behavior by making sets look approximately like finite sets. We will explain exactly what we mean by "approximately like a finite set" in Definition 17.2 and will derive a number of pleasant properties that follow from compactness in this section — whose highlight for optimization purposes is Theorem 17.3 and the ensuing remark — as well as in many later ones.

But wait... at this stage, one might be concerned that this description of compact sets as being pretty much like finite sets is rather far away from a definition of compactness that is common in elementary texts on mathematics for economists:

Definition 17.1 A subset of $\mathbb{R}^n$ with its usual distance is ***compact*** if it is closed and bounded.

This concern is well-motivated: Definition 17.1 is useful to recognize compact sets in $\mathbb{R}^n$, by far the most common case in elementary economic theory. In general metric spaces, however, boundedness and closedness do not guarantee the nice behavior we wish from compact sets. In such spaces, the definition will be more restrictive.

Definition 17.2 Let (X,d) be a metric space and let $U \subseteq X$.

- ☒ An ***(open) covering*** of U is a collection $\{O_i : i \in I\}$ of open sets O_i whose union contains U:

$$U \subseteq \cup_{i\in I} O_i. \tag{44}$$

- ☒ A ***subcovering*** from $\{O_i : i \in I\}$ is a subcollection $\{O_i : i \in J\}$ for some $J \subseteq I$ that still covers U:

$$U \subseteq \cup_{i\in J} O_i.$$

- ☒ Such a subcovering is ***finite*** if J has finitely many elements.
- ☒ The set U is ***compact*** if each covering contains a finite subcovering.

Admittedly, this is a bit difficult to read. Let's try to make it more transparent. Think of the open sets as open umbrellas. Say that a point of U is covered if it is kept dry by/contained under an umbrella. A covering of U is just a collection of open umbrellas that keeps each element of the set U dry. Compactness requires that for each such collection of umbrellas, you can throw away all but a finite number of them and *still* keep the set dry!

Example 17.1 The open interval $(0,1) \subseteq \mathbb{R}$ is not compact. Define, for each $x \in (0,1)$, the set $O_x = (\frac{1}{2}x, 1)$. Then $x \in O_x$, so the open sets O_x cover $(0,1)$. But there are not finitely many $x_1,\dots,x_n$ such that $(0,1) \subseteq O_{x_1} \cup \cdots \cup O_{x_n}$: such a finite subcollection cannot cover all numbers in $(0,1)$ close to zero, since $\min\{\frac{1}{2}x_1,\dots,\frac{1}{2}x_n\} > 0$. ◁

Example 17.2 Let X be an infinite set and d the discrete metric (see Example 8.2). Then X is closed (its complement, $X \setminus X = \emptyset$, is open), bounded (all distances are 0 or 1), but not compact: $x \in X$ lies in the open ball $B(x, \frac{1}{2})$, but no other point does. Hence, the open sets $\{B(x, \frac{1}{2}) : x \in X\}$ cover X, but because X has infinitely many elements, there is no finite subcovering. ◁

Example 17.3 In $(C[0,1], d_\infty)$, the set $U = \{f \in C[0,1] : 0 \le f(x) \le 1 \text{ for all } x \in [0,1]\}$ is closed and bounded, but not compact:

- ☒ U is bounded: $U \subseteq B(\mathbf{0}, 2)$;

☒ U is closed, because its complement is open: if $f \in U^c$, then for some $x \in [0,1]$: $f(x) < 0$ or $f(x) > 1$. In the former case, $B(f, \frac{1}{2}|f(x)|) \subseteq U^c$, in the latter, $B(f, \frac{1}{2}(f(x)-1)) \subseteq U^c$.

☒ U is not compact: each $f \in U$ lies in $B(f, \frac{1}{2})$, so the collection of open sets $\{B(f, \frac{1}{2}) : f \in U\}$ covers U, but there is no finite subcovering. Indeed, consider the sequence of functions $(f_n)_{n\in\mathbb{N}}$ in U with

$$f_n(x) = \begin{cases} 0 & \text{if } 0 \le x < \frac{1}{2^n}, \\ 2^n\left(x - \frac{1}{2^n}\right) & \text{if } \frac{1}{2^n} \le x \le \frac{2}{2^n}, \\ 1 & \text{if } \frac{2}{2^n} < x \le 1. \end{cases}$$

The function f_n increases linearly from 0 to 1 on the interval $[\frac{1}{2^n}, \frac{2}{2^n}]$; it is zero for lower values and one for higher values. Functions f_k and f_ℓ with $k \neq \ell$ are at distance one from each other. Consequently, each ball $B(f, \frac{1}{2})$ contains at most one of them: there is no finite subcovering. ◁

That's a bit frustrating: so far we only used the definition of compactness to identify sets that are *not* compact. Doing so is conceptually relatively easy: to show that a set is *not* compact, it suffices to find *one* covering without a finite subcovering. But if you want to use coverings to show that a set is compact, you have to show that *every* covering has a finite subcovering and that seems quite a lot of work. How would you do that? Here are a few examples.

Example 17.4 In a metric space (X, d), each finite subset is compact.

Let $F = \{x_1, \ldots, x_n\}$ be a finite subset of X and $\{O_i : i \in I\}$ a covering of F: for each x_i in F there is a set $O(x_i)$ in the covering that contains it. So the n sets $O(x_1), \ldots, O(x_n)$ are a finite subcovering of F. ◁

Example 17.5 In a metric space (X, d), if sequence $(x_k)_{k\in\mathbb{N}}$ has limit x, then the set $C = \{x, x_1, x_2, x_3, \ldots\}$ consisting of this limit and all terms of the sequence is a compact set.

Let $\{O_i : i \in I\}$ be a covering of C. Since limit x belongs to C, there is a set $O(x)$ in the covering that contains x. Set $O(x)$ is open, so x is an interior point: there is an $\varepsilon > 0$ with $B(x, \varepsilon) \subseteq O(x)$. Since $\lim_{k\to\infty} x_k = x$, we know that for this $\varepsilon > 0$ there is an $N \in \mathbb{N}$ such that $k \ge N$ implies $x_k \in B(x, \varepsilon) \subseteq O(x)$. So $O(x)$ contains x and all terms $x_N, x_{N+1}, x_{N+2}, \ldots$ from the N-th term onward. For each of the remaining terms $x_k \in \{x_1, \ldots, x_{N-1}\} \subseteq C$ there is a set $O(x_k)$ in the covering with $x_k \in O(x_k)$. Conclude: $O(x), O(x_1), \ldots, O(x_{N-1})$ is a finite subcovering of C. ◁

Example 17.6 In $\mathbb{R}$ with its usual distance, every closed and bounded interval $[a, b]$ is compact.

Let $\{O_i : i \in I\}$ be a covering of $[a, b]$ and let X consist of all $x \in [a, b]$ such that $[a, x]$ has a finite subcovering. We need to assure that $b \in X$.

Point a is covered by some set in the covering: a belongs to X. Since X is nonempty and bounded from above by b, it has a supremum $s = \sup X \in [a, b]$.

I first show that $[a, s]$ has a finite subcovering. Since a is covered, this is true if $s = a$. If $s > a$, let O_i be a set in the covering containing s. Since O_i is open, there is an $\varepsilon > 0$ with $(s-\varepsilon, s+\varepsilon) \subseteq O_i$. Taking ε sufficiently small, we may assume $a < s - \varepsilon$. Since s is the *least* upper bound of X, $s - \varepsilon$ is not an upper bound: there is an $x \in X$ with $s - \varepsilon < x \le s$. So $[a, x]$ has a finite subcovering. Adding O_i to that subcovering gives a finite subcovering of $[a, s]$. Therefore, $s \in X$.

Next, I show that $s = b$. Suppose $s < b$ and let O_i be a set in the covering that contains s. Since O_i is open, there is an $\varepsilon > 0$ with $(s-\varepsilon, s+\varepsilon) \subseteq O_i$. Taking a finite subcovering of $[a, s]$ and adding set O_i gives a finite subcovering of an interval (like $[a, s + \frac{1}{2}\varepsilon]$) extending to the right of s, contradicting that s is an upper bound of X. Therefore, $s = b$. ◁

Let us proceed by providing properties of sets that *are* compact. First, one more definition:

Definition 17.3 Let (X, d) be a metric space and $U \subseteq X$. The set U is ***totally bounded*** if, for each $\varepsilon > 0$, there is a covering of U by finitely many ε-balls:

$$\text{there is a finite subset } U' \subseteq U \text{ such that } U \subseteq \cup_{u\in U'} B(u, \varepsilon).$$

Equivalently, you can select the finite subset from X instead of U (Exc. 17.2). Every totally bounded set is bounded (Exc. 17.3). But a bounded set need not be totally bounded: see Examples 17.2 and 17.3.

Theorem 17.1

Let (X,d) be a metric space.

(a) Every compact set is closed.

(b) Every closed subset of a compact set is compact.

(c) The union of finitely many compact sets is compact.

(d) The intersection of arbitrarily many compact sets is compact.

(e) Every compact set is totally bounded and (hence) bounded.

(f) Every infinite subset of a compact set has an accumulation point.

Proof: Throughout the proof, let $C \subseteq X$ be compact.
(a) We show that $X \setminus C$ is open. Let $x \in X \setminus C$ (if $X = C$, there is nothing to prove). By the Hausdorff property, for each $c \in C$, there are disjoint neighborhoods $O(c)$ of c and $U(c)$ of x. By compactness of C, there are finitely many $c_1, \ldots, c_n$ with $C \subseteq O(c_1) \cup \cdots \cup O(c_n)$. But then $U = U(c_1) \cap \cdots \cap U(c_n)$ is open, contains x, but $O(c_i) \cap U(c_i) = \emptyset$ for all $i = 1, \ldots, n$. Since C is contained in the union of the $O(c_i)$, it follows that $C \cap U = \emptyset$: x is an interior point of $X \setminus C$.
(b) Let $D \subseteq C$ be closed. If $\{O_i : i \in I\}$ is an open covering of D, then $\{O_i : i \in I\} \cup \{X \setminus D\}$ is an open covering of C. Since C is compact, there is a finite subcovering $\{O_j : j \in J\} \cup \{X \setminus D\}$. Then $\{O_j : j \in J\}$ is a finite subcovering of D. So D is compact.
(c) A covering of the union is a covering of each individual set and the union of the individual finite subcoverings is the finite subcovering we want.
(d) Follows from (a) and (b), since the intersection of compact, hence closed, sets is closed.
(e) Let $\varepsilon > 0$. The open balls $\{B(c,\varepsilon) : c \in C\}$ cover C. A finite subcovering makes C totally bounded.
(f) Let $D \subseteq C$. Suppose D has no accumulation point. We show that D must be finite.
Let $x \in X$. Since x is not an accumulation point of D, there is an $\varepsilon_x > 0$ such that

$$(B(x,\varepsilon_x) \setminus \{x\}) \cap D = \emptyset. \tag{45}$$

C is compact, so covering $\{B(x,\varepsilon_x) : x \in X\}$ has a finite subcovering $\{B(x_1,\varepsilon_{x_1}), \ldots, B(x_k,\varepsilon_{x_k})\}$. By (45), each $B(x_i,\varepsilon_{x_i})$ contains at most one element (x_i is the only candidate) of D, so D must be finite. □

In Theorem 17.1(c), it was shown that the union of finitely many compact sets in a metric space is compact. The union of *infinitely* many compact sets, however, need not be compact. For instance, for each $x \in \mathbb{R}$, the set $\{x\}$ has only one element, so it is compact (Example 17.4). But the union of all these compact sets is $\mathbb{R}$, which is not compact, since it is not bounded (Theorem 17.1(e)).

Theorem 17.2 (Finite intersection property)

Let (X,d) be a metric space and $C \subseteq X$ compact. For each i in a nonempty index set I, let C_i be a closed subset of C. If for each nonempty *finite* subset $F \subseteq I$ of indices:

$$\cap_{i \in F} C_i \neq \emptyset, \tag{46}$$

then the intersection $\cap_{i \in I} C_i$ over *all* indices is nonempty.

Proof: If $\cap_{i\in I} C_i = \emptyset$, then $(\cap_{i\in I} C_i)^c = \cup_{i\in I} C_i^c = X$ gives an open covering of C. Let $\cup_{i\in F} C_i^c$ be a finite subcovering. Then $\cap_{i\in F} C_i = \emptyset$, contradicting (46). □

If you maximize or minimize a continuous, real-valued function over a nonempty, compact set, there is at least one solution. This is an example of an existence theorem: it tells you that a solution exists, not how to find it. Nevertheless, this is a crucial result: it helps to assure that you are not trying to solve a problem in vain.

Theorem 17.3 (Extreme value theorem)

Let (X,d) be a metric space, $C \subseteq X$ nonempty, compact, and $f : C \to \mathbb{R}$ continuous. Then f achieves a minimum and a maximum: there exist $m, M \in C$ such that for all $c \in C$: $f(m) \leq f(c) \leq f(M)$.

Proof: We prove that f has a maximum; the proof for a minimum is analogous. Suppose that f does not achieve a maximum: for each $x \in C$ there is a $y \in C$ with $f(x) < f(y)$, i.e., $x \in L(y) = \{c \in C : f(c) < f(y)\} = f^{-1}((-\infty, f(y))$. By continuity of f, the pre-image $L(y)$ is open. Hence, the collection of sets $\{L(y) : y \in C\}$ is an open covering of C. By compactness of C, there is a finite subset $C' = \{y_1, \ldots, y_k\} \subseteq C$ such that $\{L(y) : y \in C'\}$ covers C. Since C' is finite, it contains a y^* with highest function value: $f(y^*) = \max\{f(y_1), \ldots, f(y_k)\}$. But then $L(y^*)$ covers C. In particular, $y^* \in C \subseteq L(y^*)$, so $f(y^*) < f(y^*)$, a contradiction. □

Notice that we actually proved a stronger result: for the existence of a maximum over the compact set C, it suffices that pre-images of open sets of the form $(-\infty, a)$ (with $a \in \mathbb{R}$) are open. Functions with this property are called ***upper semicontinuous***. (Similarly, for the existence of a minimum over the compact set C, it suffices that pre-images of the open sets (a, ∞) (with $a \in \mathbb{R}$) are open.) By Theorem 11.2, every continuous function is upper semicontinuous. But upper semicontinuity is a much weaker requirement than continuity.

Example 17.7 Function $f : \mathbb{R} \to \mathbb{R}$ with $f(x) = 1$ if $x \geq 0$ and $f(x) = 0$ otherwise is upper semicontinuous:

$$f^{-1}((-\infty, a)) = \begin{cases} \mathbb{R} & \text{if } a > 1, \\ (-\infty, 0) & \text{if } 0 < a \leq 1, \\ \emptyset & \text{if } a \leq 0. \end{cases}$$

Regardless of $a \in \mathbb{R}$, these pre-images are open sets. But f is not continuous at $x = 0$. ◁

The next theorem characterizes compact sets. A subset U of a metric space is ***sequentially compact*** if each sequence in U has a convergent subsequence with a limit in U.

Theorem 17.4 (Characterizations of compactness in metric spaces)

Let (X,d) be a metric space and $U \subseteq X$. The following claims are equivalent:

(a) U is compact,

(b) U is sequentially compact,

(c) U is complete and totally bounded.

Proof: **(a) ⇒ (b)** Let U be compact and suppose there is a sequence $(x_k)_{k\in\mathbb{N}}$ in U without a convergent subsequence. Then for each $u \in U$, there must be an $\varepsilon_u > 0$ such that $B(u, \varepsilon_u)$ contains only finitely

many terms of $(x_k)_{k\in\mathbb{N}}$: otherwise we could construct a subsequence converging to u by choosing points in the sequence in ever smaller balls around u.

But then the balls $\{B(u,\varepsilon_u) : u \in U\}$ are a covering of U without a finite subcovering: any finite subcollection contains only finitely many elements of $\{x_k : k \in \mathbb{N}\} \subseteq U$.

(b) ⇒ (a) Assume that every sequence in U has a convergent subsequence with limit in U and let $\{O_i : i \in I\}$ be a covering of U.

STEP 1: U is totally bounded.

Suppose not: for some $\varepsilon > 0$, the covering $\{B(u,\varepsilon) : u \in U\}$ has no finite subcovering. Let $x_1 \in U$ and for $k \in \mathbb{N}, k > 1$, let $x_k \in U \setminus (B(x_1,\varepsilon) \cup \cdots \cup B(x_{k-1},\varepsilon))$. The sequence $(x_k)_{k\in\mathbb{N}}$ has no convergent subsequence: distinct elements lie at least distance ε apart.

STEP 2: There is an $\varepsilon > 0$ such that for each $x \in U$, there is an $i \in I$ with $B(x,\varepsilon) \subseteq O_i$.

For each $x \in U$ there is a set $O(x)$ in the covering that contains x. Since $O(x)$ is open, there is an $\varepsilon(x) > 0$ with $B(x,2\varepsilon(x)) \subseteq O(x)$. Note that the radius is $2\varepsilon(x)$, not $\varepsilon(x)$. Now $\{B(x,\varepsilon(x)) : x \in U\}$ is a covering of U. By compactness, there is a finite subcovering $\{B(x_1,\varepsilon(x_1)),\ldots,B(x_k,\varepsilon(x_k))\}$. I claim that $\varepsilon = \min\{\varepsilon(x_1),\ldots,\varepsilon(x_k)\}$ works. As the minimum of finitely many positive numbers, it is positive. And for each $x \in U$ there is an x_i with $x \in B(x_i,\varepsilon(x_i))$. So by the triangle inequality:

$$B(x,\varepsilon) \subseteq B(x,r(x_i)) \subseteq B(x_i,2r(x_i)) \subseteq O(x_i).$$

STEP 3: U is totally bounded, so for ε from step 2 there are finitely many $x_1,\ldots,x_n$ with $U \subseteq \cup_{i=1}^n B(x_i,\varepsilon)$. By step 2, each such ball is contained in some O_i, so the corresponding O_i are a finite subcovering of U.

(a) ⇒ (c) Assume U is compact. By Theorem 17.1, U is totally bounded.

To see that U is complete, let $(x_k)_{k\in\mathbb{N}}$ be a Cauchy sequence in U. By our previous step, U is sequentially compact, so the sequence has a convergent subsequence $(x_{k(n)})_{n\in\mathbb{N}}$ with limit $x \in U$. Let's show that the entire sequence converges to x. By convergence of the subsequence and definition of a Cauchy sequence, for each $\varepsilon > 0$, there are $M, N \in \mathbb{N}$ such that

$$d(x_{k(n)},x) < \varepsilon/2 \text{ for all } n \geq M \qquad \text{and} \qquad d(x_k,x_m) < \varepsilon/2 \text{ for all } k,m \geq N.$$

Let $n \geq M$ be such that $k(n) \geq N$. Then for all $k \geq N$:

$$d(x_k,x) \leq d(x_k,x_{k(n)}) + d(x_{k(n)},x) < \varepsilon/2 + \varepsilon/2 = \varepsilon.$$

So $x_k \to x$: the set U is complete.

(c) ⇒ (b) Assume that U is complete and totally bounded. To see that it is sequentially compact, let $x = (x_k)_{k\in\mathbb{N}}$ be a sequence in U. Since U is totally bounded, there is, for each $n \in \mathbb{N}$, a finite set $S_n \subseteq U$ such that $U \subseteq \cup_{u\in S_n} B(u,\frac{1}{n})$. Construct a convergent subsequence of x as follows:

- ☒ For $n = 1$: since S_1 is finite, there is a $u_1 \in S_1$ such that $B(u_1,\frac{1}{1})$ contains infinitely many terms of x. Let $x_{k(1)}$ be one of them.
- ☒ For $n = 2$: since S_2 is finite, there is a $u_2 \in S_2$ such that $B(u_1,\frac{1}{1}) \cap B(u_2,\frac{1}{2})$ contains infinitely many terms of x. Choose $k(2) > k(1)$ such that $x_{k(2)}$ is one of them.
- ☒ In general, for $n > 1$, choose $u_n \in S_n$ such that $\cap_{i=1}^n B(u_i,\frac{1}{i})$ contains infinitely many terms of x and choose $k(n) > k(n-1)$ such that $x_{k(n)}$ is one of them.

If $m > n$, then $d(x_{k(m)},x_{k(n)}) < 1/n$, so the subsequence $(x_{k(n)})_{n\in\mathbb{N}}$ is Cauchy. By completeness, it converges to a point in U. □

We are now equipped to establish the equivalence of the two definitions (Definition 17.1 and 17.2) of compactness of sets in $\mathbb{R}^n$.

Theorem 17.5 (The Heine-Borel theorem)

In $\mathbb{R}^n$ with its usual distance, a set is compact if and only if it is closed and bounded.

Proof: Each compact subset of $\mathbb{R}^n$ is closed and bounded by Theorem 17.1. Conversely, let C be closed and bounded. Consider a sequence in C. It is bounded, since C is, so it has a convergent subsequence by Theorem 13.2. C is closed, so its limit lies in C by Theorem 13.4. By Theorem 17.4, C is compact. □

Theorem 17.6

Let (X,d) and (Y,d') be metric spaces, $C \subseteq X$ compact, and $f : C \to Y$ continuous. Then:

(a) Continuous functions map compact sets to compact sets: $f(C)$ is compact;

(b) f is uniformly continuous.

Proof: **(a)** Let $\{O_i : i \in I\}$ be a covering of $f(C)$. By continuity of f, the pre-images $f^{-1}(O_i)$ are open. Thus $\{f^{-1}(O_i) : i \in I\}$ is a covering of C. By compactness of C, it contains a finite subcovering $\{f^{-1}(O_i) : i \in J\}$ and hence $\{O_i : i \in J\}$ is a finite subcovering of $f(C)$: $f(C)$ is compact.
(b) Let $\varepsilon > 0$. By continuity of f, there exists, for each $x \in C$, a $\delta(x) > 0$ such that

$$\text{if } y \in C \text{ and } d(x,y) < \delta(x), \text{ then } d'(f(x),f(y)) < \frac{1}{2}\varepsilon. \tag{47}$$

Since $x \in B\left(x, \frac{1}{2}\delta(x)\right)$, the collection $\left\{B\left(x, \frac{1}{2}\delta(x)\right) : x \in C\right\}$ covers C. By compactness of C, there exist finitely many $x_1, \ldots, x_n \in C$ such that $\left\{B\left(x_i, \frac{1}{2}\delta(x_i)\right) : i = 1, \ldots, n\right\}$ covers C.

Let $\delta = \min\left\{\frac{1}{2}\delta(x_1), \ldots, \frac{1}{2}\delta(x_n)\right\}$. We show that

$$\text{for all } y, z \in C, \text{ if } d(y,z) < \delta, \text{ then } d'(f(y),f(z)) < \varepsilon.$$

If $d(y,z) < \delta$, pick $i \in \{1, \ldots, n\}$ such that $y \in B(x_i, \frac{1}{2}\delta(x_i))$. Then

$$d(y,x_i) < \frac{1}{2}\delta(x_i) < \delta(x_i),$$
$$d(x_i,z) \leq d(x_i,y) + d(y,z) < \frac{1}{2}\delta(x_i) + \delta \leq \delta(x_i).$$

Hence, using the triangle inequality and (47):

$$d'(f(y),f(z)) \leq d'(f(y),f(x_i)) + d'(f(x_i),f(z)) \leq \frac{1}{2}\varepsilon + \frac{1}{2}\varepsilon = \varepsilon.$$

□

Finally, let us state — without proof — the following useful theorem:

Theorem 17.7 (Tychonoff's theorem; the product of compact sets is compact)

If A and B are compact subsets in $\mathbb{R}^m$ and $\mathbb{R}^n$ respectively, then $A \times B$ is compact in $\mathbb{R}^{m+n}$.

And you can use this repeatedly to conclude that for instance the product $A \times B \times C$ of compact subsets of $\mathbb{R}^k$, $\mathbb{R}^\ell$, and $\mathbb{R}^m$ is a compact subset of $\mathbb{R}^{k+\ell+m}$.

Example 17.8 Since the interval $[0,1]$ is compact in $\mathbb{R}$, the Cartesian product $[0,1] \times \cdots \times [0,1]$ of n of these sets is a compact subset of $\mathbb{R}^n$. ◁

Exercises section 17

17.1 Use Definition 17.1 to check whether the following sets are compact:

(a) $\{x \in \mathbb{R} : 0 \le x < 5\}$,

(b) the subset of $\mathbb{R}$ consisting of the integers 5 and 37,

(c) the union of the sets in (a) and (b),

(d) $\{x \in \mathbb{R}^2 : x_1 x_2 = 1\}$,

(e) $\{x \in \mathbb{R}^2 : 3x_1 - 2x_2 \le 6, x_1 + x_2 \le 3, x_1 \ge 0, x_2 \ge 0\}$,

(f) $\{x \in \mathbb{R}^2 : 0 \le x_1 \le 3, 1 \le x_2 < 4\}$,

(g) $\{x \in \mathbb{R}^3 : x_3 - x_1 x_2 = 5\}$,

(h) $\{x \in \mathbb{R}^3 : x_1^2 + 4x_2^2 = 4, x_1 + 2x_3 = 2\}$,

(i) $\{x \in \mathbb{R}^2 : 0 \le x_2 \le x_1^3, x_2 \ge 2x_1^3 - 6x_1^2 + 12x_1 - 8\}$.

17.2 Let U be a subset of metric space (X, d). Show that the following two are equivalent:

(a) For each $\varepsilon > 0$ there is a finite subset U' of U such that $U \subseteq \cup_{u \in U'} B(u, \varepsilon)$.

(b) For each $\varepsilon > 0$ there is a finite subset U' of X such that $U \subseteq \cup_{u \in U'} B(u, \varepsilon)$.

17.3 Show that every totally bounded set in a metric space is bounded.

17.4 Show that in $\mathbb{R}^n$ with its usual distance, a set U is bounded if and only if it is totally bounded.

17.5 Show that if U is a bounded subset of $\mathbb{R}^n$, then its closure $\mathrm{cl}(U)$ is compact.

17.6 The set V of functions in $C[0,1]$ of the form $f(x) = a_1 x + a_2$ with slope $a_1 \in [-1,4]$ and intercept $a_2 \in [2,3]$ is compact.

We show this using Theorem 17.6(a), which says that continuous functions map compact sets to compact sets. For instance, if we define the function $F : \mathbb{R}^2 \to C[0,1]$ that assigns to each $a = (a_1, a_2) \in \mathbb{R}^2$ the function $F(a) = f_a$ with $f_a(x) = a_1 x + a_2$, we see that V is precisely the set of functions you get from substituting vectors $a = (a_1, a_2) \in [-1,4] \times [2,3]$ into the function F:

$$V = F([-1,4] \times [2,3]).$$

So if we show that $[-1,4] \times [2,3]$ is compact and F is continuous, Theorem 17.6(a) says that V is compact.

(a) Let $a = (a_1, a_2)$ and $b = (b_1, b_2)$ belong to $\mathbb{R}^2$. Use the triangle inequality to show that the functions $f_a(x) = a_1 x + a_2$ and $f_b(x) = b_1 x + b_2$ satisfy, for all $x \in [0,1]$:

$$|f_b(x) - f_a(x)| \le |b_1 - a_1| + |b_2 - a_2|.$$

(b) Use this to show that $d_\infty(f_b, f_a) \le 2\|b - a\|_2$.

(c) Use this to show that the function $F : \mathbb{R}^2 \to C[0,1]$ is continuous.

(d) Why is $[-1,4] \times [2,3]$ compact? Use Theorem 17.6(a) to show that V is compact.

17.7 Let X be any set with at least two elements. Assume that the only open subsets of X are the empty set $\emptyset$ and X itself. (Mathematicians often call X the ***indiscrete space*** and the collection of open sets $\{\emptyset, X\}$ the ***trivial topology***; it is a common source of unexpected results.) Which subsets of X are closed? And which are compact?

18 The Stone-Weierstrass approximation theorem

Our next result gives the justification for common nonparametric regression models in econometrics and statistics, where a dependent variable y is modeled as a function $y = f(x)$ of some independent variable(s) x without knowing a priori what function f looks like. In practice, econometric models tend to use easy candidates for f, like polynomial functions. The Stone-Weierstrass approximation theorem tells why such a simplification is okay. It gives conditions under which each continuous real-valued function can be approximated arbitrarily well by one from a smaller collection F of functions. In our earlier terminology, that smaller collection F is dense in the larger set of all continuous functions:

Theorem 18.1 (The Stone-Weierstrass approximation theorem)

Let (X, d) be a compact metric space and F a collection of continuous functions from X to $\mathbb{R}$ that satisfies:

- ☒ for all g and h in F, also their sum $g+h$, their product gh, and scalar multiples αg (for each $\alpha \in \mathbb{R}$) belong to F,
- ☒ the constant function $\mathbf{1} : X \to \mathbb{R}$ with $\mathbf{1}(x) = 1$ for all x in X belongs to F,
- ☒ for all distinct x' and x'' in X there is a function g in F with $g(x') \neq g(x'')$.

Then F is dense in $(C(X,\mathbb{R}), d_\infty)$.

Even the most elementary proof I'm aware of[5] takes several pages; I'll skip it. The message is that in simple regression, with one explanatory variable, there is little loss of generality in restricting attention to polynomial functions:

Example 18.1 Each continuous function $f : [a,b] \to \mathbb{R}$ on a compact interval $X = [a,b]$ of real numbers can be approximated arbitrarily well by a polynomial function. The set F of polynomial functions p, i.e., those of the form

$$p(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_k x^k \tag{48}$$

for a nonnegative integer k and constants $a_0, \ldots, a_k \in \mathbb{R}$ satisfies the three conditions of the Stone-Weierstrass theorem: doing the corresponding calculations, you see that the sum, product, and scalar multiples of polynomial functions are again polynomial. Also the constant function that is everywhere equal to one is a polynomial: in (48), take $a_0 = 1$ and set all other coefficients a_i equal to zero. Finally, the function g with $g(x) = x$ is a polynomial and for all distinct x' and x'' in $[a,b]$ it satisfies $g(x') = x' \neq x'' = g(x'')$.

So this set F of polynomial functions is dense in $(C([a,b],\mathbb{R}), d_\infty)$: for each continuous function $f : [a,b] \to \mathbb{R}$ and each $\varepsilon > 0$, no matter how small, you can find a polynomial function $p \in F$ such that

$$d_\infty(f,p) = \sup_{x \in [a,b]} |f(x) - p(x)| < \varepsilon.$$

In words: for each x in the domain, $f(x)$ and $p(x)$ are less than ε apart. ◁

This last point is worth stressing: the Stone-Weierstrass theorem is about approximating a function using the supremum metric, so the approximation provides a good fit *at each point in the domain*. This is sometimes called 'uniform approximation' and is in contrast with the more familiar type of Taylor approximations which only guarantee a good local fit near a specific point in the domain.

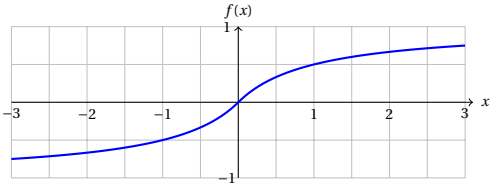
[5] B. Brosowski and F. Deutsch (1981), An elementary proof of the Stone-Weierstrass Theorem. Proceedings of the American Mathematical Society 81, 89–92.

Our example extends to functions of multiple variables: each continuous $f : X \to \mathbb{R}$ on a compact subset X of $\mathbb{R}^n$ has a good polynomial approximation, because the set F of all linear combinations of polynomial terms

$$(x_1, \ldots, x_n) \mapsto x_1^{k_1} x_2^{k_2} \cdots x_n^{k_n} \qquad (k_1, \ldots, k_n \in \{0, 1, 2, \ldots\})$$

satisfies the conditions of the Stone-Weierstrass theorem and is therefore dense.

The compactness assumption in the Stone-Weierstrass theorem is important. In contrast with our earlier example, polynomial functions no longer assure good approximations to continuous functions whose domain X is not compact. Look at the function $f : \mathbb{R} \to \mathbb{R}$ with $f(x) = \frac{x}{|x|+1}$ whose graph is drawn to the right. You can't find a polynomial that provides a close fit at each point in its domain. Constant polynomials don't work, since the function tends to -1 as $x \to -\infty$ but to a different value 1 as $x \to \infty$. And nonconstant polynomials diverge to plus or minus infinity as $x \to \infty$, so they don't give a good approximation to our bounded function either. Exercise 18.1 investigates why we need the three conditions on the collection F of candidate approximations.



Exercises section 18

18.1 For each of the following collections F of functions from $[0,1]$ to $\mathbb{R}$, *(i)* show that it violates at least one of the conditions in the Stone-Weierstrass theorem and *(ii)* argue why it isn't dense in $(C([0,1],\mathbb{R}), d_\infty)$ by finding a continuous function $f : [0,1] \to \mathbb{R}$ that doesn't have a good approximation in F.

(a) All continuous functions g with $g(0) = 0$;

(b) All constant functions;

(c) All continuous functions g with $g(0) = g(1)$;

(d) All affine functions, i.e., those of the form $g(x) = ax + b$ for some parameters a and b in $\mathbb{R}$.

19 Convex sets

Convex sets and functions play a crucial role in mathematical economics and decision theory, for instance in linear and nonlinear optimization, game theory, and the analysis of economic equilibria. We provide some of the necessary mathematical background.

19.1 Convex sets

Definition 19.1 A set C in a vector space X is ***convex*** if for each pair of elements $x, y \in C$, the entire line segment between x and y belongs to C:

$$\text{for all } x, y \in C \text{ and all } \lambda \in [0,1]: \qquad \lambda x + (1-\lambda) y \in C.$$

λ is the Greek letter 'lambda'. Since a vector space is closed under scalar multiplication, λx and $(1-\lambda) y$ are well-defined; since it is closed under addition, so is their sum $\lambda x + (1-\lambda) y$. The linear combination $\lambda x + (1-\lambda) y$ with nonnegative weights adding up to one is called a convex combination of x and y; this extends easily to more than two terms:

Definition 19.2 A vector $x \in X$ is a ***convex combination*** of elements of a set $C \subseteq X$ if there is a finite number $m \in \mathbb{N}$ of vectors $v_1, \ldots, v_m \in C$ and scalars $\lambda_1, \ldots, \lambda_m \geq 0$ with $\sum_{i=1}^m \lambda_i = 1$ such that

$$x = \lambda_1 v_1 + \cdots + \lambda_m v_m. \tag{49}$$

Note that each convex combination of $m > 2$ terms can be rewritten as a convex combination of two vectors of fewer terms: since the λ_i are nonnegative and sum to one, at least one of them, say λ_1 is smaller than one and we can write

$$\sum_{i=1}^m \lambda_i v_i = \lambda_1 v_1 + (1-\lambda_1) \sum_{i=2}^m \tfrac{\lambda_i}{1-\lambda_1} v_i.$$

By induction, we have the first part of the next theorem; its second part involves a straightforward check of the definition.

Theorem 19.1

(a) A set is convex if and only if it contains all convex combinations of its elements.

(b) The intersection of convex sets is a convex set.

Now let $S \subseteq X$ be an arbitrary, not necessarily convex, set in vector space X. There is at least one convex set containing S, namely X. Moreover, the intersection of all convex sets containing S is a convex set and consequently the *smallest* convex set containing S. This makes the following well-defined:

Definition 19.3 The ***convex hull*** $\operatorname{conv}(S)$ of a set $S \subseteq X$ is the smallest convex set containing S. It is the intersection of all convex sets containing S: $\operatorname{conv}(S) = \cap_{C \subseteq X : C \text{ is convex, } S \subseteq C} C$.

By Theorem 19.1, $\operatorname{conv}(S)$ is simply the set of all convex combinations of elements of S: the set of convex combinations of elements of S is itself a convex set containing S. Since every other convex set containing S must also include these convex combinations, it is the smallest convex set containing S.

A ***polytope*** is the convex hull of a *finite* set $S = \{v_1, \ldots, v_m\}$ of vectors, in which case

$$\text{conv}(S) = \text{conv}(\{v_1, \ldots, v_m\}) = \{\lambda_1 v_1 + \cdots + \lambda_m v_m : \lambda_1, \ldots, \lambda_m \geq 0, \sum_i \lambda_i = 1\}.$$

Here are a few other examples of convex sets:

Example 19.1 (Hyperplanes and halfspaces) A hyperplane is the set of solutions to a single linear equation. Similarly, a halfspace is the set of solutions to a single linear inequality. Formally, a ***hyperplane*** in $\mathbb{R}^n$ is a set of the form

$$\{x \in \mathbb{R}^n : c^\top x = \delta\} \qquad \text{for some vector } c \in \mathbb{R}^n, c \neq \mathbf{0}, \text{ and a number } \delta \in \mathbb{R}. \tag{50}$$

Vector c is referred to as the ***normal*** of the hyperplane. A ***halfspace*** consists of the points 'on one side' of a hyperplane, i.e., it is a set of the form

$$\{x \in \mathbb{R}^n : c^\top x \leq \delta\} \qquad \text{for some vector } c \in \mathbb{R}^n, c \neq \mathbf{0}, \text{ and a number } \delta \in \mathbb{R}.$$

Sometimes, hyperplanes and halfspaces are referred to as affine if $\delta \neq 0$ and linear if $\delta = 0$. For instance, in $\mathbb{R}^2$,

$$\{x \in \mathbb{R}^2 : 3x_1 + 4x_2 = 12\} \text{ is a hyperplane,} \qquad \{x \in \mathbb{R}^2 : 3x_1 + 4x_2 \leq 12\} \text{ is a halfspace.}$$

As pre-images of the closed sets $\{\delta\}$ and $(-\infty, \delta]$ under the continuous function $x \mapsto c^\top x$, hyperplanes and halfspaces are closed sets. They are convex as well. We prove this for hyperplanes; the proof for halfspaces is analogous. Let $x, y \in \mathbb{R}^n$ lie in the hyperplane (50): they satisfy $c^\top x = \delta$ and $c^\top y = \delta$. Let $\lambda \in [0,1]$. Then

$$c^\top(\lambda x + (1-\lambda)y) = c^\top(\lambda x) + c^\top((1-\lambda)y) = \lambda c^\top x + (1-\lambda)c^\top y = \lambda\delta + (1-\lambda)\delta = \delta,$$

so $\lambda x + (1-\lambda)y$ lies in the hyperplane as well. ◁

Example 19.2 (Balls in normed vector spaces) Each ball $B(v, \varepsilon)$ in a normed vector space X is convex: if $x, y \in B(v, \varepsilon)$ and $\lambda \in [0,1]$, then

$$\begin{aligned}\|\lambda x + (1-\lambda)y - v\| = \|\lambda(x-v) + (1-\lambda)(y-v)\| &\overset{(N4)}{\leq} \|\lambda(x-v)\| + \|(1-\lambda)(y-v)\| \\ &\overset{(N3)}{=} \lambda\|x-v\| + (1-\lambda)\|y-v\| < \lambda\varepsilon + (1-\lambda)\varepsilon = \varepsilon,\end{aligned}$$

so $\lambda x + (1-\lambda)y \in B(v, \varepsilon)$. ◁

Theorem 19.2

If C is a convex subset of a normed vector space X, then also its closure $\text{cl}(C)$ and its interior $\text{int}(C)$ are convex.

Proof: **(Closure)** If $\text{cl}(C)$ is empty, it is convex. So assume it is nonempty and let $x, y \in \text{cl}(C)$ and $\lambda \in [0,1]$. Since $x, y \in \text{cl}(C)$, there are sequences $(x_k)_{k\in\mathbb{N}}$ and $(y_k)_{k\in\mathbb{N}}$ in C with $\lim_{k\to\infty} x_k = x$ and $\lim_{k\to\infty} y_k = y$. For each $k \in \mathbb{N}$, $\lambda x_k + (1-\lambda)y_k \in C$ by convexity of C. Moreover,

$$\lim_{k\to\infty}\big(\lambda x_k + (1-\lambda)y_k\big) = \lambda \lim_{k\to\infty} x_k + (1-\lambda)\lim_{k\to\infty} y_k = \lambda x + (1-\lambda)y,$$

so $\lambda x + (1-\lambda)y \in \text{cl}(C)$.
(Interior) If $\text{int}(C)$ is empty, it is convex. So assume it is nonempty and let $x, y \in \text{int}(C)$ and $\lambda \in [0,1]$. Since $x, y \in \text{int}(C)$, there is an $\varepsilon > 0$ such that the open balls $B(x, \varepsilon)$ and $B(y, \varepsilon)$ are contained in C. We

show that also the ball $B(z,\varepsilon)$ around $z = \lambda x + (1-\lambda)y$ is contained in C. Let $v \in B(z,\varepsilon)$. To show that $v \in C$, write

$$v = z + (v-z) = \lambda \underbrace{[x + (v-z)]}_{\in B(x,\varepsilon)\subseteq C} + (1-\lambda) \underbrace{[y + (v-z)]}_{\in B(y,\varepsilon)\subseteq C},$$

which lies in C since it is a convex combination of elements of the convex set C. □

19.2 Polyhedra and Fourier-Motzkin elimination

Definition 19.4 A ***polyhedron*** or ***polyhedral set*** is a set $P \subseteq \mathbb{R}^n$ of solutions to a system of finitely many linear inequalities:

$$P = \{x \in \mathbb{R}^n : Ax \leq b\} \qquad \text{for some matrix } A \in \mathbb{R}^{m\times n} \text{ and some vector } b \in \mathbb{R}^m.$$

As the intersection of halfspaces, one for each linear inequality, a polyhedron is closed and convex. The definition of a polyhedron is easy enough, but how would you go about actually *finding* the solutions to a system of linear inequalities? Fourier-Motzkin elimination is a tool to solve systems of linear inequalities $Ax \leq b$ by removing unknowns one at a time, similar to Gaussian elimination for solving systems of linear equations $Ax = b$. It is probably best illustrated using an example.

Example 19.3 Let us solve the system of linear inequalities

$$-2x_1 - x_2 \leq -2 \tag{51}$$
$$3x_1 + x_2 \leq 9 \tag{52}$$
$$-x_1 + 2x_2 \leq 4 \tag{53}$$
$$-x_2 \leq 0 \tag{54}$$

by eliminating x_1. In inequalities (51) and (53), x_1 has a negative coefficient. They impose lower bounds on x_1:

$$1 - \tfrac{1}{2}x_2 \leq x_1$$
$$-4 + 2x_2 \leq x_1$$

Inequality (52), where x_1 has a positive coefficient, imposes an upper bound on x_1:

$$x_1 \leq 3 - \tfrac{1}{3}x_2.$$

Inequality (54), where x_1 does not appear or (fancy!) has zero coefficient, imposes no bounds on x_1:

$$-x_2 \leq 0.$$

We can squeeze in an x_1 between the upper and lower bounds if and only if the lower bounds on x_1 do not exceed any of the upper bounds on x_1. Moreover, we need to append $-x_2 \leq 0$. In other words, there is a solution (x_1, x_2) to our system of inequalities if and only if

$$1 - \tfrac{1}{2}x_2 \leq 3 - \tfrac{1}{3}x_2$$
$$-4 + 2x_2 \leq 3 - \tfrac{1}{3}x_2$$
$$-x_2 \leq 0$$

has a solution. Rearrange terms:

$$\begin{aligned} -\tfrac{1}{6}x_2 &\le 2 \\ \tfrac{7}{3}x_2 &\le 7 \\ -x_2 &\le 0 \end{aligned}$$

Now repeat the same steps to get rid of x_2: the inequalities where x_2 has a positive coefficient provide an upper bound and those where x_2 has a negative coefficient provide a lower bound:

$$\begin{aligned} -12 &\le x_2 \\ x_2 &\le 3 \\ 0 &\le x_2 \end{aligned}$$

Clearly, the latter system has a solution, because the lower bounds (0 and -12) do not exceed the upper bound (3). Also, we see that the feasible candidates for x_2 lie between 0 and 3. Substituting this back into the lower and upper bounds on x_1, we find the feasible candidates for x_1. To summarize, the set of solutions to our system of linear inequalities consists of all $x \in \mathbb{R}^2$ with $0 \le x_2 \le 3$ and

$$\max\{1 - \tfrac{1}{2}x_2, -4 + 2x_2\} \le x_1 \le 3 - \tfrac{1}{3}x_2. \qquad \triangleleft$$

The technical lingo is:

Theorem 19.3 (Fourier-Motzkin elimination)

Consider the projection $\pi : \mathbb{R}^n \to \mathbb{R}^{n-1}$ that omits the first coordinate:

$$\pi(x_1, \ldots, x_n) = (x_2, \ldots, x_n).$$

If $P \subseteq \mathbb{R}^n$ is a polyhedron, then $\pi(P)$ is a polyhedron.

Proof: Let $P = \{x : Ax \le b\}$ for some $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$. Divide the m inequalities into three sets, depending on whether the coefficient a_{i1} of the unknown x_1 in equation i is (L)ess than, (E)qual to, or (G)reater than zero:

$$L = \{i : a_{i1} < 0\}, \quad E = \{i : a_{i1} = 0\}, \quad G = \{i : a_{i1} > 0\}.$$

Inequalities in L give lower bounds on x_1: inequality $i \in L$ can be rewritten as

$$-\frac{a_{i2}}{a_{i1}}x_2 - \cdots - \frac{a_{in}}{a_{i1}}x_n + \frac{b_i}{a_{i1}} \le x_1. \tag{55}$$

Inequalities in G give upper bounds on x_1: inequality $j \in G$ can be rewritten as

$$x_1 \le -\frac{a_{j2}}{a_{j1}}x_2 - \cdots - \frac{a_{jn}}{a_{j1}}x_n + \frac{b_j}{a_{j1}}. \tag{56}$$

Inequalities in E don't involve x_1 and impose no restrictions on x_1. Now $(x_2, \ldots, x_n) \in \pi(P)$ if and only if there is an x_1 with $(x_1, x_2, \ldots, x_n) \in P$. This happens if and only if $(x_2, \ldots, x_n)$ satisfies the inequalities in E and we can 'squeeze in' a number x_1 between the lower bounds in (55) and the upper bounds in (56). Equivalently, the inequalities in E must hold, as well as the inequalities

$$-\frac{a_{i2}}{a_{i1}}x_2 - \cdots - \frac{a_{in}}{a_{i1}}x_n + \frac{b_i}{a_{i1}} \le -\frac{a_{j2}}{a_{j1}}x_2 - \cdots - \frac{a_{jn}}{a_{j1}}x_n + \frac{b_j}{a_{j1}}, \tag{57}$$

for each $i \in L$ and $j \in G$. Letting n_L, n_E, n_G denote the number of inequalities in L, E, G, respectively, we see that $\pi(P)$ is the set of solutions to a system of $n_L n_G + n_E$ linear inequalities: it is a polyhedron! $\square$

Iteratively applying this to a system of linear inequalities, eliminating variables one at a time, gives a method to check whether the system has a solution and an explicit way to find them.

19.3 Convex cones

Definition 19.5 A ***convex cone*** is a set $C \subseteq \mathbb{R}^n$ that is:

- ⊠ closed under addition: for all $x, y \in C : x + y \in C$;
- ⊠ closed under rescaling by a nonnegative scalar: for all $x \in C$ and all real numbers $\lambda \geq 0 : \lambda x \in C$.

A convex cone C — otherwise its name would be pretty bizarre — really is a convex set: if $x, y \in C$ and $\lambda \in [0,1]$, than λx and $(1-\lambda)y$ belong to C by the second property and so does their sum $\lambda x + (1-\lambda)y$, by the first property.

Recall from earlier definitions that

- ⊠ a linear combination assigns arbitrary real "weights" to a finite number of vectors,
- ⊠ a convex combination assigns nonnegative "weights" to a finite number of vectors, with the weights adding up to one.

Here is a third variant on this theme: arbitrary nonnegative weights!

Theorem 19.4

Set $C \subseteq \mathbb{R}^n$ is a convex cone if and only if it contains all nonnegative combinations of its elements:

$$\text{for all } m \in \mathbb{N}, \text{ all } v_1, \ldots, v_m \in C, \text{ all real numbers } \lambda_1, \ldots, \lambda_m \geq 0: \qquad \lambda_1 v_1 + \cdots + \lambda_m v_m \in C.$$

Proof: Exercise 19.1; induction is your friend. □

Now mimic the discussion after Theorem 19.1: let $S \subseteq \mathbb{R}^n$ be an arbitrary set. There is at least one convex cone containing S, namely $\mathbb{R}^n$. Moreover, the intersection of all convex cones containing S is once again a convex cone: since the properties in Definition 19.5 hold for each convex cone containing S, they hold for their intersection. Consequently, the intersection of all convex cones containing S is the smallest convex cone containing S. This makes the following notion well-defined.

Definition 19.6

- ⊠ The ***convex cone*** $\text{cone}(S)$ ***generated by*** a set $S \subseteq \mathbb{R}^n$ is the smallest convex cone containing S.
- ⊠ A convex cone is ***finitely generated*** if it is generated by a set with finitely many elements.

By Theorem 19.4, every finitely generated cone is of the form

$$\text{cone}(\{v_1, \ldots, v_m\}) = \{\lambda_1 v_1 + \cdots + \lambda_m v_m : \lambda_1, \ldots, \lambda_m \in \mathbb{R}, \lambda_1, \ldots, \lambda_m \geq 0\} \tag{58}$$

for some set $\{v_1, \ldots, v_m\}$ of m vectors in $\mathbb{R}^n$. If we define V to be the $n \times m$ matrix with columns $v_1, \ldots, v_m$ and $\lambda = (\lambda_1, \ldots, \lambda_m)$, then (58) can be rewritten as

$$\text{cone}(\{v_1, \ldots, v_m\}) = \{x \in \mathbb{R}^n : \text{there is a } \lambda \in \mathbb{R}^m \text{ with } x = V\lambda \text{ and } \lambda \geq \mathbf{0}\}. \tag{59}$$

Our next theorem says that this can be rewritten as the set of solutions to a system of linear inequalities. For instance, the cone generated by

$$v_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \qquad \text{and} \qquad v_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

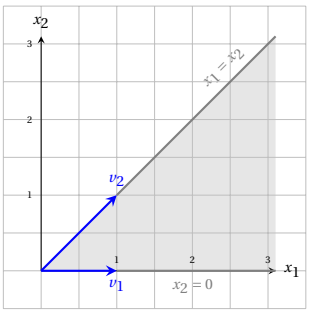
is the shaded area in the figure below; it is the set

$$\{x \in \mathbb{R}^2 : x_2 \geq 0, x_1 - x_2 \geq 0\} \tag{60}$$

of points on/above the line $x_2 = 0$ and on/below the line $x_1 = x_2$. This can be proved with Fourier-Motzkin elimination: $x \in \mathbb{R}^2$ lies in the finitely generated cone if and only if there are λ_1 and λ_2 with

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \lambda_1 \begin{bmatrix} 1 \\ 0 \end{bmatrix} + \lambda_2 \begin{bmatrix} 1 \\ 1 \end{bmatrix} \tag{61}$$

$$\lambda_1, \lambda_2 \geq 0 \tag{62}$$



To figure out what restrictions this imposes on x_1 and x_2 we use Fourier-Motzkin elimination to consecutively eliminate λ_2 and λ_1. The first row of (61) gives $\lambda_2 = x_1 - \lambda_1$ and the second that $\lambda_2 = x_2$. Together with (62) it follows that there is a solution $x_1, x_2, \lambda_1, \lambda_2$ if and only if x_1, x_2, λ_1 satisfy

$$\begin{array}{rcl} x_1 - \lambda_1 & = & x_2 \\ x_2 & \geq & 0 \\ \lambda_1 & \geq & 0 \end{array} \qquad \text{or, after rearranging,} \qquad \begin{array}{rcl} \lambda_1 & = & x_1 - x_2 \\ \lambda_1 & \geq & 0 \\ x_2 & \geq & 0 \end{array}$$

And eliminating λ_1, this has a solution if and only if x_1 and x_2 satisfy $x_2 \geq 0$ and $x_1 - x_2 \geq 0$, as in (60).

I kept the equalities to simplify calculations. In the proof of Theorem 19.5, I rewrote it in terms of linear inequalities (rewriting an equality $a = b$ as two inequalities $a \leq b$ and $b \leq a$) because that is how Fourier-Motzkin elimination is usually formulated. But of course these two approaches are equivalent.

Theorem 19.5

A finitely generated cone is a polyhedron (and hence closed).

Proof: Let $C \subseteq \mathbb{R}^n$ be a finitely generated cone. By (59), we can write

$$\begin{array}{rcl} C & = & \{x \in \mathbb{R}^n : \text{there is a } \lambda \in \mathbb{R}^m \text{ with } x = V\lambda \text{ and } \lambda \geq \mathbf{0}\} \\ & = & \{x \in \mathbb{R}^n : \text{there is a } \lambda \in \mathbb{R}^m \text{ with } x - V\lambda \leq \mathbf{0}, -x + V\lambda \leq \mathbf{0}, -\lambda \leq \mathbf{0}\}. \end{array}$$

This is the projection of the polyhedron

$$P = \{(x, \lambda) \in \mathbb{R}^{n+m} : x - V\lambda \leq \mathbf{0}, -x + V\lambda \leq \mathbf{0}, -\lambda \leq \mathbf{0}\} \tag{63}$$

by projecting away the m coordinates of λ one at a time. By Fourier-Motzkin elimination (Theorem 19.3), C is a polyhedron:

$$C = \{x : Ax \leq \mathbf{0}\} \qquad \text{for some matrix } A.$$

The righthand side of the linear inequalities must be the zero vector: this follows using inequality (57) and the fact that the righthand sides in (63) are all zero. □

So it is easy to give examples of cones that are not finitely generated, for instance, because they are not closed. A nonempty set that is both a polyhedron and a convex cone is called a ***polyhedral cone***. Such sets are always of the form $\{x : Ax \leq \mathbf{0}\}$ for some matrix A; see Exercise 19.3.

Exercises section 19

19.1 Prove Theorem 19.4.

19.2 Solve the following system of linear inequalities using Fourier-Motzkin elimination:

$$x_1 - x_2 \leq 0, \qquad x_1 - x_3 \leq 0, \qquad -x_1 + x_2 + 2x_3 \leq 2, \qquad -x_3 \leq -1.$$

19.3 Show that a nonempty set $P \subseteq \mathbb{R}^n$ is a polyhedral cone if and only if $P = \{x \in \mathbb{R}^n : Ax \leq \mathbf{0}\}$ for some $m \times n$ matrix A.

19.4 Show that a polytope in $\mathbb{R}^n$ is a polyhedron and consequently closed.

20 Farkas' Lemma and some variants

20.1 Farkas' lemma and Gordan's theorem

We prove a geometric variant of Farkas' Lemma. It states: given a convex cone generated by the columns of a matrix $A \in \mathbb{R}^{m\times n}$ and given a vector $b \in \mathbb{R}^m$, there are two mutually exclusive possibilities:

1. b belongs to the cone, in which case there are nonnegative coefficients for the columns of A to represent b;
2. b does not belong to the cone, in which case we can find a hyperplane (whose normal we call y) such that the cone lies on one side and b on the other.

Theorem 20.1 (Farkas' Lemma)

Exactly one of the following two problems has a solution:

(i) $Ax = b, x \geq \mathbf{0}$;

(ii) $y^\top A \geq \mathbf{0}^\top, y^\top b < 0$.

Proof: Firstly, (i) and (ii) cannot *both* have a solution: if $x \geq \mathbf{0}$ satisfies $Ax = b$ and y satisfies $y^\top A \geq \mathbf{0}$, then $(y^\top A)x$ is the inner product of nonnegative vectors, so

$$0 \leq (y^\top A)x = y^\top (Ax) = y^\top b.$$

Secondly, if (i) has no solution, then (ii) does: let $C = \{Ax : x \geq \mathbf{0}\}$ be the convex cone that is finitely generated by the columns of A. By Theorem 19.5, C is polyhedral:

$$C = \{x : Bx \leq \mathbf{0}\} \qquad \text{for some matrix } B.$$

No solution to (i) means $b \notin C$. Hence there is a row (denoted by vector $\hat{y}$) of B with $\hat{y}^\top b > 0$. On the other hand, each column Ae_j of A *does* belong to C, since $e_j \geq \mathbf{0}$. In particular, $\hat{y}^\top (Ae_j) = (\hat{y}^\top A)e_j \leq 0$. So $\hat{y}^\top A \leq \mathbf{0}^\top$. Let $y = -\hat{y}$. Then $y^\top b < 0$ and $y^\top A \geq \mathbf{0}^\top$. □

A variant that is useful in our treatment of optimization problems later is:

Theorem 20.2 (Gordan's theorem)

Given $k \in \mathbb{N}$ vectors $v_1, \ldots, v_k$ in $\mathbb{R}^n$, exactly one of the following is true:

(a) There is a vector d in $\mathbb{R}^n$ with

$$\begin{aligned} v_1^\top d &> 0, \\ &\vdots \\ v_k^\top d &> 0. \end{aligned}$$

(b) There are nonnegative numbers $\mu_1, \ldots, \mu_k$, not all equal to zero, with

$$\mu_1 v_1 + \cdots + \mu_k v_k = \mathbf{0}.$$

Proof: Claims (a) and (b) cannot both be true: if $v_i^\top d > 0$ for all $i = 1, \ldots, k$, then multiplication with nonnegative scalars μ_i, not all zero, gives

$$(\mu_1 v_1 + \cdots + \mu_k v_k)^\top d = \mu_1 (v_1^\top d) + \cdots + \mu_k (v_k^\top d) > 0.$$

Hence $\mu_1 v_1 + \cdots + \mu_k v_k$ cannot be the zero vector: the inner product of d and the zero vector is zero.

We now show that if (b) is false, then (a) is true. Let (b) be false. Then for each $i = 1, \ldots, k$, vector $-v_i$ does not lie in the cone generated by $v_1, \ldots, v_k$: if, to the contrary, there were nonnegative scalars $\alpha_1, \ldots, \alpha_k$ with $\alpha_1 v_1 + \cdots + \alpha_k v_k = -v_i$, then

$$\mathbf{0} = v_i + (-v_i) = 1 v_i + \alpha_1 v_1 + \cdots + \alpha_k v_k$$

would be a solution to (b). Apply Farkas' Lemma to $-v_i$ and the cone generated by $v_1, \ldots, v_k$: there is a vector y_i with $y_i^\top (-v_i) < 0$ and $y_i^\top v_j \geq 0$ for all $j = 1, \ldots, k$. So $y_i^\top v_i > 0$ and $y_i^\top v_j \geq 0$ for all other $j \neq i$. Let $d = y_1 + \cdots + y_k$. This d is our desired solution to (a), because for each $i = 1, \ldots, k$:

$$v_i^\top d = d^\top v_i = \sum_{j \neq i} \underbrace{y_j^\top v_i}_{\geq 0} + \underbrace{y_i^\top v_i}_{>0} > 0.$$

□

20.2 Other variants

There are many variants of Farkas' Lemma. They can typically derived from one another by clever rewriting. The most common tricks are:

1. An equality ($a = b$) can be rewritten as two inequalities ($a \leq b$ and $b \leq a$).
2. Real vectors can be written as the difference of two nonnegative vectors: if $x \in \mathbb{R}^n$, then $x = x^+ - x^-$ with $x^+, x^- \geq \mathbf{0}$ defined as follows:

   $$\text{for each coordinate } i = 1, \ldots, n: \quad x_i^+ = \max\{x_i, 0\} \qquad \text{and} \qquad x_i^- = \max\{-x_i, 0\}.$$

   For instance, $x = (3, -2, 4) = x^+ - x^-$ with $x^+ = (3, 0, 4)$ and $x^- = (0, 2, 0)$.
3. An inequality can be written as an equality with a 'slack' variable to fill the gap. For instance,

   $$x_1 + 2x_2 \leq 3 \Leftrightarrow x_1 + 2x_2 + s = 3 \text{ for some } s \geq 0.$$

Let's practise on one variant of Farkas' Lemma:

Theorem 20.3

Exactly one of the following two problems has a solution:

(i) $Ax \leq b$;

(ii) $y^\top A = \mathbf{0}^\top$, $y \geq \mathbf{0}$, $y^\top b < 0$.

Proof: $Ax \leq b$ has a solution if and only if $A(x^+ - x^-) + w = b$ has a nonnegative solution (x^+, x^-, w). Thus, with $B = [A, -A, I]$:

$$Ax \leq b \text{ has a solution} \quad \Longleftrightarrow \quad B \begin{bmatrix} x^+ \\ x^- \\ w \end{bmatrix} = b \text{ has a nonnegative solution.}$$

Using Farkas' Lemma with B instead of A, the second statement is true if and only if there is no y such that $y^\top B \geq \mathbf{0}^\top$ and $y^\top b < 0$. In other words, there is no y such that $y^\top A \geq \mathbf{0}^\top$, $y^\top (-A) \geq \mathbf{0}^\top$, $y^\top I \geq \mathbf{0}^\top$, and $y^\top b < 0$. Rewriting once more gives that there is no y such that $y^\top A = \mathbf{0}^\top$, $y \geq \mathbf{0}$, and $y^\top b < 0$. □

20.3 Application: stationary distributions of Markov chains

Let $n \in \mathbb{N}$. The ***unit simplex***

$$\Delta_n = \{x \in \mathbb{R}^n : x \geq \mathbf{0}, x_1 + \cdots + x_n = 1\}$$

consists of all probability vectors in $\mathbb{R}^n$: vectors whose coordinates are nonnegative and add up to one. A square matrix $A \in \mathbb{R}^{n \times n}$ is a ***stochastic matrix*** if each column (or each row, depending on an arbitrary choice of direction) is a probability vector. In the theory on Markov chains, stochastic matrices are used to describe transition probabilities: a_{ij} is the probability (notice the order) of moving from state j in the current period ('today') to state i in the next period ('tomorrow'). For instance, suppose there are two states, state 1 being good weather and state 2 being bad, and the transition probability matrix is

$$\begin{bmatrix} \frac{4}{5} & \frac{1}{3} \\ \frac{1}{5} & \frac{2}{3} \end{bmatrix}.$$

The first column says that if the weather is good today, the probability of the weather being good tomorrow is $4/5$ and the probability of the weather being bad is $1/5$. The second column is interpreted likewise.

If a probability vector $x \in \Delta_n$ specifies the probability of being in any of the n states today, then Ax is the probability distribution over the states tomorrow. Observe that Ax indeed lies in Δ_n: it is a convex combination of the columns of A. Since each column of A lies in the convex set Δ_n, so does their convex combination.

We call $x \in \Delta_n$ a ***stationary distribution*** if the distribution over states remains unchanged over time: $Ax = x$. We now prove that *each stochastic matrix has a stationary distribution.*

A stationary distribution is a nonnegative solution x to $Ax = x$ whose coordinates sum to one (probabilities!), i.e., a nonnegative solution to

$$\begin{bmatrix} a_{11}-1 & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22}-1 & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn}-1 \\ 1 & 1 & \cdots & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix} \tag{64}$$

We show that such a solution exists via Farkas' Lemma, arguing that the second alternative in that lemma cannot hold. This second alternative says that there is a vector $y = (y_1, \ldots, y_n, y_{n+1})$ whose inner product with each column of the matrix in (64) is nonnegative:

$$\begin{aligned} (a_{11}-1)y_1 + a_{21}y_2 + \cdots + a_{n1}y_n + 1y_{n+1} &\geq 0, \\ a_{12}y_1 + (a_{22}-1)y_2 + \cdots + a_{n2}y_n + 1y_{n+1} &\geq 0, \\ &\vdots \\ a_{1n}y_1 + a_{2n}y_2 + \cdots + (a_{nn}-1)y_n + 1y_{n+1} &\geq 0, \end{aligned} \tag{65}$$

but whose inner product with the vector on the right side of (64) is less than zero:

$$0y_1 + \cdots + 0y_n + 1y_{n+1} = y_{n+1} < 0.$$

Let y_k be the largest number among $y_1, \ldots, y_n$. Since the numbers in the k-th column of A are probabilities (i.e., nonnegative with sum one), we find that

$$y_k = a_{1k}y_k + \cdots + a_{nk}y_k \geq a_{1k}y_1 + \cdots + a_{nk}y_n.$$

But rewriting the k-th inequality in (65) gives the opposite:

$$y_k \le a_{1k}y_1 + \cdots + a_{nk}y_n + 1\underbrace{y_{n+1}}_{<0} < a_{1k}y_1 + \cdots + a_{nk}y_n.$$

So the second alternative in Farkas' Lemma has no solution: a stationary distribution exists.

Exercises section 20

20.1 (a) Use Fourier-Motzkin elimination to find all solutions x to the following system of linear inequalities $Ax \le b$:

$$x_1 + x_2 + x_3 \le 4, \quad x_1 - 2x_2 - x_3 \le 0, \quad -x_1 + x_2 + x_3 \le 1, \quad -x_1 - 3x_2 - 4x_3 \le -7.$$

(b) By Theorem 20.3, there is no solution y to $y^\top A = \mathbf{0}^\top, y \ge \mathbf{0}, y^\top b < 0$. Verify this explicitly.
HINT: first solve $y^\top A = \mathbf{0}^\top$ by Gaussian elimination. Then try to get the inequalities right.

20.2 Let $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$. Show that exactly one of the two following systems has a solution:

	system 1	system 2
(a)	$Ax \ge b$	$y^\top A = \mathbf{0}^\top, y^\top b > 0, y \ge \mathbf{0}$
(b)	$Ax = b$	$y^\top A = \mathbf{0}^\top, y^\top b < 0$
(c)	$Ax \le b, x \ge \mathbf{0}$	$y^\top A \ge \mathbf{0}^\top, y^\top b < 0, y \ge \mathbf{0}$
(d)	$Ax = \mathbf{0}, \mathbf{0} \ne x \ge \mathbf{0}$	$y^\top A > \mathbf{0}^\top$
(e)	$Ax = \mathbf{0}, x > \mathbf{0}$	$\mathbf{0}^\top \ne y^\top A \ge \mathbf{0}^\top$
(f)	$Ax \le \mathbf{0}, \mathbf{0} \ne x \ge \mathbf{0}$	$y^\top A > \mathbf{0}^\top, y > \mathbf{0}$
(g)	$Ax \le \mathbf{0}, x > \mathbf{0}$	$\mathbf{0}^\top \ne y^\top A \ge \mathbf{0}^\top, y \ge \mathbf{0}$

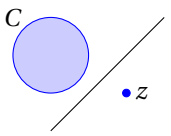
20.3 Let $A \in \mathbb{R}^{m_1\times n}, B \in \mathbb{R}^{m_2\times n}, C \in \mathbb{R}^{m_3\times n}$. Show that exactly one of the following sets is nonempty:

$$\{x \in \mathbb{R}^n : Ax < \mathbf{0}, Bx \le \mathbf{0}, Cx = \mathbf{0}\} \quad \text{or} \quad \{(y_1, y_2, y_3) \in \mathbb{R}^{m_1} \times \mathbb{R}^{m_2} \times \mathbb{R}^{m_3} : y_1^\top A + y_2^\top B + y_3^\top C = \mathbf{0}^\top, \mathbf{0} \ne y_1 \ge \mathbf{0}, y_2 \ge \mathbf{0}\}.$$

20.4 Let $A \in \mathbb{R}^{m\times n}$. Show that there exist $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^m$ with $Ax \le \mathbf{0}, x \ge \mathbf{0}, y^\top A \ge \mathbf{0}^\top$, and $Ax + y > \mathbf{0}$.

21 Separating hyperplane theorems

In convex analysis, the idea behind separation of two sets or between a point and a set — like the set C and the point z in the figure — is very roughly that you can build a straight "wall" between the two, so that each of the two lies on a distinct side of the wall. In two dimensions, the "wall" is simply a line. In three dimensions, it is modelled by a plane and in higher dimensions by a hyperplane. Some preparation that is of interest in its own right:



Theorem 21.1

Let $C \subseteq \mathbb{R}^n$ be nonempty, closed, and convex, and let $z \in \mathbb{R}^n \setminus C$. There is a unique $x \in C$ with minimal (Euclidean) distance to z. It satisfies

$$\text{for all } y \in C: \qquad (z-x)^\top (y-x) \leq 0. \tag{66}$$

Proof: EXISTENCE OF x: There is a vector $x \in C$ that lies as close as possible to z. To see this, let $y \in C$. A point with minimal distance to z cannot lie further away than y does: it must lie in the set $\{x \in \mathbb{R}^n : \|x-z\| \leq \|y-z\|\} \cap C$. This set is nonempty (it contains y), closed (as the intersection of two closed sets) and bounded (as it is contained in a closed ball), hence compact. The continuous function $x \mapsto \|x-z\|$ on this set achieves a minimum by the Extreme Value Theorem.
UNIQUE x: See Exercise 21.1.
PROPERTY (66). Let $y \in C$. By convexity of C: $\lambda y + (1-\lambda)x \in C$ for all $\lambda \in (0,1)$. By definition of x:

$$\text{for all } \lambda \in (0,1): \qquad \|\lambda y + (1-\lambda)x - z\|^2 = \|\lambda(y-x) + (x-z)\|^2 \geq \|x-z\|^2.$$

Rewrite in terms of inner products:

$$\text{for all } \lambda \in (0,1): \qquad \lambda^2 (y-x)^\top (y-x) + 2\lambda (y-x)^\top (x-z) + (x-z)^\top (x-z) \geq (x-z)^\top (x-z).$$

Simplifying this expression and dividing by $\lambda \in (0,1)$ gives:

$$\text{for all } \lambda \in (0,1): \qquad 2(z-x)^\top (y-x) \leq \lambda \|y-x\|^2.$$

Letting λ go down to zero, the right side becomes arbitrarily small, so $(z-x)^\top (y-x) \leq 0$, proving (66).□

Here is the first separation result. One commonly speaks of strict point-set separation: it separates a point from a set by means of a hyperplane and it does so strictly, in the sense that C and z lie in the interior of their respective halfspaces.

Theorem 21.2 (Strict point-set separation)

Let $C \subseteq \mathbb{R}^n$ be a nonempty, closed, convex set, and let $z \notin C$. Then there is a vector $c \in \mathbb{R}^n, c \neq \mathbf{0}$, and a number $\delta \in \mathbb{R}$ such that

$$\text{for all } y \in C: \qquad c^\top y < \delta < c^\top z.$$

Proof: By Theorem 21.1, there is an $x \in C$ with minimal distance to z. Define $c = z - x \neq \mathbf{0}$. By (66), $c^\top y \leq c^\top x$ for all $y \in C$. Since $c \neq \mathbf{0}$: $0 < \|c\|^2 = c^\top c = c^\top (z-x)$. So $c^\top x < c^\top z$. Hence, any δ with $c^\top x < \delta < c^\top z$ will do the trick. □

Using this result, we can characterize closed convex sets as the intersection of (affine) halfspaces.

Theorem 21.3

Set $C \subseteq \mathbb{R}^n$ is closed and convex if and only if there is a collection $\mathcal{H}$ of halfspaces such that $C = \cap_{H \in \mathcal{H}} H$.

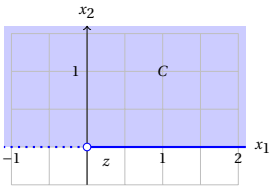
Proof: $\Rightarrow$: Let $\mathcal{H}$ be the collection of halfspaces H with $C \subseteq H$. Clearly, $C \subseteq \cap_{H \in \mathcal{H}} H$. To show that $C \supseteq \cap_{H \in \mathcal{H}} H$, let $x \in \cap_{H \in \mathcal{H}} H$ and suppose that $x \notin C$. By Theorem 21.2, there is a halfspace H' with $C \subseteq H'$ and $x \notin H'$. So $H' \in \mathcal{H}$ and $x \notin \cap_{H \in \mathcal{H}} H$, a contradiction.

$\Leftarrow$: Let C be the intersection of halfspaces. Each halfspace is closed and convex. As the intersection of closed sets, C is closed; as the intersection of convex sets, C is convex. □

In Theorem 21.2, we need to assume that C is closed:

Example 21.1 Consider the convex set

$$C = \{x \in \mathbb{R}^2 : x_2 \geq 0, \text{ and if } x_2 = 0, \text{ then } x_1 > 0\}$$



and $z = \mathbf{0} = (0,0)$. Then $z \notin C$, but there are no $c \in \mathbb{R}^2$ and $\delta \in \mathbb{R}$ with $c^\top z > \delta$ and $c^\top x < \delta$ for all $x \in C$. ◁

We can, however, find a weaker form of separation for arbitrary convex sets:

Theorem 21.4 (Weak point-set separation)

Let $C \subseteq \mathbb{R}^n$ be nonempty, convex and let $z \notin C$. Then there exists a vector $c \in \mathbb{R}^n, c \neq \mathbf{0}$, and a number $\delta \in \mathbb{R}$ such that $c^\top x \leq \delta \leq c^\top z$ for all $x \in C$.

Proof: Translating by $-z$ if necessary, we may assume that $z = \mathbf{0} \notin C$ and we show that there is a vector $c \in \mathbb{R}^n, c \neq \mathbf{0}$ with $c^\top x \geq 0$ for all $x \in C$.

Define for each $x \in C$ the nonempty, closed set $S_x = \{y \in \mathbb{R}^n : \|y\| = 1, y^\top x \geq 0\}$. Let $\{x^1, \ldots, x^m\}$ be a nonempty, finite set of points in C. Since $\mathbf{0} \notin C$, there is no solution $\lambda \in \mathbb{R}^m$ to

$$\sum_{i=1}^m \lambda_i x^i = \mathbf{0}, \sum_{i=1}^m \lambda_i = 1, \lambda \geq \mathbf{0}.$$

Equivalently, there is no solution $\lambda \in \mathbb{R}^m$ to

$$\sum_{i=1}^m \lambda_i x^i = \mathbf{0}, \mathbf{0} \neq \lambda \geq \mathbf{0}.$$

By Exercise 20.2, there is a vector $y \in \mathbb{R}^n$ with $y^\top x^i > 0$ for all i. Obviously, $y \neq \mathbf{0}$ and we can rescale y such that $\|y\| = 1$. Hence, $y \in \cap_{i=1}^m S_{x^i} \neq \emptyset$. Since the sets S_x are closed subsets of the compact set $\{y \in \mathbb{R}^n : \|y\| = 1\}$, and the intersection of finitely many of them is nonempty, the finite intersection property (Theorem 17.2) assures that $\cap_{x \in X} S_x \neq \emptyset$. Let c be any point in this intersection. Then $\|c\| = 1$ gives $c \neq \mathbf{0}$, and by construction $c^\top x \geq 0$ for all $x \in C$: we have found the desired hyperplane. □

Here is an application to the separation of two convex sets:

Theorem 21.5 (Weak set-set separation)

Let C_1 and C_2 be two nonempty, convex sets in $\mathbb{R}^n$ with $C_1 \cap C_2 = \emptyset$. Then there exists a vector $c \in \mathbb{R}^n, c \neq \mathbf{0}$, and a number $\delta \in \mathbb{R}$ such that $c^\top x \leq \delta \leq c^\top y$ for all $x \in C_1$ and $y \in C_2$.

Proof: The set $C = C_1 - C_2 = \{x - y : x \in C_1, y \in C_2\}$ is nonempty and convex and $\mathbf{0} \notin C$. By Theorem 21.4, there exists a vector $c \in \mathbb{R}^n, c \neq \mathbf{0}$, such that $c^\top v \leq c^\top \mathbf{0} = 0$ for all $v \in C$. It follows that $c^\top (x - y) \leq 0$ or, equivalently, that $c^\top x \leq c^\top y$ for all $x \in C_1, y \in C_2$. Taking $\delta = \sup\{c^\top x : x \in C_1\}$ does the trick. □

Exercises section 21

21.1 Prove that the vector $x \in C$ minimizing the distance to z in Theorem 21.1 is unique. Argue by contradiction: suppose x_1 and x_2 both have minimal distance to z. Consider $x = \frac{1}{2}(x_1 + x_2)$ and apply the parallelogram law with $\frac{1}{2}(x_1 - z)$ and $\frac{1}{2}(x_2 - z)$ in the place of x and y.

21.2 Separating hyperplane theorems typically say when two *convex* sets (in Theorems 21.2 and 21.4, one of these consists of a single point z) can be separated by a hyperplane. If one of the sets is *not* convex, sometimes you can separate them by a hyperplane, sometimes not:

(a) The circle $C = \{x \in \mathbb{R}^2 : x_1^2 + x_2^2 = 1\}$ around the origin in $\mathbb{R}^2$ with radius one is not convex. Why?

(b) Find a point $z \in \mathbb{R}^2$ that can be separated from C with a hyperplane. A clear drawing suffices.

(c) Find a point $z \in \mathbb{R}^2$ that cannot be separated from C with a hyperplane. Try to prove this.

Separating hyperplane theorems typically require that the convex sets have little in common (see requirement $z \notin C$ in Theorems 21.2 and 21.4 and $C_1 \cap C_2 = \emptyset$ in Theorem 21.5). Here is what can go wrong if they have even one point in common:

(d) In $\mathbb{R}^2$, draw the closed convex sets $C_1 = \{x \in \mathbb{R}^2 : x_1 = 0\}$ and $C_2 = \{x \in \mathbb{R}^2 : x_2 = 0\}$. Show that they have only one point in common, but that they cannot be separated by a hyperplane: there is no nonzero vector $c \in \mathbb{R}^2$ with $c^\top x \leq c^\top y$ for all $x \in C_1$ and $y \in C_2$.

Finally, let us use Farkas' Lemma to find a separating hyperplane in a specific case:

(e) Consider the convex cone $\text{cone}\{v_1, v_2\}$ with $v_1 = (2,2)$ and $v_2 = (-1,-2)$ and the point $z = (-1,1)$. Use Farkas' lemma to (1) show that z does not belong to the cone and (2) find a hyperplane that separates z from the cone. (It may be helpful to draw a sketch first.)

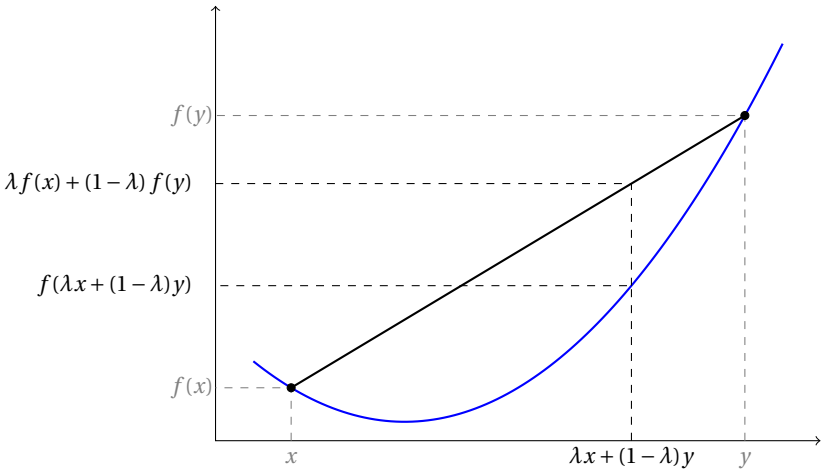
22 Convex functions and variants

In addition to convex *sets*, there are also convex *functions*. This section contains results about (variants of) convex functions. You will have a good big-picture grasp of the material if you understand:

- ☒ the definitions;
- ☒ the idea behind the figures in subsection 22.1;
- ☒ that there are various ways to transform convex functions to new ones (Theorem 22.6);
- ☒ that variants of convex functions facilitate and add structure to the solution of optimization problems (subsection 22.4).

22.1 Basic properties of convex functions

A function $f : C \to \mathbb{R}$ on a convex domain C, like the real line in our picture below, is called convex if the line piece connecting any two points $(x, f(x))$ and $(y, f(y))$ on its graph has no points below the graph.



Definition 22.1 Let $C \subseteq \mathbb{R}^n$ be a convex set. A function $f : C \to \mathbb{R}$ is ***convex*** if

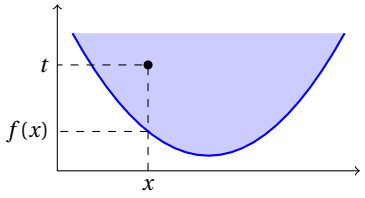
$$\text{for all } x, y \in C \text{ and all } \lambda \in [0,1]: \qquad f(\lambda x + (1-\lambda) y) \leq \lambda f(x) + (1-\lambda) f(y). \tag{67}$$

Example 22.1 We use the definition to show that the function $f : \mathbb{R} \to \mathbb{R}$ with $f(x) = x^2$ is convex. Let $x, y \in \mathbb{R}$ and $\lambda \in [0,1]$. Then

$$\begin{aligned} f\big(\lambda x + (1-\lambda) y\big) \leq \lambda f(x) + (1-\lambda) f(y) \quad &\iff \quad \big(\lambda x + (1-\lambda) y\big)^2 \leq \lambda x^2 + (1-\lambda) y^2 \\ &\iff \quad \lambda^2 x^2 + 2\lambda(1-\lambda) x y + (1-\lambda)^2 y^2 \leq \lambda x^2 + (1-\lambda) y^2 \\ &\iff \quad \lambda(1-\lambda)(x^2 - 2xy + y^2) \geq 0 \\ &\iff \quad \lambda(1-\lambda)(x-y)^2 \geq 0. \end{aligned}$$

The final inequality is true because all three terms in the product are nonnegative. ◁

This definition might make it evident that there is a close connection between convex *sets* and convex *functions*: the set of points on/above its graph is a convex set. Using the Greek prefix 'epi-' for 'on/above', this set is called the epigraph of the function. Since points *on* the graph are of the form $(x, f(x))$, those *above* the graph (see our figure) are of the form (x, t) for some $t \geq f(x)$. To summarize:



Definition 22.2 The ***epigraph*** of a function $f : C \to \mathbb{R}$ with domain $C \subseteq \mathbb{R}^n$ is the set

$$\operatorname{epi}(f) = \{(x,t) \in C \times \mathbb{R} : t \geq f(x)\} \subseteq \mathbb{R}^{n+1}.$$

Theorem 22.1 (A function is convex if and only if area above its graph is convex)

Let $C \subseteq \mathbb{R}^n$ be convex. A function $f : C \to \mathbb{R}$ is convex if and only if $\operatorname{epi}(f)$ is a convex set.

Proof: **(⇒)** Assume f is convex. Let $(x,t),(y,t') \in \operatorname{epi}(f)$, and $\lambda \in [0,1]$. To show:

$$\lambda(x,t) + (1-\lambda)(y,t') = (\lambda x + (1-\lambda)y, \lambda t + (1-\lambda)t') \in \operatorname{epi}(f).$$

Since $(x,t),(y,t') \in \operatorname{epi}(f)$ and f is convex, it follows that

$$\lambda t + (1-\lambda)t' \geq \lambda f(x) + (1-\lambda) f(y) \geq f(\lambda x + (1-\lambda)y).$$

Since the first term is larger than the third, it follows that $(\lambda x + (1-\lambda)y, \lambda t + (1-\lambda)t') \in \operatorname{epi}(f)$.

(⇐) Assume $\operatorname{epi}(f)$ is a convex set. Then for all $x, y \in C$ and $\lambda \in [0,1]$: $(x, f(x)), (y, f(y)) \in \operatorname{epi}(f)$, so $(\lambda x + (1-\lambda)y, \lambda f(x) + (1-\lambda) f(y)) \in \operatorname{epi}(f)$. The latter means $f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y)$. □

Example 22.2 The absolute-value function $f : \mathbb{R} \to \mathbb{R}$ with $f(x) = |x| = \max\{-x, x\}$ is convex: its epigraph

$$\operatorname{epi}(f) = \left\{(x,t) \in \mathbb{R}^2 : t \geq f(x) = \max\{-x,x\}\right\} = \left\{(x,t) \in \mathbb{R}^2 : t \geq -x, t \geq x\right\}$$

is polyhedral (it is the set of solutions to two linear inequalities), hence convex. ◁

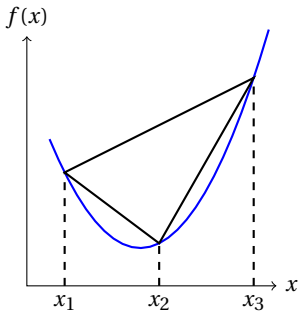
Analogously, we may replace the weak inequality in the epigraph with a strict one:

$$f : C \to \mathbb{R} \text{ is convex} \quad \Longleftrightarrow \quad \{(x,t) \in C \times \mathbb{R} : t > f(x)\} \text{ is a convex set.} \tag{68}$$

In the figure to the right, look at the line piece connecting the points $(x_1, f(x_1))$ and $(x_3, f(x_3))$. Its slope is given by the difference quotient

$$\frac{f(x_3) - f(x_1)}{x_3 - x_1}$$

of the change $f(x_3) - f(x_1)$ in the function value relative to the change $x_3 - x_1$ in the argument of the function. The function is convex, so the function value at the intermediate point x_2 lies below the line piece: the line piece connecting $(x_1, f(x_1))$ to $(x_2, f(x_2))$ moves down steeper than the line piece we started with, i.e., its slope is smaller:



$$\frac{f(x_2) - f(x_1)}{x_2 - x_1} \leq \frac{f(x_3) - f(x_1)}{x_3 - x_1}.$$

And the third line piece from $(x_2, f(x_2))$ to $(x_3, f(x_3))$ then needs to move pretty steeply up to get back to $(x_3, f(x_3))$. This gives:

Theorem 22.2

Let $f : I \to \mathbb{R}$ be a function on an interval I of real numbers.

(a) If f is convex, then for all $x_1, x_2, x_3 \in I$ with $x_1 < x_2 < x_3$:

$$\frac{f(x_2)-f(x_1)}{x_2-x_1} \leq \frac{f(x_3)-f(x_1)}{x_3-x_1} \leq \frac{f(x_3)-f(x_2)}{x_3-x_2}. \tag{69}$$

(b) Conversely, each of the inequalities in (69) implies that f is convex.

Proof: **(a)** Let $x_1, x_2, x_3 \in I$ have $x_1 < x_2 < x_3$. Since x_2 lies between x_1 and x_3 we can write $x_2 = \lambda x_1 + (1-\lambda)x_3$ for some (do you see which?) $\lambda \in (0,1)$. By convexity of f:

$$\begin{aligned}
\frac{f(x_2)-f(x_1)}{x_2-x_1} &= \frac{f(\lambda x_1 + (1-\lambda)x_3) - f(x_1)}{\lambda x_1 + (1-\lambda)x_3 - x_1} \\
&\leq \frac{\lambda f(x_1) + (1-\lambda) f(x_3) - f(x_1)}{\lambda x_1 + (1-\lambda)x_3 - x_1} \\
&= \frac{(1-\lambda)\big(f(x_3)-f(x_1)\big)}{(1-\lambda)(x_3-x_1)} \\
&= \frac{f(x_3)-f(x_1)}{x_3-x_1}.
\end{aligned}$$

This proves the first inequality in (69); the second is proved the same way.
(b) The arguments for the different inequalities are all similar, so I will just show that the first inequality in (69) implies convexity of f. Let $x, y \in I$ have $x < y$ and let $\lambda \in (0,1)$. Consequently

$$x < \lambda x + (1-\lambda)y < y.$$

Replacing '$x_1 < x_2 < x_3$' with these three values, the first inequality in (69) becomes

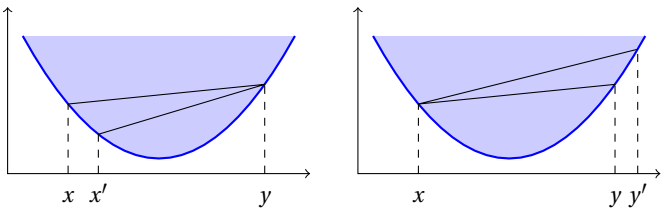
$$\frac{f\big(\lambda x + (1-\lambda)y\big) - f(x)}{\lambda x + (1-\lambda)y - x} = \frac{f\big(\lambda x + (1-\lambda)y\big) - f(x)}{(1-\lambda)(y-x)} \leq \frac{f(y)-f(x)}{y-x}.$$

Rearranging terms gives (67). □

Rewriting the inequalities in (69), it follows that the difference quotient

$$\frac{f(y)-f(x)}{y-x}$$

is (weakly) increasing in both x and y (see figure below): if you increase x to x' or y to y', the associated line piece becomes steeper.



Theorem 22.3 (Convex functions have weakly increasing difference quotients)

A function $f : I \to \mathbb{R}$ on an interval I of real numbers is convex if and only if the difference quotient

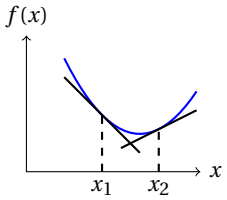
$$Q(x,y) = \frac{f(y)-f(x)}{y-x} \qquad (\text{with } x,y \in I, x \neq y)$$

is a weakly increasing function of each of its two variables.

Proof: Assume f is convex. Since $Q(x,y) = Q(y,x)$ for all distinct x and y, it is enough to prove that Q is weakly increasing in its first variable. So pick distinct $x, x', y \in I$ with $x < x'$. We must show that $Q(x,y) \leq Q(x',y)$. There are three similar cases, depending on whether y is below, above, or between x and x'. I'll do only one: $x < x' < y$. (Verify the other cases yourself.) Substituting these values for x_1, x_2, and x_3 in (69) gives $Q(x,y) \leq Q(x',y)$.

Conversely, if the difference quotient is weakly increasing in both variables, then $Q(x_1,x_2) \leq Q(x_1,x_3)$ for all $x_1, x_2, x_3 \in I$ with $x_1 < x_2 < x_3$: the first inequality in (69) holds, so f is convex by Theorem 22.2. □

Remember that the derivative $f'(x)$ in x of a function f from and to the real numbers is the limit of the difference quotient $\frac{f(y)-f(x)}{y-x}$ as y tends to x. And we just argued that as you move to the right in the domain of f, these difference quotients are weakly increasing. So convex functions have weakly increasing derivatives: in our figure, the slope of the tangent line at the low point x_1 is less than that at the high point x_2.



Theorem 22.4 (Derivative tests for convexity)

Let I be an open interval of real numbers.

(a) A differentiable function $f : I \to \mathbb{R}$ is convex if and only if its derivative f' is a weakly increasing function.

(b) A twice differentiable function $f : I \to \mathbb{R}$ is convex if and only if its second derivative f'' is a nonnegative function.

Proof: **(a)** Assume that f is convex. Take $x, y \in I$ with $x < y$. To show: $f'(x) \leq f'(y)$. For all $h > 0$ with $x < x+h \leq y-h < y$, Theorem 22.3 gives

$$Q(x,x+h) = \frac{f(x+h)-f(x)}{h} \leq Q(y-h,y) = \frac{f(y)-f(y-h)}{h}.$$

Letting h tend to zero, these terms converge to $f'(x)$ and $f'(y)$, respectively, so $f'(x) \leq f'(y)$.

Conversely, assume that f' is a weakly increasing function. Let $x_1, x_2, x_3 \in I$ satisfy $x_1 < x_2 < x_3$. By the Mean Value Theorem,

$$\frac{f(x_2)-f(x_1)}{x_2-x_1} = f'(\alpha) \qquad \text{and} \qquad \frac{f(x_3)-f(x_2)}{x_3-x_2} = f'(\beta)$$

for some numbers α and β with $x_1 < \alpha < x_2 < \beta < x_3$. Since f' is nondecreasing, $f'(\alpha) \leq f'(\beta)$. By Theorem 22.2, f is convex.

(b) Recall that a differentiable function on an open interval is nondecreasing if and only if its derivative is nonnegative. Apply this to the function f' from (a). □

Example 22.3 The second derivatives of the functions $f, g, h : \mathbb{R} \to \mathbb{R}$ with $f(x) = x^2$, $g(x) = x^3$, and $h(x) = e^x$ are $f''(x) = 2$, $g''(x) = 6x$, and $h''(x) = e^x$. Since f'' and h'' are nonnegative functions, f and h are convex. But g'' achieves negative values if $x < 0$, so g is not convex. ◁

Example 22.4 (A slight extension) For the derivative test to work, the interval I need not be open. If f is continuous on the interval I and f' is nondecreasing on the *interior* of I, the proof above — via the Mean Value Theorem — remains valid and f is convex. For instance, the function $f : [0,\infty) \to \mathbb{R}$ with $f(x) = -\sqrt{x}$ is convex: it is continuous on $I = [0,\infty)$ and although it isn't differentiable at $x = 0$ (where the graph is infinitely steep), its derivative $f'(x) = -\frac{1}{2\sqrt{x}}$ on $(0,\infty)$ is a nondecreasing function of x. The latter also follows from its second derivative $f''(x) = \frac{1}{4x\sqrt{x}}$ being nonnegative on $(0,\infty)$. ◁

As we saw in the figure before Theorem 22.4, convex functions lie above their tangents:

Theorem 22.5 (Convex functions lie above their tangents)

Let $f : I \to \mathbb{R}$ be a convex function on an interval I of real numbers. If f is differentiable at a point $x^* \in I$, then the graph of f lies above the tangent line at x^*:

$$\text{for all } x \in I: \qquad f(x) \geq f(x^*) + f'(x^*)(x - x^*).$$

Proof: Let $x \in I$. By convexity of f we have, for each $\lambda \in (0,1]$:

$$f\big(\lambda x + (1-\lambda)x^*\big) \leq \lambda f(x) + (1-\lambda) f(x^*) \qquad \Longleftrightarrow \qquad f\big(x^* + \lambda(x - x^*)\big) \leq f(x^*) + \lambda\big(f(x) - f(x^*)\big).$$

Rearranging terms and dividing by λ gives

$$\frac{f(x^* + \lambda(x - x^*)) - f(x^*)}{\lambda} \leq f(x) - f(x^*).$$

The left side is the difference quotient at $\lambda = 0$ of the function $\lambda \mapsto f(x^* + \lambda(x - x^*))$. By the chain rule, this function is differentiable at $\lambda = 0$ and as λ tends to zero, the difference quotient goes to $f'(x^*)(x - x^*)$, proving the inequality in the theorem. □

The following result indicates how to construct convex functions from others.

Theorem 22.6

Let $C \subseteq \mathbb{R}^n$ be a nonempty, convex set.

(a) If $f : C \to \mathbb{R}$ is a convex function and $\alpha \geq 0$, then αf is a convex function.

(b) If $f : C \to \mathbb{R}$ and $g : C \to \mathbb{R}$ are convex functions, then their sum $f + g$ is a convex function.

(c) If $\{f_i : i \in I\}$ is a collection of convex functions $f_i : C \to \mathbb{R}$ and there is a function $g : C \to \mathbb{R}$ that bounds them from above:

$$\text{for all } i \in I: \qquad f_i \leq g,$$

then the pointwise supremum $f : C \to \mathbb{R}$ with $f(x) = \sup_{i \in I} f_i(x)$ is convex.

(d) If $f : C \to \mathbb{R}$ is convex and $g : U \to \mathbb{R}$ is convex and nondecreasing on some convex set $U \supseteq f(C)$, then $(g \circ f) : C \to \mathbb{R}$ is convex.

Proof: (a) and (b) follow easily from the definition of convexity; (c) follows because $\text{epi}(f) = \cap_{i\in I} \text{epi}(f_i)$ is the intersection of convex sets and hence convex. For (d), let $x, y \in C$ and $\lambda \in [0,1]$. By convexity of f:

$$f(\lambda x + (1-\lambda)y) \le \lambda f(x) + (1-\lambda) f(y).$$

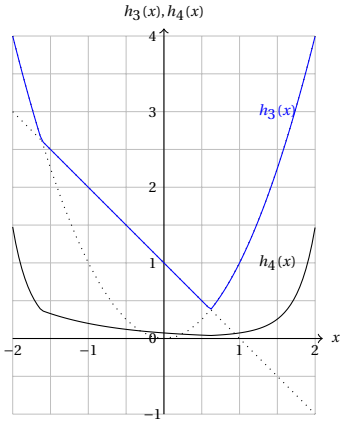
Using that g is nondecreasing to prove the first inequality and convex to prove the second, we find

$$g\big(f(\lambda x + (1-\lambda)y)\big) \le g\big(\lambda f(x) + (1-\lambda) f(y)\big) \le \lambda g(f(x)) + (1-\lambda) g(f(y)). \qquad \square$$

Example 22.5 Using for instance the derivative test, we see that the functions $f : \mathbb{R} \to \mathbb{R}$ and $g : \mathbb{R} \to \mathbb{R}$ with $f(x) = x^2$ and $g(x) = -x + 1$ are convex. By Theorem 22.6, also the following functions $h_1, \ldots, h_4 : \mathbb{R} \to \mathbb{R}$ are convex:

$$\begin{aligned} h_1(x) &= 5x^2, \\ h_2(x) &= x^2 - x + 1, \\ h_3(x) &= \max\{x^2, -x+1\} \\ h_4(x) &= \tfrac{1}{37} \exp(h_3(x)). \end{aligned}$$

The graphs of the final two functions are drawn in the figure to the right. $\triangleleft$



22.2 Variants of convex functions

There are a few common variants of convex functions that pop up in economics; here are some of them.

Definition 22.3 Let $C \subseteq \mathbb{R}^n$ be a convex set. A function $f : C \to \mathbb{R}$ is:

- ☒ ***concave*** if $-f$ is convex. Rewriting makes this equivalent to:

$$\text{for all } x, y \in C \text{ and all } \lambda \in [0,1]: \qquad f(\lambda x + (1-\lambda)y) \ge \lambda f(x) + (1-\lambda) f(y).$$

- ☒ ***quasiconcave*** if

$$\text{for all } x, y \in C \text{ and all } \lambda \in [0,1]: \qquad f(\lambda x + (1-\lambda)y) \ge \min\{f(x), f(y)\}.$$

Sometimes ***strictly*** convex/concave/quasiconcave functions are considered. These require that the defining inequalities are strict ($<$ or $>$ instead of $\le$ or $\ge$) whenever $x \ne y$ and $\lambda \in (0,1)$.

Since f is concave if and only if $-f$ is convex, each result about convex functions translates to a corresponding result about concave functions: there is no need to prove them separately. For instance, by Theorem 22.1, a function is concave if and only if the set of points *under* its graph is a convex set; and by Theorem 22.5, concave functions lie *below* their tangent lines.

Theorem 22.7

Let $C \subseteq \mathbb{R}^n$ be a nonempty, convex set and $f : C \to \mathbb{R}$ a function.

(a) If f is concave, then f is quasiconcave.

(b) f is quasiconcave if and only if for each $r \in \mathbb{R}$, the set $\{x \in C : f(x) \ge r\}$ is a convex set.

Proof: **(a)** Let $x, y \in C$ and $\lambda \in [0,1]$. If f is concave, then

$$f(\lambda x + (1-\lambda)y) \geq \lambda f(x) + (1-\lambda) f(y) \geq \lambda \min\{f(x), f(y)\} + (1-\lambda)\min\{f(x), f(y)\} = \min\{f(x), f(y)\}.$$

(b) Assume f is quasiconcave. Let $r \in \mathbb{R}$, $x, y \in C$, and $\lambda \in [0,1]$. If $f(x) \geq r$ and $f(y) \geq r$, then

$$f(\lambda x + (1-\lambda)y) \geq \min\{f(x), f(y)\} \geq r,$$

establishing convexity of the set $\{x \in C : f(x) \geq r\}$.

Conversely, assume $\{x \in C : f(x) \geq r\}$ is convex for each $r \in \mathbb{R}$. Let $x, y \in C$ and $\lambda \in [0,1]$. Taking $r = \min\{f(x), f(y)\}$ gives $f(\lambda x + (1-\lambda)y) \geq \min\{f(x), f(y)\}$. □

Quasiconcave functions of a single real variable are easy to recognize: they are either monotonic or first go up and then go down.

Theorem 22.8 (Quasiconcave functions of one variable)

A function $f : I \to \mathbb{R}$ on an interval I of real numbers is quasiconcave if and only if (at least) one of the following conditions is true:

(qc1) f is weakly increasing;

(qc2) f is weakly decreasing;

(qc3) there is a point x^* in I such that f is weakly increasing on $I \cap (-\infty, x^*)$ and weakly decreasing on $I \cap [x^*, \infty)$;

(qc4) there is a point x^* in I such that f is weakly increasing on $I \cap (-\infty, x^*]$ and weakly decreasing on $I \cap (x^*, \infty)$.

22.3 More on continuity and differentiability

Convex functions are continuous on the interior of their domain. Exercise 22.1 shows that they need not be continuous in boundary points.

Theorem 22.9 (Convex functions are continuous on the interior of their domain)

Let $C \subseteq \mathbb{R}^n$ be a nonempty, convex set and let $f : C \to \mathbb{R}$ be a convex function. Then f is continuous in each interior point of C.

In Theorem 22.5, we saw that convex functions lie above their tangent lines. We showed this under a differentiability assumption, but the result holds more generally, at least at interior points:

Theorem 22.10

Let $f : C \to \mathbb{R}$ be a convex function on a convex domain C in $\mathbb{R}^n$ and let z be an interior point of f. Then we can find a tangent line at z such that f lies entirely above this tangent: there is a vector $a \in \mathbb{R}^n$ such that

$$\text{for all } x \in C: \qquad f(x) \geq f(z) + a^\top (x - z). \tag{70}$$

Since the tangent line $x \mapsto f(z) + a^\top (x - z)$ lies below (in Latin: 'sub') the graph of f and its gradient is a, the vector a is sometimes called a ***subgradient*** of f in the point z. So such subgradients exist at interior points of the domain. There are two caveats here. First of all, this is about interior points: sometimes there are no subgradients in boundary points, for instance if the function is infinitely steep. Secondly, it says that there is at least one such subgradient in interior points, but there may be more. For instance, the absolute-value function $x \mapsto |x|$ is convex and at $x = 0$, its set of subgradients is $[-1, 1]$.

If the function happens to be differentiable at an interior point, there is only one subgradient, the derivative:[6]

Theorem 22.11 (Derivative as unique subgradient)

Let $f : C \to \mathbb{R}$ be a convex function on a convex domain C in $\mathbb{R}^n$. If f is differentiable at an interior point $z \in C$, then $f'(z)$ is a subgradient, i.e.,

$$\text{for all } x \in C: \qquad f(x) \geq f(z) + f'(z)(x - z).$$

There are no other subgradients at z.

The characterization of convex functions in terms of second derivatives (Theorem 22.4) extends to convex functions

$$(x_1, \ldots, x_n) \mapsto f(x_1, \ldots, x_n)$$

of several variables. Recall that if such a function f is twice differentiable at a point y in its domain, then the ***Hessian*** of f at y is the $n \times n$ matrix

$$H_f(y) = \begin{bmatrix} \frac{\partial^2 f(y)}{\partial x_1 \partial x_1} & \frac{\partial^2 f(y)}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f(y)}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f(y)}{\partial x_2 \partial x_1} & \frac{\partial^2 f(y)}{\partial x_2 \partial x_2} & \cdots & \frac{\partial^2 f(y)}{\partial x_2 \partial x_n} \\ \vdots & \vdots & & \vdots \\ \frac{\partial^2 f(y)}{\partial x_n \partial x_1} & \frac{\partial^2 f(y)}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f(y)}{\partial x_n \partial x_n} \end{bmatrix}$$

of second-order partial derivatives.

Theorem 22.12

Let $C \subseteq \mathbb{R}^n$ be open and convex and let $f : C \to \mathbb{R}$ be twice differentiable. Function f is convex if and only if the Hessian $H_f(y)$ is positive semidefinite at each point y in its domain.

22.4 Applications to optimization

In optimization problems, assumptions related to convexity or concavity are often imposed to make them easier to solve (we will see this in detail in the part on constrained optimization) or to make qualitative statements about the set of solutions. The following two theorems are often used.

[6] This presumes that you know a little about differentiable functions of several variables. If you don't, Section 23 gives a quick recap.

Recall that a point $x \in C$ is a ***local maximum*** of a function $f : C \to \mathbb{R}$ if it has the largest function value among all points in a sufficiently small neighborhood of x: there is a neighborhood U (in particular, some open ball around x) with $f(x) \geq f(y)$ for all $y \in U$. And it is a ***global maximum*** if it has the highest function value among all points in the domain: $f(x) \geq f(y)$ for all $y \in C$.

Theorem 22.13 (Maximizers of concave functions)

Let $f : C \to \mathbb{R}$ be a concave function on a convex domain C in $\mathbb{R}^n$.

(a) If $x \in C$ is a local maximum of f, then it is also a global maximum.

(b) If f is differentiable at an interior point x of C and the first-order condition $\nabla f(x) = \mathbf{0}$ holds, then x is a global maximum.

Proof: **(a)** Since x is a local maximum, $f(x) \geq f(z)$ for all $z \in C$ sufficiently close to x. Let $y \in C$. Since $\lambda x + (1-\lambda)y$ is close to x if λ is close to one, we have that for $\lambda \in (0,1)$ sufficiently large:

$$f(x) \geq f(\lambda x + (1-\lambda)y).$$

By concavity of f,

$$f(\lambda x + (1-\lambda)y) \geq \lambda f(x) + (1-\lambda) f(y).$$

Combining the two inequalities, $f(x) \geq \lambda f(x) + (1-\lambda) f(y)$. Rearranging terms and dividing by $1-\lambda > 0$ gives $f(x) \geq f(y)$. Since this holds for arbitrary $y \in C$, x is a global maximum.
(b) Applying Theorem 22.11 to the convex function $-f$ we have for each $y \in C$:

$$f(y) \leq f(x) + \nabla f(x)(y-x) = f(x) + \mathbf{0}(y-x) = f(x). \qquad \square$$

Theorem 22.14 (Maximizers of quasiconcave functions)

Let $f : C \to \mathbb{R}$ be a function on a convex domain C in $\mathbb{R}^n$. Denote its set of global maxima by

$$C_{\max} = \{x \in C : f(x) \geq f(y) \text{ for all } y \in C\}.$$

(a) If f is quasiconcave, then the set $C_{\max}$ of maximizers is convex.

(b) If f is strictly quasiconcave, then the set $C_{\max}$ of maximizers has at most one element.

Proof: **(a)** If $C_{\max}$ is empty, it is convex. If it is nonempty, let x^* be one of its elements. It achieves maximal value $r = f(x^*)$. By quasiconcavity, the set $C_{\max} = \{x \in X : f(x) \geq r\}$ is convex.
(b) If, to the contrary, $C_{\max}$ has more than one element, we can pick two distinct ones, x and x'. The domain C is convex, so $\frac{1}{2}x + \frac{1}{2}x'$ lies in C as well. By strict quasiconcavity,

$$f\left(\tfrac{1}{2}x + \tfrac{1}{2}x'\right) > \min\{f(x), f(x')\} = f(x) = f(x').$$

This contradicts that x and x' maximize f. $\square$

Quasiconcavity is a popular assumption in traditional results about economic and game-theoretic equilibria: Theorem 22.14 tells that optimizing behavior leads to convex sets of solutions, which in its turn is required in the fixed-point theorem of Kakutani — one of the most commonly invoked theorems to establish the existence of equilibria.

22.5 Postponed proofs

22.5.1 Proof of Theorem 22.8

STEP 1: If f satisfies one of the properties (qc1) to (qc4), then f is quasiconcave.

I will only do the proof for (qc3): the proof for (qc4) is similar and those for (qc1) and (qc2) are baked into this proof. So assume f satisfies (qc3): there is a point x^* in I such that f is weakly increasing on $I \cap (-\infty, x^*)$ and weakly decreasing on $I \cap [x^*, \infty)$. Take $x, y \in I$ and $\lambda \in (0,1)$. Without loss of generality, $x < y$. Consequently,

$$x < \lambda x + (1-\lambda) y < y.$$

Distinguish two cases, depending on where the convex combination $\lambda x + (1-\lambda) y$ lies. If it lies in the set $I \cap (-\infty, x^*)$ where f is weakly increasing, then so does the smaller number x. Hence,

$$f(\lambda x + (1-\lambda) y) \geq f(x) \geq \min\{f(x), f(y)\}.$$

If it lies in the set $I \cap [x^*, \infty)$ where f is weakly decreasing, then so does the larger number y. Hence

$$f(\lambda x + (1-\lambda) y) \geq f(y) \geq \min\{f(x), f(y)\}.$$

Conclude that f is quasiconcave.

STEP 2: If f is quasiconcave, then it satisfies one of the properties (qc1) to (qc4).

We repeatedly appeal to the following.

Claim: if $x, y \in I$ satisfy $x < y$ and $f(x) > f(y)$, then f is weakly decreasing from y onward, i.e., on $I \cap [y, \infty)$.

Proof (of claim): Suppose, to the contrary, that there are $v, w \in I$ with $y \leq v < w$ and $f(v) < f(w)$.

- ☒ If $f(y) \leq f(v)$, then $x < y < w$, but $f(y) < f(x)$ and $f(y) < f(w)$, contradicting quasiconcavity: y lies on the linepiece between x and w, so its function value cannot be below that of both endpoints.
- ☒ If $f(y) > f(v)$, then $y < v < w$, but $f(v) < f(y)$ and $f(v) < f(w)$, again contradicting quasiconcavity. □

Now the main argument: we must show that one of the properties (qc1) to (qc4) is true. If f is weakly increasing, then (qc1) holds. So from now, suppose that f is not weakly increasing: there are $x, y \in I$ with $x < y$ and $f(x) > f(y)$. By the claim, f is weakly decreasing from y onward, so the set

$$A = \{a \in I : f \text{ is weakly decreasing on } I \cap [a, \infty)\}$$

is nonempty: it contains y.

If A does not have a lower bound in I, then f is weakly decreasing on I: (qc2) holds.

If A does have a lower bound in I, let $x^* \in I$ be its infimum. Firstly, if x^* lies in A, (qc3) holds:

- ☒ Since $x^* \in A$, f is weakly decreasing on $I \cap [x^*, \infty)$.
- ☒ And it is weakly increasing on $I \cap (-\infty, x^*)$: otherwise there would exist $v, w \in I \cap (-\infty, x^*)$ with $v < w$ and $f(v) > f(w)$. By our claim, f is then weakly decreasing on $I \cap [w, \infty)$, i.e., the point $w < x^*$ lies in A, contradicting that x^* is a lower bound on A.

Secondly, if x^* does not lie in A, (qc4) holds:

- ☒ To see that f is weakly decreasing on $I \cap (x^*, \infty)$, pick two elements v and w in this set with $v < w$. We want to argue that $f(v) \geq f(w)$. Since $x^* < v$ and x^* is the greatest lower bound on A, v is not a lower bound on A: there is an element $a \in A$ with $a < v$. Consequently, f is weakly decreasing on the interval $I \cap [a, \infty)$, which contains v and w. Hence, $f(v) \geq f(w)$.

⊠ Finally, f is weakly increasing on $I \cap (-\infty, x^*]$. Otherwise, there are v and w in this set with $v < w$ and $f(v) > f(w)$. Our claim then implies that the element $w \le x^*$ lies in A. If $w = x^*$, this contradicts our assumption that $x^* \notin A$. And if $w < x^*$, this contradicts x^* being a lower bound on A.

22.5.2 Proof of Theorem 22.9

This is one of those instances where the use of different norms than the standard Euclidean one is beneficial: we use the 'box-shaped' balls of the supremum norm to establish that there is a finite collection of points such that each point in this ball can be described as a suitably chosen convex combination. This helps to show that the convex function remains bounded on the ball and from there to continuity is only a small step.

Let $c \in \text{int}(C)$: there is a $\delta > 0$ such that $c + x \in C$ for all $x \in D = \{x \in \mathbb{R}^n : \|x\|_\infty \le \delta\}$. Define auxiliary function $h : D \to \mathbb{R}$ with $h(x) = f(c+x) - f(c)$. Then h is convex, $\mathbf{0} \in \text{int}(D)$, and $h(\mathbf{0}) = 0$. It suffices to prove that h is continuous at $\mathbf{0}$.

By construction:

$$\text{for all } i = 1, \dots, n: \qquad \delta e_i \in D \text{ and } \delta(-e_i) \in D$$

and if $x \in D$, then $\mathbf{0} = \frac{1}{2}x + \frac{1}{2}(-x)$, so $0 = h(\mathbf{0}) \le \frac{1}{2}h(x) + \frac{1}{2}h(-x)$ implies

$$\text{if } x \in D, \text{ then:} \qquad h(x) \ge -h(-x). \tag{71}$$

Let $x \in \mathbb{R}^n$ have $\|x\|_\infty < \frac{\delta}{n}$. Then x is a convex combination of $\delta e_1, \dots, \delta e_n, \delta(-e_1), \dots, \delta(-e_n), \mathbf{0} \in D$:

$$x = \sum_{i:x_i>0} \frac{x_i}{\delta} \delta e_i + \sum_{i:x_i<0} \frac{-x_i}{\delta} \delta(-e_i) + \left(1 - \sum_i \frac{|x_i|}{\delta}\right)\mathbf{0}.$$

Let $\beta = \frac{1}{\delta}\max\{h(\delta e_1), \dots, h(\delta e_n), h(\delta(-e_1)), \dots, h(\delta(-e_n))\}$. By (71), $\beta \ge 0$. By convexity of h:

$$h(x) \le \sum_{i:x_i>0} \frac{|x_i|}{\delta} h(\delta e_i) + \sum_{i:x_i<0} \frac{|x_i|}{\delta} h(\delta(-e_i)) \le \beta \sum_i |x_i| = \beta \|x\|_1.$$

Replacing x by $-x$, we have $h(-x) \le \beta\|-x\|_1 = \beta\|x\|_1$. With (71), this gives $h(x) \ge -h(-x) \ge -\beta\|x\|_1$. So

$$|h(x) - h(\mathbf{0})| = |h(x) - 0| = |h(x)| \le \beta\|x\|_1 = \beta\|x - \mathbf{0}\|_1,$$

so h is continuous at 0. Hence f is continuous at c.

22.5.3 Proof of Theorem 22.10

By expression (68), the set

$$D = \{(x, t) \in C \times \mathbb{R} : t < f(x)\} \subseteq \mathbb{R}^{n+1}$$

is convex and $(z, f(z)) \notin D$. By Theorem 21.4, there is a vector $(a, \tau) \neq \mathbf{0}$ such that

$$\text{for all } x \in C: \qquad a^\top x + \tau t \le a^\top z + \tau f(z). \tag{72}$$

Then $\tau \neq 0$: if, to the contrary, τ were zero, then $a^\top x \le a^\top z$ for all $x \in C$. Since $z + \varepsilon a \in C$ for sufficiently small $\varepsilon > 0$, it follows that $a^\top z + \varepsilon\|a\|^2 \le a^\top z$, a contradiction. So $\tau \neq 0$.

Since $(z, f(z) + 1) \in D$, we also know that

$$a^\top z + \tau(f(z) + 1) \le a^\top z + \tau f(z).$$

So $\tau < 0$. Dividing the vector (a, τ) by $-\tau$ if necessary, we may assume w.l.o.g. that $\tau = -1$. Substituting this in (72) and rewriting gives that

$$\text{for all } (x, t) \in D: \qquad t \ge f(z) + a^\top(x - z).$$

Letting t move down to $f(x)$ we obtain (70).

22.5.4 Proof of Theorem 22.11

That $f'(z)$ is a subgradient follows as in Theorem 22.5: Let $x \in C$. By convexity of f, for each $\lambda \in (0,1]$:

$$f(\lambda x + (1-\lambda)z) \leq \lambda f(x) + (1-\lambda)f(z) \quad \Longleftrightarrow \quad f(z + \lambda(x-z)) \leq f(z) + \lambda(f(x) - f(z)).$$

Rearranging terms and dividing by λ gives:

$$\frac{f(z+\lambda(x-z)) - f(z)}{\lambda} \leq f(x) - f(z).$$

As λ goes down to zero, the left-hand side converges to $f'(z)(x-z)$.

We now prove that there are no other subgradients. Suppose that also a is a subgradient of f at z:

$$\text{for all } x \in C: \qquad f(x) \geq f(z) + a^\top (x-z).$$

Pick any vector $v \in \mathbb{R}^n$. Since z lies in the interior of C, $z + tv$ lies in C as long as scalar t is sufficiently close to zero. Substituting $x = z + tv$ in the inequality above, we find that

$$f(z+tv) \geq f(z) + a^\top (tv).$$

So on a neighborhood of $t = 0$, the function

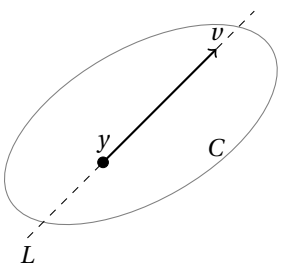
$$t \mapsto f(z+tv) - f(z) - a^\top (tv)$$

is well-defined and nonnegative. At $t = 0$, the function value is zero. So this function achieves its global maximum at the interior point 0 of its domain. By the usual first-order condition, its derivative — which exists, since f is differentiable at z — must be zero. By the chain rule, this derivative is $f'(z)v - a^\top v$.

Since $f'(z)v - a^\top v = 0$ for each v, substituting the standard basis vectors $e_1, \ldots, e_n$ shows that $f'(z) = a^\top$.

22.5.5 Proof of Theorem 22.12

Function f is convex if and only if at each point y in C and each direction $v \neq \mathbf{0}$, its restriction to the straight line $L = \{y + tv : t \in \mathbb{R}\}$ or more precisely, to $L \cap C$, is convex. In other words, for all such y and v, the single-variable function $t \mapsto f(y+tv)$ is convex: its second derivative in $t = 0$ is nonnegative according to Theorem 22.4; this derivative is $v^\top H_f(y) v$. And $v^\top H_f(y) v \geq 0$ for all $y \in C$ and all $v \neq \mathbf{0}$ simply means that $H_f(y)$ is positively semidefinite for each $y \in C$.



Exercises section 22

22.1 For which real numbers a is the function $f : [0,1] \to \mathbb{R}$ with $f(1) = a$ and $f(x) = 0$ for all other x:

(a) convex?

(b) concave?

(c) quasiconcave?

22.2 Give an example of:

(a) a quasiconcave function from and to $\mathbb{R}$ that is not continuous;

(b) two quasiconcave functions from and to $\mathbb{R}$ whose sum is not quasiconcave;

(c) a function that is both quasiconcave and convex, but not concave.

22.3 Consider a function $f : C \to \mathbb{R}$ on a convex domain C in $\mathbb{R}^n$ whose range we denote by $R = \{f(x) : x \in C\}$ and a strictly increasing function $g : R \to \mathbb{R}$. Are the following claims necessarily true?

(a) If f is concave, then the composition $g \circ f$ is concave.

(b) If f is quasiconcave, then the composition $g \circ f$ is quasiconcave.

22.4 Give an example of a quasiconcave function with a local maximum that is not a global maximum.

22.5 Let $f : C \to \mathbb{R}$ be a quasiconcave function on a convex domain $C \subseteq \mathbb{R}^n$. Assume $x \in C$ is a strict local maximum, i.e., $f(x) > f(y)$ for all other $y \in C$ in a neighborhood U of x.

(a) Show that x is also a global maximum.

(b) Can f have other global maxima than x?

22.6 (Subgradients imply convexity) Let $f : C \to \mathbb{R}$ be a function on an open domain C in $\mathbb{R}^n$. Suppose that f has a subgradient at each point in its domain, i.e., for each $z \in C$ there is a vector $a_z \in \mathbb{R}^n$ such that

$$\text{for all } x \in C: \qquad f(x) \geq f(z) + a_z^\top (x - z).$$

Prove that f is a convex function. HINT: Show that f is the pointwise supremum of the affine functions $x \mapsto f(z) + a_z^\top (x - z)$ and use Theorem 22.6.

22.7 Prove expression (68).

22.8 (Which quasiconcave functions are concave?) We saw that each concave function is quasiconcave, but that the converse is false. Here we show that a quasiconcave function is concave precisely when it remains quasiconcave after tilting it (by adding some linear function).

Formally, let $f : C \to \mathbb{R}$ be a quasiconcave function on a convex domain $C \subseteq \mathbb{R}^n$. Show that f is concave if and only if for each vector $a \in \mathbb{R}^n$, the function $g : C \to \mathbb{R}$ with $g(x) = f(x) + a^\top x$ is quasiconcave.

23 Differentiability

We assume familiarity with differentiation of real-valued functions of a single real variable: a function $f : \mathbb{R} \to \mathbb{R}$ is differentiable at a point $x \in \mathbb{R}$ if there is a number $f'(x)$ such that

$$\lim_{h \to 0} \frac{f(x+h) - f(x)}{h} = f'(x). \tag{73}$$

The number $f'(x)$ is referred to as the derivative of f at x. We can rewrite the definition as

$$\lim_{h \to 0} \frac{f(x+h) - f(x) - f'(x)h}{h} = 0. \tag{74}$$

This expression tells us that if we want to guess the impact on the function value if we start at x and change it by h, then a reasonable estimate would be to take the 'slope' $f'(x)$ of the function and multiply it by h: the difference goes to zero faster than h does.

If h is small, the linear function $h \mapsto f'(x)h$ provides a good estimate of $f(x+h) - f(x)$.

These ideas behind the slope and linear approximation extend to distinct notions of differentiability of a function f of several real variables $x_1, \ldots, x_n$:

- ☒ If we change only one variable, keeping all others fixed, we obtain *partial* derivatives;
- ☒ If we look at changes in a specific direction, we obtain *directional* derivatives;
- ☒ If we allow arbitrary small changes in the variables, we obtain the function's *derivative* and call the function differentiable.

Remark 23.1 (Differentiability assumptions) Partial derivatives assume that we can change the value of one variable, keeping the others fixed. If we can't do that — for instance if the function's domain is a circle — it makes no sense to talk about partial derivatives. Similar comments apply to the other types of differentiation. So when discussing them, we explicitly need to assume that it is possible to make certain small changes in a function's variables. Rather than writing down such an assumption in each and every result separately, let's decide right here that throughout this section we are in points of the function's domain where it is feasible to talk about the relevant notions of differentiability. A common sufficient condition is that such points lie in the interior of the domain. ◁

The big picture (and our agenda for this section) is that these three notions are increasingly demanding:

$$\begin{array}{c} f \text{ is differentiable} \\ \Downarrow \\ f \text{ has directional derivatives in each direction} \\ \Downarrow \\ f\text{'s partial derivatives with respect to each variable } x_1, \ldots, x_n \text{ exist} \end{array}$$

and the implications in the other direction are false in general, but true under additional assumptions.

23.1 Partial derivatives and the gradient

Let $f : X \to \mathbb{R}$ be a function on a domain $X \subseteq \mathbb{R}^n$ of n real variables. Since we know how to differentiate functions of a single variable, the idea behind partial derivatives is to do exactly that: we simply keep all but one of the variables (say x_i) fixed and differentiate the resulting function of just one variable x_i. So we change x_i by an amount h, look at the difference quotient

$$\frac{f(x_1, \ldots, x_{i-1}, x_i + h, x_{i+1}, \ldots, x_n) - f(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_n)}{h}$$

and see what happens as h tends to zero. This notation is hardly pleasing on the eye and rather time-consuming to write. Fortunately, our notation for the standard basis vector

$$e_i = (0,\ldots,0,\underbrace{1}_{\text{coord. } i},0,\ldots,0)$$

comes to the rescue. Adding a number h to the i-th coordinate of vector x is the same as adding the vector he_i. With this, our definition becomes:

Definition 23.1 If it exists, the ***partial derivative*** of f with respect to its i-th variable at a point x in its domain is the number

$$\frac{\partial f(x)}{\partial x_i} = \lim_{h\to 0} \frac{f(x+he_i)-f(x)}{h}.$$

Other notations for the partial derivative include $f_i'(x)$, $f_{x_i}'(x)$, and for functions with values of the form $f(x,y,z)$, the self-explanatory $\partial_x f(x,y,z)$, $\partial_y f(x,y,z)$, $\partial_z f(x,y,z)$.

The corresponding row vector of all n partial derivatives is called the ***gradient*** of f and is denoted

$$\nabla f(x) = \left(\frac{\partial f(x)}{\partial x_1},\ldots,\frac{\partial f(x)}{\partial x_n}\right).$$

Symbol ∇ is pronounced 'nabla'; this name comes — due to the symbol's shape — from a Hellenistic Greek word $\nu\acute{\alpha}\beta\lambda\alpha$ for a Phoenician harp.

Example 23.1 Consider $f:\mathbb{R}^3\to\mathbb{R}$ with $f(x)=3x_1x_3^2-2x_1^4x_2^6$. Keeping coordinates x_2 and x_3 fixed and differentiating this function with respect to x_1, we see that $\partial f(x)/\partial x_1 = 3x_3^2-8x_1^3x_2^6$. Analogously, we find the partial derivatives with respect to its second and third variables to obtain

$$\nabla f(x) = \left(\frac{\partial f(x)}{\partial x_1},\frac{\partial f(x)}{\partial x_2},\frac{\partial f(x)}{\partial x_3}\right) = \left(3x_3^2-8x_1^3x_2^6,\ -12x_1^4x_2^5,\ 6x_1x_3\right). \qquad \triangleleft$$

23.2 Directional derivatives

Directional derivatives measure a function's slope if you stand at a point x in its domain and move a little bit in a direction d:

Definition 23.2 Let $x\in X\subseteq\mathbb{R}^n$ be a point in the domain of the function $f:X\to\mathbb{R}$ and $d\in\mathbb{R}^n, d\neq\mathbf{0}$ a direction. If it exists, the ***directional derivative of*** f ***at*** x ***in direction*** d is the number

$$D_d f(x) = \lim_{h\to 0}\frac{f(x+hd)-f(x)}{h}.$$

If you look along the i-th coordinate axis, in direction $d=e_i$, then you get precisely the definition of the partial derivative with respect to the function's i-th variable. So if at a certain point x the directional derivatives exist in each direction d it follows that in particular all its partial derivatives exist. The converse is not true: knowing the partial derivatives only tells you about the directional derivatives along the coordinate axes, but nothing about other directions.

Example 23.2 The function $f:\mathbb{R}^2\to\mathbb{R}$ with

$$f(x) = \begin{cases} 0 & \text{if } x_1=0 \text{ or } x_2=0,\\ 1 & \text{otherwise}\end{cases}$$

has value zero at each point along the x_1-axis and the x_2-axis, so its partial derivatives in the origin $x = (0,0)$ are equal to zero. But directional derivatives in any direction $d \in \mathbb{R}^2$ with both d_1 and d_2 distinct from zero do not exist:

$$\frac{f(x+hd)-f(x)}{h} = \frac{f(d_1,d_2)-f(0,0)}{h} = \frac{1-0}{h}$$

diverges to $+\infty$ if h tends to zero from above and to $-\infty$ if it does so from below, so it has no limit. ◁

23.3 Differentiability

The definition of differentiability for a function of n variables closely mimics the one-variable case (74):

Definition 23.3

- ⊠ Let $x \in X \subseteq \mathbb{R}^n$ be a point in the domain of the function $f : X \to \mathbb{R}$. We call f ***differentiable at*** x if there is a (row) vector $f'(x) \in \mathbb{R}^n$ with

$$\lim_{h \to \mathbf{0}} \frac{|f(x+h)-f(x)-f'(x)h|}{\|h\|} = 0. \tag{75}$$

  Vector $f'(x)$ is the ***derivative*** of f at x.
- ⊠ Function f is ***differentiable*** if it is differentiable at each point in its domain.

In (75), the expression $f'(x)h$ is the inner product of $f'(x)$ and h. Our next theorem says that differentiability implies the existence of directional and partial derivatives. According to its second part differentiation boils down to computing partial derivatives: the derivative $f'(x)$ is the gradient and the directional derivative in direction d is $D_d f(x) = \nabla f(x) d$.

Theorem 23.1

Assume that $f : X \to \mathbb{R}$ is differentiable at an interior point x of its domain $X \subseteq \mathbb{R}^n$. Then:

(a) For each direction d, the directional derivative $D_d f(x)$ exists and equals $f'(x)d$.

(b) In particular, $f'(x)$ is simply the gradient $\nabla f(x)$.

Proof: **(a)** Fix a direction $d \in \mathbb{R}^n$, $d \neq \mathbf{0}$. Since f is differentiable at x, there is, for each $\varepsilon > 0$, a number $\delta > 0$ such that each scalar h with $0 < \|hd\| < \delta$ has

$$\frac{|f(x+hd)-f(x)-f'(x)hd|}{\|hd\|} < \frac{\varepsilon}{\|d\|}.$$

Rewriting gives

$$0 < |h| < \frac{\delta}{\|d\|} \qquad \Longrightarrow \qquad \left| \frac{f(x+hd)-f(x)}{h} - f'(x)d \right| < \varepsilon.$$

So the directional derivative is

$$D_d f(x) = \lim_{h \to 0} \frac{f(x+hd)-f(x)}{h} = f'(x)d.$$

(b) The partial derivative $\partial f(x)/\partial x_i$ with respect to coordinate i is the directional derivative in direction $d = e_i$. Substituting this direction in the equation $D_d f(x) = f'(x)d$ from our previous step shows that this partial derivative is the i-th coordinate of $f'(x)$. □

But differentiability is a more demanding requirement than having directional derivatives:

Example 23.3 Consider $f:\mathbb{R}^2\to\mathbb{R}$ with

$$f(x_1,x_2)=\begin{cases}\frac{x_1|x_2|}{\sqrt{x_1^2+x_2^2}} & \text{if } x\neq\mathbf{0},\\ 0 & \text{if } x=\mathbf{0}.\end{cases}$$

At $x=\mathbf{0}$, all directional derivatives exist: for each direction $d\neq(0,0)$ and each $t\neq 0$:

$$\frac{1}{t}\big(f(\mathbf{0}+td)-f(\mathbf{0})\big)=\frac{1}{t}\frac{td_1\,|td_2|}{\sqrt{(td_1)^2+(td_2)^2}}=\frac{t|t|}{t\sqrt{t^2}}\frac{d_1\,|d_2|}{\sqrt{d_1^2+d_2^2}}=\frac{d_1\,|d_2|}{\sqrt{d_1^2+d_2^2}},$$

so the directional derivative is

$$D_df(0,0)=\frac{d_1\,|d_2|}{\sqrt{d_1^2+d_2^2}}. \tag{76}$$

But f is not differentiable at $x=(0,0)$: if it were, then Theorem 23.1 says that the directional derivative must be a linear function of the direction,

$$D_df(0,0)=f'(0,0)d.$$

However, (76) is not a linear function of d. ◁

We saw that differentiability at interior points implies the existence of directional derivatives in each direction. And the latter implies that its partial derivatives exist. Our earlier examples indicate that the implications in the other direction are false. But partial derivatives are easy to compute, so it would be convenient to have a criterion on partial derivatives that implies differentiability. Here it is:

Theorem 23.2

If the partial derivatives of $f:X\to\mathbb{R}$ exist and are continuous at all points in a neighborhood of some $x\in X\subseteq\mathbb{R}^n$ in its domain, then f is differentiable at x.

Proof: Let's prove it for functions of two variables. The n-variable case is the same but notationally more painful. Let U be the mentioned neighborhood of x. If $f'(x)$ exists, Theorem 23.1 says that it must be the gradient $\nabla f(x)$. We need to show that for each $\varepsilon>0$ there is a $\delta>0$ such that all $h\in\mathbb{R}^2$ with $0<\|h\|<\delta$, $x+h\in U$, satisfy

$$|f(x+h)-f(x)-\nabla f(x)h|<\varepsilon\|h\|.$$

Changing variables from x_i+h_i to x_i one at a time we can write

$$\begin{aligned}f(x+h)-f(x)=f(x_1+h_1,x_2+h_2)-f(x_1,x_2+h_2)\\ +f(x_1,x_2+h_2)-f(x_1,x_2).\end{aligned}$$

By the mean-value theorem there is a z_1 between x_1+h_1 and x_1 with

$$f(x_1+h_1,x_2+h_2)-f(x_1,x_2+h_2)=\frac{\partial f}{\partial x_1}(z_1,x_2+h_2)h_1.$$

With a similar expression for the second coordinate we obtain

$$f(x+h)-f(x)=\frac{\partial f}{\partial x_1}(z_1,x_2+h_2)h_1+\frac{\partial f}{\partial x_2}(x_1,z_2)h_2.$$

Since

$$\nabla f(x)h = \frac{\partial f}{\partial x_1}(x_1,x_2)h_1 + \frac{\partial f}{\partial x_2}(x_1,x_2)h_2,$$

the triangle inequality and the fact that for each coordinate i, $|h_i| \le \|h\|$ give

$$|f(x+h) - f(x) - \nabla f(x)h| \le \left(\left| \frac{\partial f}{\partial x_1}(z_1, x_2+h_2) - \frac{\partial f}{\partial x_1}(x_1,x_2) \right| + \left| \frac{\partial f}{\partial x_2}(x_1,z_2) - \frac{\partial f}{\partial x_2}(x_1,x_2) \right| \right) \|h\|.$$

The partial derivatives are continuous and each z_i lies between $x_i + h_i$ and x_i, so there is a $\delta > 0$ such that the term in braces is smaller than ε whenever $0 < \|h\| < \delta$. That inequality finishes our proof. □

If f is differentiable at a point x^*, then for vectors h close to $\mathbf{0}$ the function value $f(x^*+h)$ is close to $f(x^*) + f'(x^*)h$. Replacing x^*+h by x and using $f'(x^*) = \nabla f(x^*)$, it follows that for x near x^*, $f(x)$ is close to

$$f(x^*) + \nabla f(x^*)(x - x^*).$$

The function $x \mapsto f(x^*) + \nabla f(x^*)(x-x^*)$ is called the ***linear approximation*** to f at x^*.

Example 23.4 Function $f : \mathbb{R}^2 \to \mathbb{R}$ with $f(x_1,x_2) = 2x_1 + x_1^2 x_2$ has gradient

$$\nabla f(x_1,x_2) = (2 + 2x_1x_2, x_1^2).$$

In the point $x^* = (1,1)$, we have $f(x^*) = 3$ and $\nabla f(x^*) = (4,1)$, so the linear approximation to f at x^* is the function $\ell : \mathbb{R}^2 \to \mathbb{R}$ with

$$\ell(x_1,x_2) = f(x^*) + \nabla f(x^*)(x-x^*) = 3 + 4(x_1-1) + (x_2-1).$$ ◁

23.4 Differentiable functions are continuous

We just argued that if a function is differentiable at a point in its domain, then near that point the associated linear approximation provides a good fit. And since the linear approximation is continuous, f must be continuous as well:

Theorem 23.3 (Differentiability implies continuity)

If $f : X \to \mathbb{R}$ is differentiable at a point x in its domain $X \subseteq \mathbb{R}^n$, then it is continuous at x.

Proof: For continuity at x we need to show that for each $\varepsilon > 0$ there is a $\delta > 0$ such that for each $y \in X$:

$$\|y - x\| < \delta \qquad \Longrightarrow \qquad \big|f(y) - f(x)\big| < \varepsilon.$$

To make this resemble (75), note that y with $\|y-x\| < \delta$ is of the form $y = x+h$ with $h = y-x$ satisfying $\|h\| < \delta$, so this can be rewritten as

$$\|h\| < \delta \qquad \Longrightarrow \qquad \big|f(x+h) - f(x)\big| < \varepsilon.$$

Let $\varepsilon > 0$. By differentiability there is a δ with $0 < \delta < \varepsilon/(1 + \|f'(x)\|)$ and

$$0 < \|h\| < \delta \qquad \Longrightarrow \qquad \frac{\big|f(x+h) - f(x) - f'(x)h\big|}{\|h\|} < 1.$$

The triangle inequality implies that for $\|h\| < \delta$:

$$\begin{aligned} \big|f(x+h) - f(x)\big| &\le \big|f(x+h) - f(x) - f'(x)h\big| + \big|f'(x)h\big| \\ &< \|h\| + \big|f'(x)h\big| \le \|h\| + \|f'(x)\|\|h\| = \|h\|(1 + \|f'(x)\|) \\ &< \delta(1 + \|f'(x)\|) < \varepsilon. \end{aligned}$$ □

23.5 Steepest ascent

Consider a function $f : \mathbb{R}^n \to \mathbb{R}$ and vectors $x, d \in \mathbb{R}^n$. Then d is a ***direction of ascent*** at x if small deviations from x in the direction d lead to higher function values. Formally: there is a $\delta > 0$ such that for all $h \in (0, \delta)$: $f(x + hd) > f(x)$. In particular, if the directional derivative

$$\lim_{h \to 0} \frac{f(x + hd) - f(x)}{h}$$

is positive, then d is a direction of ascent. Recall that this directional derivative is the slope of the function f at x in direction d; let us try to find the largest slope:

Theorem 23.4 (The gradient points in the direction of maximal ascent)

If f is differentiable at a point x in the interior of its domain and $\nabla f(x) \neq \mathbf{0}$, then the solution to the problem of finding the largest directional derivative by choosing d with $\|d\| \leq 1$ is given by $\nabla f(x) / \|\nabla f(x)\|$.

Proof: Differentiability, the Cauchy-Schwarz inequality, and the assumption that $\|d\| \leq 1$, give

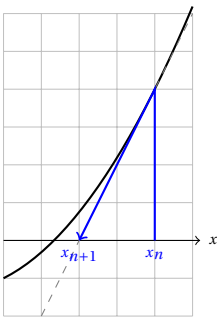
$$D_d f(x) = \nabla f(x) d \leq \|\nabla f(x)\| \, \|d\| \leq \|\nabla f(x)\|,$$

with equality throughout if and only if $d = \nabla f(x) / \|\nabla f(x)\|$. □

Thus, when you try to maximize a function and find a point that doesn't quite do the job, a clever trick is to move a bit in the direction of the gradient to find a better candidate. This is the rationale behind ascent/descent methods in numerical optimization.

23.6 Newton's method

Newton's method (aka the Newton-Raphson method) is an algorithm for solving equations of the form $f(x) = \mathbf{0}$, when f is a differentiable function. We illustrate its logic first for a function $f : \mathbb{R} \to \mathbb{R}$. Suppose that after n iterations of the algorithm we have found a candidate x_n with $f(x_n) \neq 0$. We try to find a correction h to x_n such that $f(x_n + h) = 0$. By differentiability, as long as h is small, $f(x_n + h)$ is reasonably approximated by $f(x_n) + f'(x_n)h$, so ignoring the approximation error and solving $f(x_n) + f'(x_n)h = 0$, we find $h = -f(x_n)/f'(x_n)$, provided of course that we're not dividing by zero. This gives our next candidate



$$x_{n+1} = x_n - f(x_n)/f'(x_n). \tag{77}$$

In our figure, this means drawing the linear approximation to f at x_n (the dotted line) and taking x_{n+1} to be the point where it intersects the horizontal axis. The challenge is to find a suitable starting point and conditions such that the algorithm converges to a desired solution x with $f(x) = \mathbf{0}$.

Example 23.5 (Heron's formula) Consider the equation $f(x) = x^2 - a = 0$ for given $a > 0$. Then (77) becomes

$$x_{n+1} = x_n - \frac{x_n^2 - a}{2x_n} = \frac{2x_n^2 - x_n^2 + a}{2x_n} = \frac{1}{2}\left(x_n + \frac{a}{x_n}\right). \tag{78}$$

This iteration scheme to find approximations for $\sqrt{a}$ is known as Heron's formula and was known to Greek mathematicians in the first century CE.

For each $x_0 > 0$, the sequence of Newton iterates recursively defined by (78) converges to $\sqrt{a}$. Using standard algebra, we find that for all $n \in \mathbb{N}$: $\sqrt{a} \leq x_{n+1} \leq x_n$. Thus, the sequence $x_1, x_2, \ldots$ is weakly decreasing and bounded from below by $\sqrt{a}$. Consequently, it converges to a limit x. To show that $x = \sqrt{a}$, use continuity of Heron's formula to deduce that

$$x = \lim_{n\to\infty} x_n = \lim_{n\to\infty} \frac{1}{2}\left(x_n + \frac{a}{x_n}\right) = \frac{1}{2}\left(x + \frac{a}{x}\right).$$

Rewriting gives $x = \sqrt{a}$. ◁

For a differentiable function $f : \mathbb{R}^n \to \mathbb{R}^n$, the same reasoning leads to the iteration scheme

$$x_{n+1} = x_n - \left[f'(x_n)\right]^{-1} f(x_n),$$

assuming that the inverse of $f'(x_n)$ exists. Establishing conditions for convergence is the domain of numerical mathematics (the socalled Newton-Kantorovich theorem). It is not uncommon in applied economics to take a leap of faith and just apply Newton's method in the hope that it converges.

Exercises section 23

23.1 Consider $f : \mathbb{R}^2 \to \mathbb{R}$ with

$$f(x) = \begin{cases} \frac{x_1^2 x_2}{x_1^4 + x_2^2}\|x\| & \text{if } x \neq \mathbf{0}, \\ 0 & \text{if } x = \mathbf{0}. \end{cases}$$

(a) Show that f is continuous at $x = \mathbf{0}$. HINT: Use that $(x_1^2 - x_2)^2 \geq 0$.

(b) Show that at $x = \mathbf{0}$, function f has a directional derivative in each direction $d \neq \mathbf{0}$ and that the directional derivatives are a linear function of d.

(c) Show that f is not differentiable at $x = \mathbf{0}$. HINT: Approximate the origin via points of the form (a, a^2).

24 Static optimization

In this section, we formulate necessary and sufficient conditions for the solutions of optimization problems under some differentiability assumptions. The archetypical problem is of the form

$$\text{maximize } f(x) \text{ with } x \in F.$$

Here, f is the real-valued ***goal function*** we aim to maximize. The '$x \in F$'-part stresses that the arguments x we can choose from must belong to a feasible set $F \subseteq \mathbb{R}^n$ for some $n \in \mathbb{N}$, which allows us to impose ***constraints*** on which x are actually permissible: only vectors $x \in F$ are ***feasible***.

We develop the theory for maximization problems. To solve a minimization problem, turn it into maximization with a sign change on the goal function: the minimizers of f are the maximizers of $-f$.

The constraints often come in the form of (in)equality constraints like (draw!)

$$x_2 \geq x_1^2 \qquad \text{and} \qquad x_2 - x_1 = 2.$$

For notational simplicity, we write inequality constraints in the form $h(x) \leq 0$ and equality constraints as $g(x) = 0$ for some functions h and g. After minor manipulations, also other restrictions can be rewritten this way. Like the ones above:

$$\underbrace{x_1^2 - x_2}_{h(x)} \leq 0 \qquad \text{and} \qquad \underbrace{x_2 - x_1 - 2}_{g(x)} = 0.$$

Inequality constraint $h(x) \leq 0$ is ***binding*** at a feasible point x if it holds with equality ($h(x) = 0$) and ***nonbinding*** or ***slack*** otherwise ($h(x) < 0$). Equality constraints are, of course, always binding in feasible points. For instance, given the two constraints above:

- ☒ Point $x = (1,1)$ is not feasible: it violates the equality constraint, since $g(1,1) = 1 - 1 - 2 = -2 \neq 0$.
- ☒ Point $x = (2,4)$ is feasible. Both constraints are binding: $h(2,4) = 2^2 - 4 = 0$ and $g(2,4) = 4 - 2 - 2 = 0$.
- ☒ Point $x = (0,2)$ is feasible. The inequality constraint is nonbinding/slack since $h(0,2) = 0^2 - 2 = -2 < 0$. The equality constraint is binding since $g(0,2) = 2 - 0 - 2 = 0$.

Recall that a feasible point x^* is a ***local maximum*** if it has the highest function value among all nearby feasible points:

$$f(x^*) \geq f(x) \text{ for all feasible } x \text{ in a neighborhood of } x^*$$

and a ***global maximum*** if it has the highest function value on the entire feasible set:

$$f(x^*) \geq f(x) \text{ for all } x \in F.$$

Each global maximum is a local maximum as well.

24.1 First-order conditions at interior solutions

Our first result says that if x^* is a local maximum, there is no feasible direction in which the function increases (positive directional derivative). In an interior solution, this gives the usual first-order condition that the function's partial derivatives must be zero.

Theorem 24.1

Let $X \subseteq \mathbb{R}^n$ be a convex set and let $x^* \in X$ be a local maximum of $f : X \to \mathbb{R}$. For any point $x \in X$, if the directional derivative in direction $x - x^*$ exists, it must be nonpositive:

$$D_{x-x^*} f(x^*) \leq 0.$$

In particular, if f is differentiable at x^*, then

$$f'(x^*)(x - x^*) \leq 0.$$

If, moreover, x^* is an interior point of X, the partial derivatives at x^* must be zero: $\nabla f(x^*) = \mathbf{0}$.

Proof: Suppose that for some $x \in X$ the directional derivative $D_{x-x^*} f(x^*)$ — the slope of f if we stand at x^* and look towards x — is positive. Moving a bit in that direction, which is possible since $x^* + \varepsilon(x - x^*) \in U$ for all $\varepsilon \in (0, 1)$ by convexity of X, we find a higher function value, contradicting that x^* is a local maximum.

If f is differentiable at x^*, the reasoning behind Theorem 23.1 tells us that the directional derivative is $f'(x^*)(x - x^*)$. And that the derivative is simply the gradient $\nabla f(x^*)$.

Finally, if x^* is an interior point, we can move slightly in all directions d. Importantly, if d is a feasible direction, then so is $-d$. Applying this to direction $d = e_i$, we see that the i-th coordinate of $\nabla f(x^*)$, i.e., the partial derivative of f with respect to its i-th variable, must be zero. And this holds for all $i = 1, \ldots, n$, so $\nabla f(x^*) = \mathbf{0}$. □

24.2 Problems with inequality constraints: Fritz John and Karush-Kuhn-Tucker conditions

In this subsection we consider a maximization problem with $p \in \mathbb{N}$ inequality constraints:

$$\text{maximize } f(x) \text{ with } h_1(x) \leq 0, \ldots, h_p(x) \leq 0. \tag{79}$$

The real-valued functions $f, h_1, \ldots, h_p$ are defined on a domain $X \subseteq \mathbb{R}^n$ containing all feasible points. We will formulate (Fritz John or Karush-Kuhn-Tucker) conditions that must hold at a maximum. Briefly, the argument is this: if x^* is a maximum, you cannot move in a direction that keeps you inside the feasible set and leads to a higher value of the goal function. With Gordan's theorem (Thm. 20.2), this can be rewritten to the so-called Fritz John conditions for a maximum. Now in detail:

Suppose x^* is a local maximum of problem (79) and that the first r constraints are binding: $h_1(x^*) = \cdots = h_r(x^*) = 0$ and $h_{r+1}(x^*) < 0$, …, $h_p(x^*) < 0$. Rearranging the constraints if necessary, this is without loss of generality. Recall that under suitable differentiability assumptions (e.g., Theorem 23.1) — which we take for granted throughout this discussion — if we stand at x^* and look in direction $d \in \mathbb{R}^n, d \neq \mathbf{0}$, the slope of the function f is given by directional derivative $\nabla f(x^*)d$. If there is a direction d with

$$\nabla f(x^*)d > 0 \qquad \text{but} \qquad \nabla h_1(x^*)d < 0, \ldots, \nabla h_r(x^*)d < 0, \tag{80}$$

then f increases in that direction (positive slope!), but constraints h_1 to h_r decrease (negative slope!). So moving a bit in that direction gives a feasible point with a higher value of the goal function f, contradicting that x^* is a local maximum. Rewriting (80), we know that there is no solution $d \in \mathbb{R}^n$ to

the system of linear inequalities

$$\begin{aligned} \nabla f(x^*)d &> 0 \\ -\nabla h_1(x^*)d &> 0 \\ &\vdots \\ -\nabla h_r(x^*)d &> 0 \end{aligned}$$

By Gordan's theorem, there are numbers $\mu_0, \mu_1, \ldots, \mu_r \geq 0$, not all zero, such that

$$\mu_0 \nabla f(x^*) - \mu_1 \nabla h_1(x^*) - \cdots - \mu_r \nabla h_r(x^*) = \mathbf{0}. \tag{81}$$

These μ's are called ***Lagrange multipliers***. If we also introduce Lagrange multipliers $\mu_{r+1}, \ldots, \mu_p$ for the nonbinding constraints and set them to zero ($\mu_{r+1} = \cdots = \mu_p = 0$), we find that

$$\mu_0 \nabla f(x^*) - \mu_1 \nabla h_1(x^*) - \cdots - \mu_p \nabla h_p(x^*) = \mathbf{0}.$$

and for each constraint j,

$$\mu_j \geq 0 \text{ and } \mu_j h_j(x^*) = 0.$$

Indeed, $\mu_j h_j(x^*) = 0$ for binding constraints because they have $h_j(x^*) = 0$ and for nonbinding constraints because we set the corresponding μ_j to zero. This proves:

Theorem 24.2 (Fritz John conditions for problems with inequality constraints)

If x^* is a local maximum of optimization problem (79), then it satisfies the following ***Fritz John (FJ) conditions***: there are numbers $\mu_0, \mu_1, \ldots, \mu_p \geq 0$, not all zero, with

(gradient condition)	$\mu_0 \nabla f(x^*) - \mu_1 \nabla h_1(x^*) - \cdots - \mu_p \nabla h_p(x^*) = \mathbf{0}$	(82)
(feasibility)	$h_j(x^*) \leq 0$ for all $j = 1, \ldots, p$	(83)
(complementary slackness)	$\mu_j \geq 0$ and $\mu_j h_j(x^*) = 0$ for all $j = 1, \ldots, p$	(84)

To see why conditions (84) are called ***complementary slackness*** conditions, look at the j-th constraint $h_j(x^*) \leq 0$ and its corresponding Lagrange multiplier $\mu_j \geq 0$. The complementary slackness condition $\mu_j h_j(x^*) = 0$ says that $h_j(x^*) = 0$ or $\mu_j = 0$, possibly both. So *at most one* of the two inequalities $h_j(x^*) \leq 0$ and $\mu_j \geq 0$ is nonbinding/slack.

The Fritz John conditions easily give rise to other necessary conditions for a maximum x^*. Recall that the multipliers $\mu_0, \mu_1, \ldots, \mu_p$ are nonnegative and not all equal to zero. Distinguish two cases:

If $\mu_0 = 0$, plug this into expression (81) with the goal function and the binding constraints to find

$$-\mu_1 \nabla h_1(x^*) - \cdots - \mu_r \nabla h_r(x^*) = \mathbf{0},$$

where not all $\mu_1, \ldots, \mu_r$ are zero. So the gradients $\nabla h_1(x^*), \ldots, \nabla h_r(x^*)$ of the binding constraints are linearly dependent.

And if $\mu_0 > 0$, then we can divide all Lagrange multipliers in Theorem 24.2 by μ_0 and see that x^* also satisfies the FJ conditions with rescaled multipliers $\frac{\mu_0}{\mu_0}, \frac{\mu_1}{\mu_0}, \ldots, \frac{\mu_p}{\mu_0}$. In particular, the coefficient in front of the gradient $\nabla f(x^*)$ of the goal function is $\frac{\mu_0}{\mu_0} = 1$. These conditions — the FJ conditions with $\mu_0 = 1$ — are called the Karush-Kuhn-Tucker conditions.

To summarize:

Theorem 24.3 (Karush-Kuhn-Tucker conditions for problems with inequality constraints)

If x^* is a local maximum of optimization problem (79), then the gradients of the binding constraints at x^* are linearly dependent or x^* satisfies the following ***Karush-Kuhn-Tucker (KKT) conditions***: there are numbers $\mu_1, \ldots, \mu_p \geq 0$ with

$$\begin{aligned} &\text{(gradient condition)} && \nabla f(x^*) - \mu_1 \nabla h_1(x^*) - \cdots - \mu_p \nabla h_p(x^*) = \mathbf{0} && (85)\\ &\text{(feasibility)} && h_j(x^*) \leq 0 \text{ for all } j = 1, \ldots, p && (86)\\ &\text{(complementary slackness)} && \mu_j \geq 0 \text{ and } \mu_j h_j(x^*) = 0 \text{ for all } j = 1, \ldots, p && (87) \end{aligned}$$

Under additional assumptions, we can dispense with the linear dependence scenario in the previous theorem and each maximum must satisfy the KKT conditions:

Theorem 24.4

If x^* is a local maximum of optimization problem (79) and at least one of the following three cases is true, then x^* satisfies the KKT conditions.
CASE 1: The gradients of the binding constraints at x^* are linearly independent.
CASE 2: $h_1, \ldots, h_p$ are convex functions and there is an x_0 with $h_1(x_0) < 0, \ldots, h_p(x_0) < 0$.
CASE 3: $h_1, \ldots, h_p$ are affine functions, i.e., the feasible set is of the form $\{x \in \mathbb{R}^n : Ax \leq b\}$ for some matrix A and vector b.

The first case just rephrases Theorem 24.3. The proof for the other two cases is postponed to subsection 24.7. The three cases impose additional assumptions/qualifications on the constraints and are therefore often referred to as ***constraint qualifications***.

So far our theorems about maximization under inequality constraints have given *necessary* conditions for a maximum: if x^* is a maximum then it must be among the candidates satisfying a bunch of conditions (FJ, KKT, ...). But solving those conditions may give spurious candidates that aren't maxima at all. Hence, when you face an optimization problem you always need to argue whether a solution actually exists. The Extreme Value Theorem is a useful tool and so is our next result:

Theorem 24.5

If x^* satisfies the Karush-Kuhn-Tucker conditions of optimization problem (79) and

- ☒ goal function f is concave and differentiable,
- ☒ constraint functions $h_1, \ldots, h_p$ are convex and differentiable,

then x^* is a (global) maximum.

If we introduce four sets of (feasible) points, namely

X_{MAX},	those solving our maximization problem,
X_{FJ},	those satisfying the Fritz John conditions,
X_{LD},	those where the gradients of the binding constraints are linearly dependent,
X_{KKT},	those satisfying the Karush-Kuhn-Tucker conditions,

then Theorems 24.2 and 24.3 can be summarized as follows:

$$X_{MAX} \subseteq X_{FJ} \subseteq X_{LD} \cup X_{KKT}.$$

This gives us two methods (using either FJ or KKT) to find candidate maxima. The latter is more common, so I will spell out the steps in detail:

STEP 1: Find the elements of the set X_{LD}. (Do Exercise 24.1 to practise.)
STEP 2: Write down and solve the KKT conditions to find the set X_{KKT}.
STEP 3: Compute the function value $f(x)$ of each candidate $x \in X_{LD} \cup X_{KKT}$. If the maximization problem has solutions, they are the candidates $x \in X_{LD} \cup X_{KKT}$ with the highest function value.

You can save a lot of time if you can easily verify that you are in case 2 or 3 from Theorem 24.4, in which case you can skip step 1.

The gradient condition

$$\nabla f(x^*) - \mu_1 \nabla h_1(x^*) - \cdots - \mu_p \nabla h_p(x^*) = \mathbf{0}$$

is often stated in terms of an auxiliary function, the ***Lagrangian***, which is defined as

$$\mathscr{L}(x, \mu) = f(x) - \mu_1 h_1(x) - \cdots - \mu_p h_p(x).$$

Indeed, the gradient condition says that in an optimum x^*, the partial derivatives

$$\frac{\partial \mathscr{L}(x^*, \mu)}{\partial x_1}, \ldots, \frac{\partial \mathscr{L}(x^*, \mu)}{\partial x_n}$$

of the Lagrangian with respect to the n coordinates $x_1, \ldots, x_n$ must be zero.

24.3 A worked example with inequality constraints

Consider the problem:

$$\text{maximize } x_1 + \ln(1 + x_2) \text{ with } 16x_1 + x_2 \leq 495, \quad x_1 \geq 0, \quad x_2 \geq 0.$$

Goal function $f : \mathbb{R}^2_+ \to \mathbb{R}$ with $f(x) = x_1 + \ln(1 + x_2)$ is a composition of continuous functions, hence continuous. The feasible set $\{x \in \mathbb{R}^2 : 16x_1 + x_2 \leq 495, x_1 \geq 0, x_2 \geq 0\}$ is nonempty: it contains $(0,0)$, closed, since it is the intersection of three closed halfspaces, and bounded: $0 \leq x_1 \leq 495/16$ and $0 \leq x_2 \leq 495$. So by the Heine-Borel theorem, the feasible set is compact. By the Extreme Value Theorem, a maximum exists.

Rewrite the problem in the standard form (79): maximize $f(x) = x_1 + \ln(1 + x_2)$ subject to $h_1(x) = 16x_1 + x_2 - 495 \leq 0, h_2(x) = -x_1 \leq 0, h_3(x) = -x_2 \leq 0$. The constraints h_1, h_2, h_3 are affine functions, so we are in Case 3 of Theorem 24.4: a maximum must satisfy the KKT conditions (85), (86), and (87). The Lagrangian is

$$\mathscr{L}(x, \mu) = f(x) - \sum_{j=1}^{3} \mu_j h_j(x) = x_1 + \ln(1 + x_2) - \mu_1(16x_1 + x_2 - 495) - \mu_2(-x_1) - \mu_3(-x_2).$$

In a local maximum x, the following KKT-conditions must hold:

☒ Gradient condition: partial derivatives of $\mathscr{L}$ w.r.t. x_1, x_2 are zero:

$$1 - 16\mu_1 + \mu_2 = 0 \tag{88}$$

$$\tfrac{1}{1+x_2} - \mu_1 + \mu_3 = 0 \tag{89}$$

☒ Feasibility:

$$16x_1 + x_2 \leq 495 \tag{90}$$

$$x_1 \geq 0 \tag{91}$$

$$x_2 \geq 0 \tag{92}$$

☒ Complementary slackness:

$$\mu_1, \mu_2, \mu_3 \geq 0 \tag{93}$$
$$\mu_1(16x_1 + x_2 - 495) = 0 \tag{94}$$
$$\mu_2 x_1 = 0 \tag{95}$$
$$\mu_3 x_2 = 0 \tag{96}$$

By (88) and (93): $16\mu_1 = 1 + \mu_2 \geq 1$, so $\mu_1 \geq 1/16$. By (94): $16x_1 + x_2 = 495$.

Now distinguish four cases, depending on whether the nonnegativity constraints are binding:

1. $x_1 = x_2 = 0$: this contradicts $16x_1 + x_2 = 495$.
2. $x_1 = 0, x_2 > 0$. Then $x_2 = 495$ and $\mu_3 = 0$ by (96). By (89): $\mu_1 = 1/496$, contradicting $\mu_1 \geq 1/16$.
3. $x_1 > 0, x_2 = 0$. Then $\mu_2 = 0$ by (95) and $\mu_1 = 1/16$ by (88). But (89) gives $\mu_1 = 1 + \mu_3 \geq 1$, contradicting $\mu_1 = 1/16$.
4. $x_1 > 0, x_2 > 0$. Then $\mu_2 = \mu_3 = 0$ by complementary slackness. Substitution in (88) gives $\mu_1 = 1/16$. Substitution in (89) gives $x_2 = 15$. Substitution in $16x_1 + x_2 = 495$ gives $x_1 = 30$. This gives one candidate: $(x_1, x_2, \mu_1, \mu_2, \mu_3) = (30, 15, 1/16, 0, 0)$.

We showed that there is a maximum. We found only one candidate. Conclude: the function is maximized if $x = (30, 15)$ and the maximal value is $f(30, 15) = 30 + \ln(16)$.

24.4 Problems with mixed constraints

Before, we established optimality conditions for problems with inequality constraints. Here, we also allow equality constraints. Having two types of constraints, such problems are often called problems with mixed constraints. In standard form:

$$\text{maximize } f(x) \text{ with } g_1(x) = 0, \ldots, g_m(x) = 0, h_1(x) \leq 0, \ldots, h_p(x) \leq 0, \tag{97}$$

where $f, g_1, \ldots, g_m, h_1, \ldots, h_p$ are real-valued functions on a domain $X \subseteq \mathbb{R}^n$ containing all feasible points. We assume that these functions are differentiable on a neighborhood of each feasible point.

This problem formulation allows problems with only equality constraints ($p = 0$), only inequality constraints ($m = 0$), both types of constraints ($p, q \neq 0$), or unconstrained optimization ($p = m = 0$).

To state the corresponding Fritz John conditions we again associate Lagrange multipliers μ_0 with the goal function f and $\mu_1, \ldots, \mu_p$ with the inequality constraints $h_1(x) \leq 0, \ldots, h_p(x) \leq 0$. And we introduce new multipliers $\lambda_1, \ldots, \lambda_m$ for the equality constraints $g_1(x) = 0, \ldots, g_m(x) = 0$.

Theorem 24.6 (Fritz John conditions for problems with mixed constraints)

If x^* is a local maximum of optimization problem (97), then it satisfies these Fritz John conditions: there are numbers $\lambda_1, \ldots, \lambda_m$ and $\mu_0, \mu_1, \ldots, \mu_p \geq 0$, not all zero, satisfying gradient condition

$$\mu_0 \nabla f(x^*) - \lambda_1 \nabla g_1(x^*) - \cdots - \lambda_m \nabla g_m(x^*) - \mu_1 \nabla h_1(x^*) - \cdots - \mu_p \nabla h_p(x^*) = \mathbf{0}, \tag{98}$$

and

$$\text{(feasibility)} \qquad g_1(x^*) = 0, \ldots, g_m(x^*) = 0, h_1(x^*) \leq 0, \ldots, h_p(x^*) \leq 0 \tag{99}$$
$$\text{(compl. slackness)} \qquad \mu_j \geq 0 \text{ and } \mu_j h_j(x^*) = 0 \text{ for all } j = 1, \ldots, p \tag{100}$$

The proofs for this subsection are less insightful, so I postpone them to subsection 24.7. Multipliers

μ_j for the inequality constraints $h_j(x) \leq 0$ are nonnegative ($\mu_j \geq 0$), but multipliers λ_i for the equality constraints $g_i(x) = 0$ have no sign restriction ($\lambda_i \in \mathbb{R}$). Also, there is no complementary slackness condition $\lambda_i g_i(x^*) = 0$ for the equality constraints: we already know that must be the case by feasibility ($g_i(x^*) = 0$).

Remark 24.1 (Problems with equality constraints) In a problem with only equality constraints,

$$\text{maximize } f(x) \text{ with } g_1(x) = 0, \ldots, g_m(x) = 0, \tag{101}$$

these conditions simplify considerably, because there are no complementary slackness conditions: if x^* is a maximum of (101), then it is feasible and there are Lagrange multipliers $\mu_0 \geq 0$ and $\lambda_1, \ldots, \lambda_m$, not all zero, satisfying the gradient condition

$$\mu_0 \nabla f(x^*) - \lambda_1 \nabla g_1(x^*) - \cdots - \lambda_m \nabla g_m(x^*) = \mathbf{0}.$$

This special case of Theorem 24.6 is sometimes called ***Lagrange's theorem***. ◁

Arguing as before, it suffices to verify the Fritz John conditions for μ_0 equal to zero or one. This gives the following generalization of Theorem 24.3:

Theorem 24.7

If x^* is a local maximum of optimization problem (97), then the gradients of the binding constraints at x^* are linearly dependent or there are numbers $\lambda_1, \ldots, \lambda_m$ and $\mu_1, \ldots, \mu_p \geq 0$ such that

$$\nabla f(x^*) - \lambda_1 \nabla g_1(x^*) - \cdots - \lambda_m \nabla g_m(x^*) - \mu_1 \nabla h_1(x^*) - \cdots - \mu_p \nabla h_p(x^*) = \mathbf{0}, \tag{102}$$

and

$$\begin{aligned} &\text{(feasibility)} && g_1(x^*) = 0, \ldots, g_m(x^*) = 0, h_1(x^*) \leq 0, \ldots, h_p(x^*) \leq 0 && (103)\\ &\text{(compl. slackness)} && \mu_j \geq 0 \text{ and } \mu_j h_j(x^*) = 0 \text{ for all } j = 1, \ldots, p && (104) \end{aligned}$$

By definition, the equality constraints are always binding (hold with equality). With the ***Lagrangian***

$$\mathscr{L}(x, \lambda, \mu) = f(x) - \lambda_1 g_1(x) - \cdots - \lambda_m g_m(x) - \mu_1 h_1(x) - \cdots - \mu_p h_p(x),$$

the gradient restriction says that the partial derivatives of the Lagrangian with respect to the coordinates of vector x must be zero.

With this theorem we once again have a three-step recipe for finding candidate maxima:
STEP 1: Find all feasible points where the gradients of the binding constraints are linearly dependent.
STEP 2: Write down conditions (102), (103), (104) and find all points solving them.
STEP 3: Compute the function value $f(x)$ of all candidates x from the previous two steps. If the maximization problem has solutions, they are the candidates with the highest function value.

Conditions (102), (103), (104) are the Fritz John conditions with $\mu_0 = 1$. With some extra structure on the constraints, only those conditions are necessary and you can skip step 1:

Theorem 24.8 (Fritz John conditions under concave/affine constraints: $\mu_0 = 1$)

Let x^* be a local maximum of (97). If the equality constraints $g_1, \ldots, g_m$ are affine functions and the inequality constraints $h_1, \ldots, h_p$ are concave, then x^* satisfies the Fritz John conditions with $\mu_0 = 1$.

In particular, the theorem above implies that the Fritz John conditions with $\mu_0 = 1$ are necessary for local

maxima of optimization problems with linear constraints, like the maximization of a utility function over a budget set of the form $B(p,w) = \{x \in \mathbb{R}^n : p^\top x \leq w, x \geq \mathbf{0}\}$. One final observation:

Theorem 24.9 (Sufficient conditions: maximizing the Lagrangian)

If (x^*, λ, μ) solves the Fritz John conditions with $\mu_0 = 1$ and x^* maximizes the corresponding Lagrangian, i.e.,

$$\mathscr{L}(x^*, \lambda, \mu) \geq \mathscr{L}(x, \lambda, \mu) \text{ for all feasible } x,$$

then x^* also solves the optimization problem (97).

24.5 A first worked example with mixed constraints

Consider the problem:

$$\text{maximize } x_1^2 - 3x_2^2 \text{ with } x_1^2 + x_2^2 = 17, \quad x_1 - x_2 \leq 3.$$

In standard notation, the problem becomes maximize $f(x) = x_1^2 - 3x_2^2$ with $g(x) = x_1^2 + x_2^2 - 17 = 0$ and $h(x) = x_1 - x_2 - 3 \leq 0$. Since there is only one equality constraint and one inequality constraint, I simplify notation by omitting the subscript 1 in $g_1(x), h_1(x)$, etc. Using the Extreme Value Theorem, you can argue that a maximum exists. We follow the three-step algorithm from page 127:
STEP 1: Are there feasible points where the gradients of the binding constraints are linearly dependent? Distinguish two cases:

1. Only the equality constraint $g(x) = x_1^2 + x_2^2 - 17 = 0$ is binding. Its gradient $\nabla g(x) = (2x_1, 2x_2)$ is equal to the zero vector only in the point $x = \mathbf{0}$, which is not feasible.

2. Both constraints are binding: $g(x) = x_1^2 + x_2^2 - 17 = 0$ and $h(x) = x_1 - x_2 - 3 = 0$. So $x_2 = x_1 - 3$. Substitution in the first constraint gives

   $$x_1^2 + (x_1 - 3)^2 = 2x_1^2 - 6x_1 + 9 = 17 \quad \Leftrightarrow \quad 2(x_1^2 - 3x_1 - 4) = 2(x_1 - 4)(x_1 + 1) = 0 \quad \Leftrightarrow \quad x_1 = 4 \text{ or } x_1 = -1.$$

   This gives two points: $x = (4,1)$ with gradients $\nabla g(4,1) = (8,2)$ and $\nabla h(4,1) = (1,-1)$, which are linearly independent, or the point $x = (-1,-4)$ with gradients $\nabla g(-1,-4) = (-2,-8)$ and $\nabla h(-1,-4) = (1,-1)$, which are linearly independent.

Conclude: no feasible points where the gradients of the binding constraints are linearly dependent.
STEP 2: We write down and solve the Fritz John conditions (102), (103), and (104) with $\mu_0 = 1$. The Lagrangian is

$$\mathscr{L}(x, \lambda, \mu) = f(x) - \lambda g(x) - \mu h(x) = x_1^2 - 3x_2^2 - \lambda(x_1^2 + x_2^2 - 17) - \mu(x_1 - x_2 - 3).$$

In an optimum, the following conditions must be satisfied:

☒ Gradient condition: partial derivatives of the Lagrangian w.r.t. x_1 and x_2 are zero:

$$2x_1 - 2\lambda x_1 - \mu = 0 \tag{105}$$
$$-6x_2 - 2\lambda x_2 + \mu = 0 \tag{106}$$

☒ Feasibility:

$$x_1^2 + x_2^2 = 17 \tag{107}$$
$$x_1 - x_2 \leq 3 \tag{108}$$

☒ Complementary slackness for the inequality constraint:

$$\mu \geq 0 \tag{109}$$
$$\mu(x_1 - x_2 - 3) = 0 \tag{110}$$

Case 1: $\mu = 0$. Equations (105) and (106) then give

$$\begin{array}{rcl} 2x_1(1-\lambda) = 0 & \iff & x_1 = 0 \text{ or } \lambda = 1 \\ -2x_2(3+\lambda) = 0 & \iff & x_2 = 0 \text{ or } \lambda = -3. \end{array}$$

This gives four possibilities:

1. $x_1 = 0$ and $x_2 = 0$. This is not feasible: it contradicts (107).
2. $x_1 = 0$ and $\lambda = -3$. Then (107) gives $x_2^2 = 17$, so $x_2 = -\sqrt{17}$ or $x_2 = \sqrt{17}$. The first violates (108), the second gives candidate solution $(x_1, x_2, \lambda, \mu) = (0, \sqrt{17}, -3, 0)$ with function value $f(0, \sqrt{17}) = -51$.
3. $\lambda = 1$ and $x_2 = 0$. Then (107) gives $x_1^2 = 17$, so $x_1 = -\sqrt{17}$ or $x_1 = \sqrt{17}$. The second violates (108), the first gives candidate solution $(x_1, x_2, \lambda, \mu) = (-\sqrt{17}, 0, 1, 0)$ with function value $f(-\sqrt{17}, 0) = 17$.
4. $\lambda = 1$ and $\lambda = -3$. This is not feasible: λ can't be both at the same time.

Case 2: $\mu > 0$. Complementary slackness (110) gives $x_1 - x_2 = 3$, so $x_2 = x_1 - 3$. Substitution in (107) gives

$$x_1^2 + (x_1 - 3)^2 = 2x_1^2 - 6x_1 + 9 = 17 \quad \Leftrightarrow \quad 2(x_1^2 - 3x_1 - 4) = 2(x_1 - 4)(x_1 + 1) = 0 \quad \Leftrightarrow \quad x_1 = 4 \text{ or } x_1 = -1.$$

This gives two possibilities:

1. $x_1 = 4$. Then $x_2 = x_1 - 3 = 1$. Substitution in (105) and (106) gives a system of two linear equations with two unknowns, which we can solve by Gaussian elimination:

$$\begin{aligned} 8 - 8\lambda - \mu &= 0 \\ -6 - 2\lambda + \mu &= 0 \end{aligned}$$

   with solution $(\lambda, \mu) = (1/5, 32/5)$. So we have candidate solution $(x_1, x_2, \lambda, \mu) = (4, 1, 1/5, 32/5)$ with function value $f(4, 1) = 13$.

2. $x_1 = -1$. Then $x_2 = x_1 - 3 = -4$. Substitution in (105) and (106) gives a system of two linear equations with two unknowns, which we can solve by Gaussian elimination:

$$\begin{aligned} -2 + 2\lambda - \mu &= 0 \\ 24 + 8\lambda + \mu &= 0 \end{aligned}$$

   with solution $(\lambda, \mu) = (-11/5, -32/5)$. This violates (109).

STEP 3: Comparing all solution candidates, we find maximal value 17 in feasible point $(x_1, x_2) = (-\sqrt{17}, 0)$.

24.6 A second worked example with mixed constraints

Consider the problem:

$$\text{maximize } -x_1^2 - x_2^2 - \cdots - x_n^2 \qquad \text{subject to } x_1 \geq 0, \ldots, x_n \geq 0, \quad x_1 + \cdots + x_n = 1.$$

Since $-x_1^2 - x_2^2 - \cdots - x_n^2 = -\|x\|^2$, we are searching for the shortest vector in $\mathbb{R}^n$ with nonnegative coordinates summing to one.

In standard notation, the problem becomes maximize $f(x) = -x_1^2 - x_2^2 - \cdots - x_n^2$ with $g(x) = x_1 + \cdots + x_n - 1 = 0$ and $h_i(x) = -x_i \leq 0$ for all coordinates $i = 1, \ldots, n$. Since there is only one equality constraint, I simplify notation by omitting the subscript 1 in $g_1(x)$. By the Extreme Value Theorem (verify this yourself) the problem has a solution. Moreover, the inequality constraints are linear, the equality constraint is affine, so Theorem 24.8 says that a solution must satisfy a solution must satisfy the Fritz John conditions (102), (103), and (104) with $\mu_0 = 1$.

Assigning multiplier λ to the equality constraint $g(x) = 0$ and μ_i to the inequality constraint $h_i(x) \leq 0$, the Lagrangian is

$$\mathscr{L}(x, \lambda, \mu) = f(x) - \lambda g(x) - \sum_{i=1}^{n} \mu_i h_i(x) = -x_1^2 - x_2^2 - \cdots - x_2^2 - \lambda(x_1 + \cdots + x_n - 1) + \mu_1 x_1 + \cdots + \mu_n x_n.$$

In an optimum, the following conditions must be satisfied:

☒ Partial derivatives of the Lagrangian w.r.t. x_i are zero, i.e., for each $i = 1, \ldots, n$:

$$-2x_i - \lambda + \mu_i = 0. \tag{111}$$

☒ Feasibility:

$$x_1 + \cdots + x_n = 1, \tag{112}$$

$$x_1 \geq 0, \ldots, x_n \geq 0. \tag{113}$$

☒ Complementary slackness for the inequality constraints, i.e., for each $i = 1, \ldots, n$:

$$\mu_i \geq 0 \qquad \text{and} \qquad \mu_i x_i = 0. \tag{114}$$

By (112) and (113), *some* coordinate i must have $x_i > 0$. By (114), $\mu_i = 0$. By (111), $\lambda = -2x_i < 0$. But then *all* coordinates must have $x_j > 0$: otherwise $x_j = 0$ and (111) would give $\mu_j = \lambda < 0$, contradicting (114). Since all coordinates of x are positive, all μ_i are zero by complementary slackness, so all coordinates of x are equal by (111). Since they add up to one, we find $x = (1/n, \ldots, 1/n)$.

We found only one solution candidate, $x = (1/n, \ldots, 1/n)$. We also argued that there must be a solution. So the function is maximized in $x = (1/n, \ldots, 1/n)$. Its maximal value is $-(1/n)^2 - \cdots - (1/n)^2 = -1/n$.

24.7 Postponed proofs

24.7.1 Proof of Theorem 24.4

As mentioned, the first case just restates the preceding theorem. So it remains to prove:
CASE 2: For each constraint $j = 1, \ldots, p$, Theorem 22.11 gives

$$0 > h_j(x_0) \geq h_j(x^*) + \nabla h_j(x^*)(x_0 - x^*).$$

Assume, without loss of generality, that $h_1(x^*) = \cdots = h_r(x^*) = 0$ and $h_{r+1}(x^*) < 0, \ldots, h_p(x^*) < 0$. Then

$$\nabla h_j(x^*)(x_0 - x^*) < 0$$

for $j = 1, \ldots, r$. If there is no nonnegative solution $\mu_1, \ldots, \mu_r$ to the linear equations

$$\nabla f(x^*) - \mu_1 \nabla h_1(x^*) - \cdots - \mu_r \nabla h_r(x^*) = \mathbf{0},$$

there is by Farkas' Lemma (Theorem 20.1; with a sign change) a vector y with

$$\begin{aligned} \nabla h_1(x^*) y &\leq 0 \\ &\vdots \\ \nabla h_r(x^*) y &\leq 0 \\ \nabla f(x^*) y &> 0 \end{aligned}$$

Moving a little bit from x^* in:

- ⊠ direction $x_0 - x^*$ makes h_j ($j = 1, \ldots, r$) smaller since $\nabla h_j(x^*)(x_0 - x^*) < 0$;
- ⊠ direction y makes f larger since $\nabla f(x^*) y > 0$.

Combine these directions into one: for $\varepsilon \in (0,1)$ sufficiently small, direction $d = (1-\varepsilon)y + \varepsilon(x_0 - x^*)$ satisfies $\nabla h_j(x^*) d < 0$ for $j = 1, \ldots, r$ (regardless of ε) and $\nabla f(x^*) d > 0$. As in the proof of Theorem 24.2, moving slightly in direction d gives a feasible point with a function value higher than $f(x^*)$, contradicting that x^* is a local maximum.

CASE 3 is virtually identical to the previous one. An affine constraint $h_j(x) = \alpha_1 x_1 + \cdots + \alpha_n x_n + \beta \leq 0$ can be written as $h_j(x) = \nabla h_j(x^*) x + \beta \leq 0$. By Farkas' Lemma, if there is no nonnegative solution $\mu_1, \ldots, \mu_r \geq 0$ to

$$\nabla f(x^*) - \mu_1 \nabla h_1(x^*) - \cdots - \mu_r \nabla h_r(x^*) = \mathbf{0},$$

there is a $y \in \mathbb{R}^n$ with $\nabla f(x^*) y > 0, \nabla h_1(x^*) y \leq 0, \ldots, \nabla h_r(x^*) y \leq 0$. By affinity, $h_j(x^* + ty) = h_j(x^*) + t \nabla h_j(x^*) y$, so $x^* + ty$ is feasible for small $t > 0$, but has a function value higher than $f(x^*)$, contradicting that x^* is a local maximum.

24.7.2 Proof of Theorem 24.5

Let x^* satisfy the KKT conditions with nonnegative multipliers $\mu_1, \ldots, \mu_p$. Let x be feasible. To see that x^* is a maximum, we show that $f(x) - f(x^*) \leq 0$:

$$\begin{aligned} f(x) - f(x^*) &\leq \nabla f(x^*)(x - x^*) && \text{(by concavity of } f\text{, Thm. 22.11)} \\ &= \sum_{j=1}^{p} \mu_j \nabla h_j(x^*)(x - x^*) && \text{(by the KKT conditions)} \\ &\leq \sum_{j=1}^{p} \mu_j \big(h_j(x) - h_j(x^*)\big) && \text{(by convexity of } h_j\text{, Thm. 22.11)} \\ &= \sum_{j=1}^{p} \mu_j h_j(x) && \text{(by complementary slackness)} \\ &\leq 0 && \text{(since } \mu_j \geq 0 \text{ and } h_j(x) \leq 0\text{)} \end{aligned}$$

24.7.3 Proof of Theorem 24.6

We derive the result as a consequence of the following more general theorem:

Theorem 24.10 (Fritz John conditions)

Let x^* be a local maximum of (97). Then there exist numbers $\lambda_1,\ldots,\lambda_m$ and $\mu_0,\mu_1,\ldots,\mu_p$, not all zero, such that

(a) $\mu_0\nabla f(x^*)-\sum_{i=1}^m \lambda_i\nabla g_i(x^*)-\sum_{j=1}^p \mu_j\nabla h_j(x^*)=\mathbf{0}$.

(b) $\mu_j\geq 0$ for all $j=0,1,\ldots,p$.

(c) In each open neighborhood N of x^*, there is an $x\in N$ with $\lambda_i g_i(x)>0$ for all $i=1,\ldots,m$ and $\mu_j h_j(x)>0$ for all j with $\mu_j\neq 0$.

Proof: STEP 1: We assumed at the beginning of section 24.4 that all functions are differentiable on a neighborhood of the feasible points. Define, for each $k\in\mathbb{N}$, the ***penalty function*** F_k with

$$F_k(x)=f(x)-\frac{1}{2}k\sum_{i=1}^m\big(g_i(x)\big)^2-\frac{1}{2}k\sum_{j=1}^p\Big(h_j^+(x)\Big)^2-\frac{1}{2}\|x-x^*\|^2.$$

Here, $h_j^+(x)=\max\{h_j(x),0\}$ is the positive part of the function h_j. The first summand assures that we are punished for choosing $g_i(x)\neq 0$, the second summand that we are punished for choosing $h_j(x)>0$. These punishments increase with k. The final term punishes us the further away we move from x^*.

Since x^* is a local maximum, there is an $\varepsilon>0$ such that $f(x^*)\geq f(x)$ for all feasible x in the nonempty, compact set $S=\{x:\|x-x^*\|\leq\varepsilon\}$. Let x^k maximize F_k over S; such an x^k exists by the Extreme Value Theorem (Theorem 17.3).
STEP 2: Sequence $(x^k)_{k\in\mathbb{N}}$ converges to x^*. For each $k\in\mathbb{N}$, x^k maximizes F_k, while x^* is feasible, so $F_k(x^k)\geq F_k(x^*)$:

$$f(x^*)=F_k(x^*)\leq F_k(x^k)=f(x^k)-\frac{1}{2}k\sum_{i=1}^m\Big(g_i(x^k)\Big)^2-\frac{1}{2}k\sum_{j=1}^p\Big(h_j^+(x^k)\Big)^2-\frac{1}{2}\|x^k-x^*\|^2. \tag{115}$$

Rearranging terms and dividing both sides by k gives

$$\frac{f(x^*)}{k}\leq\frac{f(x^k)}{k}-\frac{1}{2}\sum_{i=1}^m\Big(g_i(x^k)\Big)^2-\frac{1}{2}\sum_{j=1}^p\Big(h_j^+(x^k)\Big)^2-\frac{\|x^k-x^*\|^2}{2k}\leq\frac{f(x^k)}{k}. \tag{116}$$

Since f is bounded on the compact set S and $\|x^k-x^*\|\leq\varepsilon$, the left- and right-hand side of (116) converge to zero, so that

$$\lim_{k\to\infty}\sum_{i=1}^m\Big(g_i(x^k)\Big)^2=\lim_{k\to\infty}\sum_{j=1}^p\Big(h_j^+(x^k)\Big)^2=0. \tag{117}$$

Sequence $(x^k)_{k\in\mathbb{N}}$ lies in the bounded set S and consequently has a convergent subsequence. Let $x\in S$ be its limit; by (117), this limit is feasible in the original optimization problem, where x^* is optimal. Combining this with (115) gives

$$f(x)\leq f(x^*)\leq f(x^k)-\frac{1}{2}\|x^k-x^*\|^2.$$

Taking limits as $k\to\infty$ gives that $\|x-x^*\|^2=0$, i.e., $x=x^*$.
STEP 3: Since $x^k\to x^*$ and x^* is an interior point of S, x^k lies in the interior of S for k sufficiently large. So the usual first-order condition is $\nabla F_k(x^k)=\mathbf{0}$. Computing this gradient explicitly gives

$$\nabla f(x^k)-k\sum_{i=1}^m g_i(x^k)\nabla g_i(x^k)-k\sum_{j=1}^p h_j^+(x^k)\nabla h_j(x^k)-(x^k-x^*)=\mathbf{0}.$$

Writing $\lambda_i^k = kg_i(x^k)$ and $\mu_j^k = kh_j^+(x^k) \geq 0$ gives

$$\nabla f(x^k) - \sum_{i=1}^{m} \lambda_i^k \nabla g_i(x^k) - \sum_{j=1}^{p} \mu_j^k \nabla h_j(x^k) - (x^k - x^*) = \mathbf{0}. \tag{118}$$

STEP 4: We find a convergent subsequence. The sequence

$$\left\| \left(1, \lambda^k, \mu^k\right) \right\|^{-1} \left(1, \lambda^k, \mu^k\right) \in \mathbb{R}_+ \times \mathbb{R}^m \times \mathbb{R}_+^p$$

is bounded (all its terms have length one) and consequently has a convergent subsequence with limit $(\mu_0, \lambda, \mu) \in \mathbb{R}_+ \times \mathbb{R}^m \times \mathbb{R}_+^p$. Since its length must be one, not all terms are zero. Since $\mu \in \mathbb{R}_+^p$, (b) holds.

Now divide expression (118) by $\left\| \left(1, \lambda^k, \mu^k\right) \right\|$ and consider the limit of the convergent subsequence. Keeping in mind that $\left\| \left(1, \lambda^k, \mu^k\right) \right\|^{-1} \to \mu_0 \in [0,1]$ and $x^k \to x^*$, we find that

$$\left\| \left(1, \lambda^k, \mu^k\right) \right\|^{-1} (x^k - x^*) \to \mathbf{0},$$

so

$$\mu_0 \nabla f(x^*) - \sum_{i=1}^{m} \lambda_i \nabla g_i(x^*) - \sum_{j=1}^{p} \mu_j \nabla h_i(x^*) = \mathbf{0},$$

proving (a). To prove (c), define $I = \{i : \lambda_i \neq 0\}$ and $J = \{j : \mu_j \neq 0\}$. By convergence (Step 4), we have, for k sufficiently large that $\lambda_i \lambda_i^k > 0$ if $i \in I$ and $\mu_j \mu_j^k > 0$ if $j \in J$. For such k it then follows that $\lambda_i g_i(x^k) > 0$ and $\mu_j h_j^+(x^k) > 0$; the latter in its turn implies that $\mu_j h_j(x^k) > 0$. Since each neighborhood N of x^* contains some point x^k for k sufficiently large, also condition (c) holds. □

To derive Theorem 24.6 we only need to show that complementary slackness is implied by the conditions of Theorem 24.10. If $h_j(x^*) = 0$, this is trivial. If $h_j(x^*) < 0$, then $h_j^+(x^k) = 0$ for large k, so $\mu_j = 0$.

24.7.4 Proof of Theorem 24.8

By Theorem 24.10, there are numbers $\lambda_1, \ldots, \lambda_m$ and $\mu_0, \mu_1, \ldots, \mu_p$, not all zero, such that

$$\mu_0 \nabla f(x^*) - \sum_{i=1}^{m} \lambda_i \nabla g_i(x^*) - \sum_{j=1}^{p} \mu_j \nabla h_j(x^*) = \mathbf{0}.$$

We show that μ_0 cannot be zero. Suppose $\mu_0 = 0$. By linearity of $g_1, \ldots, g_m$ and concavity of $h_1, \ldots, h_p$ we have — for any feasible x — that

$$\begin{aligned} g_i(x) &= g_i(x^*) + \nabla g_i(x^*)(x - x^*) \qquad \text{for } i = 1, \ldots, m, \\ h_j(x) &\leq h_j(x^*) + \nabla h_j(x^*)(x - x^*) \qquad \text{for } j = 1, \ldots, p. \end{aligned}$$

By complementary slackness and the equality constraints being binding by definition, it follows that

$$\begin{aligned} &\sum_{i=1}^{m} \lambda_i g_i(x) + \sum_{j=1}^{p} \mu_j h_j(x) \\ &\leq \sum_{i=1}^{m} \lambda_i g_i(x^*) + \sum_{j=1}^{p} \mu_j h_j(x^*) + \left(\sum_{i=1}^{m} \lambda_i \nabla g_i(x^*) + \sum_{j=1}^{p} \mu_j \nabla h_j(x^*) \right)^\top (x - x^*) \\ &= 0. \end{aligned}$$

Since $\mu_0 = 0$ and not all multipliers are zero, it must be that $\lambda_i \neq 0$ for some i or $\mu_j > 0$ for some j. By Theorem 24.10(c), there is a feasible point x with $\lambda_i g_i(x) > 0$ for all i with $\lambda_i \neq 0$ and $\mu_j g_j(x) > 0$ for all j with $\mu_j > 0$. It follows that such an x satisfies $\sum_{i=1}^{m} \lambda_i g_i(x) + \sum_{j=1}^{p} \mu_j h_j(x) > 0$, contradicting the inequality above.

24.7.5 Proof of Theorem 24.9

For all feasible x, writing out that $\mathscr{L}(x^*, \lambda, \mu) \geq \mathscr{L}(x, \lambda, \mu)$ and rearranging terms gives

$$f(x^*) - f(x) \geq \sum_{i=1}^{m} \lambda_i \big(g_i(x^*) - g_i(x)\big) + \sum_{j=1}^{p} \mu_j \big(h_j(x^*) - h_j(x)\big). \tag{119}$$

By feasibility of x^* and x: $g_i(x^*) = g_i(x) = 0$ for all $i = 1, \ldots, m$ and $h_j(x) \leq 0$ for all $j = 1, \ldots, p$. By complementary slackness, $\mu_j \geq 0$ and $\mu_j h_j(x^*) = 0$ for all $j = 1, \ldots, p$. Combining all this, the right side of (119) is nonnegative. Hence, $f(x^*) \geq f(x)$ for all feasible x, making x^* the desired maximum.

Exercises section 24

24.1 Suppose that the following inequalities define the feasible set of an optimization problem with two variables. Rewrite the inequalities in standard form $h_j(x) \leq 0$ and find the set X_{LD} of feasible points where the gradients of the binding constraints are linearly dependent. HINT: distinguish cases depending on which/how many constraints are binding.

(a) $x_2 \geq x_1^3, x_2 \leq 3x_1 + 2$ (sketch the feasible set);

(b) $0 \leq x_2 \leq x_1^3, x_2 \geq 2x_1^3 - 6x_1^2 + 12x_1 - 8$

24.2 Solve the following problems:

(a) maximize $(x_1 + x_2)^2 + 2x_1 + x_2^2$ with $x_1 \geq 0, x_2 \geq 0, x_1 + 3x_2 \leq 4, 2x_1 + x_2 \leq 3$.

(b) minimize $4x_1 - 3x_2$ with $4 - x_1 - x_2 \geq 0, x_2 + 7 \geq 0, -(x_1 - 3)^2 + x_2 \geq -1$.

(c) maximize $x_2 - 2x_1^3 + 2x_1^2 - x_1$ with $0 \leq x_2 \leq x_1^3, x_2 \geq 2x_1^3 - 6x_1^2 + 12x_1 - 8$.

24.3 Consider the problem: minimize $x_1^2 - x_2^2 + 4x_3^2$ with $x_2 \geq -1, x_1 + x_3 \geq 1, x_3 \geq -10$.

(a) Rewrite as a maximization problem in standard form and verify whether the Karush-Kuhn-Tucker conditions are satisfied in $x = (4/5, 0, 1/5)$, in $x = (4/5, -1, 1/5)$, and in $x = (2, -1, -1)$.

(b) Is any of these three points a solution to the problem?

24.4 Consider the problem: minimize $(x_1 - x_2 + x_3)^2$ with $x_1 + 2x_2 - x_3 = 5, x_1 - x_2 - x_3 = -1$.

(a) Rewrite as a maximization problem in standard form and verify whether the Fritz John conditions are satisfied in the point $x = (3/2, 2, 1/2)$.

(b) Is this point a solution to the problem?

24.5 Suppose we optimize $x_1^2 + x_2^2 + x_3^2$ with the restriction $x_3 - x_1 x_2 = 5$.

(a) There is no maximum. Why?

(b) There is a minimum. Why?

(c) Find the minima using the Fritz John conditions.

(d) Find the minima using $x_3 = 5 + x_1 x_2$ and rewriting it as an unconstrained optimization problem over two variables, x_1 and x_2.

24.6 Use the Fritz John conditions to find a rectangle with perimeter equal to one and with maximal area.

24.7 Find, if possible, the maxima and minima of the following problems:

(a) optimize $f(x_1, x_2) = x_1^2 + 6x_1 x_2 + 4x_2^2$ subject to $x_1^2 + 4x_2^2 = 72$.

(b) optimize $f(x_1, x_2, x_3) = x_1^3 + x_2 x_3$ subject to $x_1 - x_2 = -1, x_1 - 2x_2 + x_3 = -3$.

24.8 Find, if possible, the maxima and minima of the following problems:

(a) optimize $f(x_1, x_2, x_3) = x_1^2 + 2x_2^2 + 2x_3^2$ subject to $x_1^2 + 4x_2^2 = 4, x_1 + 2x_3 = 2$.

(b) optimize $f(x_1, x_2, x_3, x_4) = x_1^3 + x_2 + x_3^2 + 3x_4$ subject to $x_2 x_3 = -2, x_1^2 + x_4^2 = 1$.

24.9 Solve the following optimization problems:

(a) maximize $f(x_1, x_2) = 2x_1 - x_1^2 + x_2$ subject to $3x_1 - 2x_2 \leq 6, x_1 + x_2 \leq 3, x_1 \geq 0, x_2 \geq 0$.

(b) maximize $f(x_1, x_2, x_3) = x_1^3 + x_2$ subject to $x_1 + x_3^2 \leq 1, x_1^2 + x_2^2 \leq 2/3$.

(c) maximize $f(x_1, x_2, x_3) = x_1 + 2x_2$ subject to $x_1^2 + x_2^2 + x_3^2 \leq 5, x_1^2 + x_3^2 \leq 1$.

25 Mapping theorems and sensitivity analysis

25.1 Differentiability of vector-valued functions

In Section 23, a function $f : X \to \mathbb{R}$ of several real variables ($X \subseteq \mathbb{R}^n$) was called differentiable at a point $x \in X$ if it had a good linear approximation, i.e., if there is a (row) vector $f'(x)$ with

$$\lim_{h \to \mathbf{0}} \frac{|f(x+h) - f(x) - f'(x)h|}{\|h\|} = 0.$$

Moreover, at interior points, this derivative was simply the gradient $\nabla f(x)$, the vector of partial derivatives (Thm. 23.1). This extends straightforwardly to *vector-valued* functions $f : X \to \mathbb{R}^m$ of several real variables ($X \subseteq \mathbb{R}^n$): again we insist on a good linear approximation. A linear function from $\mathbb{R}^n$ to $\mathbb{R}^m$ can be written as matrix multiplication for a suitable $m \times n$ matrix, again denoted as $f'(x)$. And it must be a good linear approximation:

$$\lim_{h \to \mathbf{0}} \frac{\|f(x+h) - f(x) - f'(x)h\|}{\|h\|} = 0.$$

As in Theorem 23.1, at interior points we have a relation with gradients: write the m coordinates of $f(x) \in \mathbb{R}^m$ as m separate coordinate functions,

$$f(x) = \begin{bmatrix} f_1(x) \\ \vdots \\ f_m(x) \end{bmatrix}.$$

Then

$$f'(x) = \begin{bmatrix} \nabla f_1(x) \\ \vdots \\ \nabla f_m(x) \end{bmatrix},$$

i.e., the derivative $f'(x)$ is the $m \times n$ matrix where each row $i = 1, \ldots, m$ is simply the gradient of the i-th coordinate function $f_i(x)$. This matrix of gradients is also called the ***Jacobian*** of f at x, denoted $J_f(x)$.

Example 25.1 The function $f : \mathbb{R}^3 \to \mathbb{R}^2$ with $f(x_1, x_2, x_3) = (x_1^2 + x_2 - 4x_3, 3x_1x_2)$ has

- ⊠ first coordinate function $f_1(x_1, x_2, x_3) = x_1^2 + x_2 - 4x_3$ with gradient $\nabla f_1(x) = (2x_1, 1, -4)$,
- ⊠ second coordinate function $f_2(x_1, x_2, x_3) = 3x_1x_2$ with gradient $\nabla f_2(x) = (3x_2, 3x_1, 0)$,

so its Jacobian is the 2×3 matrix with these gradients as its rows:

$$J_f(x) = \begin{bmatrix} 2x_1 & 1 & -4 \\ 3x_2 & 3x_1 & 0 \end{bmatrix}.$$

◁

25.2 Local mapping theorems

The main thing to remember about the three theorems below is this: at a point x_0 in its domain, a differentiable function f has a good linear approximation

$$\ell(x) = f(x_0) + f'(x_0)(x - x_0).$$

As long as x is close to x_0, $\ell(x)$ is close to $f(x)$, often denoted at $f(x) \approx \ell(x)$. An educated guess is that near x_0, nice behavior of f is related to nice behavior of its linear approximation ℓ. The three

local mapping theorems below confirm this: under suitable differentiability assumptions, *if the linear approximation ℓ is injective/surjective/bijective, then so is f, provided you stay close to x_0.*

We called a function $f : X \to Y$ injective if different arguments give different function values: for each $y \in Y$ there is at most one $x \in X$ with $f(x) = y$. Here is a local variant:

Definition 25.1 Function $f : X \to Y$ is a ***local injection at*** $x_0 \in X$ if there are neighborhoods A of x_0 and B of $f(x_0) = y_0$ such that for each $y \in B$ there is at most one $x \in A$ with $f(x) = y$.

To talk about neighborhoods, we implicitly assume X and Y to be metric/topological spaces.

Example 25.2 Function $f : \mathbb{R} \to \mathbb{R}$ with $f(x) = x^2$ is a local injection at each point $x_0 \neq 0$. For instance, for $x_0 > 0$, the function is strictly increasing — and consequently injective — on the neighborhood $A = (0, \infty)$ of x_0 (no matter what B you choose). But it is not locally injective at x_0: each neighborhood A of $x_0 = 0$ contains points ε and $-\varepsilon$ slightly above and slightly below zero with the same function value $f(\varepsilon) = f(-\varepsilon) = \varepsilon^2$, so f fails to be injective on such neighborhoods. ◁

The following theorem says that under suitable differentiability assumptions, if the linear approximation is injective in a point, then the function f is a local injection.

Theorem 25.1 (Local injection)

If $f : \mathbb{R}^n \to \mathbb{R}^m$ is continuously differentiable at x_0 and the $m \times n$ matrix $f'(x_0)$ is injective, then f is a local injection at x_0.

(Remember: 'continuously differentiable at x_0' means that at x_0 the derivative exists and is continuous; 'the $m \times n$ matrix $f'(x_0)$ is injective' means that the $m \times n$ matrix $f'(x_0)$ gives rise to an injective linear function $h \mapsto f'(x_0)h$, i.e., the columns of the matrix are linearly independent.)

We called a function $f : X \to Y$ surjective if for each $y \in Y$ there is at least one $x \in X$ with $f(x) = y$: each $y \in Y$ is attained as a function value. Our local variant says that whenever you are near x_0, function f attains all function values near $y_0 = f(x_0)$.

Definition 25.2 A function $f : X \to Y$ is a ***local surjection*** at $x_0 \in X$ if for each neighborhood A of x_0, its image $f(A) = \{f(a) : a \in A\}$ is a neighborhood of $y_0 = f(x_0)$.

Example 25.3 The quadratic function from our previous example is a local surjection at each point $x_0 \neq 0$ (why?), but not at $x_0 = 0$, because each neighborhood of $f(0) = 0$ contains negative numbers, which are not in the range of the quadratic function. ◁

Theorem 25.2 (Local surjection)

If $f : \mathbb{R}^n \to \mathbb{R}^m$ is continuously differentiable at x_0 and the $m \times n$ matrix $f'(x_0)$ is surjective, then f is a local surjection at x_0.

(Remember: 'the $m \times n$ matrix $f'(x_0)$ is surjective' means that the $m \times n$ matrix $f'(x_0)$ gives rise to a surjective linear function $h \mapsto f'(x_0)h$, i.e., the columns of the matrix span $\mathbb{R}^m$.) Combining the two theorems, we get:

Theorem 25.3 (Inverse function theorem: local bijection)

If $f : \mathbb{R}^n \to \mathbb{R}^m$ is continuously differentiable at x_0 and the $m \times n$ matrix $f'(x_0)$ is bijective, then f is both a local injection and a local surjection at x_0: near $(x_0, f(x_0))$, function f is invertible.

This result is useful when we solve systems of equations of the form $f(x) = y$: if we find a solution

$f(x_0) = y_0$ and the conditions of the inverse-function theorem hold, then small changes in y_0 imply that there is still a solution near x_0. In common applications, vector y is a bunch of parameters and since these are rarely known precisely, such a result leads to the helpful conclusion that solutions don't change too radically. Our next result is about solving more general systems of equations.

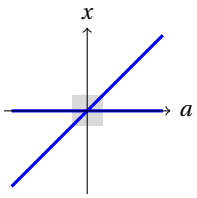
25.3 The implicit function theorem

The implicit function theorem is about solving a system of equations of the form $f(x,a) = \mathbf{0}$. Here, $f : \mathbb{R}^n \times \mathbb{R}^k \to \mathbb{R}^n$ is a function of some 'variables' $x \in \mathbb{R}^n$ and some 'parameters' $a \in \mathbb{R}^k$.

If we have a solution (x_0, a_0) to these equations, our aim is to vary the parameters and search, for a near a_0, for a new solution x to $f(x,a) = \mathbf{0}$ that is

1. unique, so that we can write this x as a well-defined function of a and
2. differentiable as a function of a.

Is that possible? Not always. Look at the function $f : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ with $f(x,a) = x(x-a)$ and notice that $f(x,a) = 0$ if and only if $x = 0$ or $x = a$ (see figure). If we ask if $f(x,a) = 0$ at least locally gives a unique solution, we see that it works near solutions (x,a) with $a \neq 0$, but not if $a = 0$: no matter what neighborhood (shaded) of $(x,a) = (0,0)$ you choose, it will always contain points on the horizontal axis and on the diagonal line, so it contains *multiple* (two) solutions for each a near zero, not a unique one. And to have any chance of differentiability, we'll need to impose differentiability assumptions on f to begin with.



Theorem 25.4 (Implicit function theorem)

Assume:

- ☒ $f : \mathbb{R}^n \times \mathbb{R}^k \to \mathbb{R}^n$ is continuously differentiable;
- ☒ $f(x_0, a_0) = \mathbf{0}$;
- ☒ the derivative of f with respect to x (keeping a_0 fixed) is bijective at x_0 (an invertible $n \times n$ matrix).

Then there are neighborhoods A of a_0 and B of x_0 with:

UNICITY: for each $a \in A$ there is a unique $x \in B$ with $f(x,a) = \mathbf{0}$.

Denoting this x by $h(a)$, we have 'implicitly' defined $x = h(a)$ as a function $h : A \to \mathbb{R}^n$ of a. Moreover:

DIFFERENTIABILITY: h is continuously differentiable.

Typical applications involve changes in a single parameter. Suppose that $n = k = 1$, f is continuously differentiable, $f(x_0, a_0) = 0$, and $\partial f(x_0, a_0)/\partial x \neq 0$. By the Implicit function theorem, there is a neighborhood $A \subseteq \mathbb{R}$ of a_0 and a continuously differentiable function $h : A \to \mathbb{R}$ such that $f(h(a), a) = 0$ for all $a \in A$. So for all a in an open set around a_0, the left-hand side $f(h(a), a)$ is constant and equal to zero. But then its derivative is zero as well. We can compute that derivative using the chain rule (notice that we find a in two places in the expression!):

$$0 = \frac{\mathrm{d}}{\mathrm{d}a} f(h(a), a) = \frac{\partial f(h(a), a)}{\partial x} \cdot h'(a) + \frac{\partial f(h(a), a)}{\partial a}.$$

Substituting $a = a_0$ and $h(a_0) = x_0$, we can solve for $h'(a_0)$ to obtain

$$h'(a_0) = -\frac{\partial f(x_0, a_0)/\partial a}{\partial f(x_0, a_0)/\partial x}.$$

Thus, the implicit function theorem tells us not only that there is an implicitly defined continuously differentiable function h that maps parameter a to the corresponding solution x of $f(x,a)=0$, but also allows us to compute the derivative of that function: it tells us both the sign and size of a small change in a on the corresponding solution x.

In the next subsection we consider a parametric optimization problem.

25.4 A variant of the envelope theorem

Let us look at a parametric variant of the standard nonlinear optimization problem in (79):

$$\text{maximize } f(x,a) \text{ with } h_1(x,a)\leq 0,\ldots,h_p(x,a)\leq 0, \tag{120}$$

where the only change is that we now allow both the goal function f and the constraints $h_1,\ldots,h_p$ to depend on some parameter $a\in\mathbb{R}$ in addition to the usual decision variables $x=(x_1,\ldots,x_n)$. For each parameter a, we define

$$m(a)=\max\{f(x,a) : h_1(x,a)\leq 0,\ldots,h_p(x,a)\leq 0\}$$

to be the maximal value of the goal function given parameter a. We make two assumptions throughout: (1) that the differentiability assumptions in the KKT theorem hold and, in line with the Implicit function theorem (2) that for each a' in a neighborhood of some fixed parameter value a there is a unique and differentiable solution $x^*(a')$ to the optimization problem. *How does the value $m(a)$ change when the parameter changes?* We distinguish two cases.

CASE 1: $x^*(a)$ is an interior point of the feasible set.

Hence, the partial derivatives $\partial f(x^*(a),a)/\partial x_i$ must be zero. With the Chain Rule, we find:

$$\begin{aligned} m'(a) &= \frac{\mathrm{d}}{\mathrm{d}a}f(x^*(a),a)=\frac{\mathrm{d}}{\mathrm{d}a}f(x_1^*(a),\ldots,x_n^*(a),a)\\ &= \underbrace{\frac{\partial}{\partial x_1}f(x^*(a),a)}_{=0}\cdot\frac{\mathrm{d}}{\mathrm{d}a}x_1^*(a)+\cdots+\underbrace{\frac{\partial}{\partial x_n}f(x^*(a),a)}_{=0}\cdot\frac{\mathrm{d}}{\mathrm{d}a}x_n^*(a)+\frac{\partial}{\partial a}f(x^*(a),a)\\ &= \frac{\partial}{\partial a}f(x^*(a),a). \end{aligned} \tag{121}$$

This result, that $m'(a)=\frac{\partial}{\partial a}f(x^*(a),a)$, is often referred to as the *envelope theorem* for this case. It says that if we change a, this affects the optimal value of the goal function

- ☒ *directly*, because a is the final argument of the goal function $f(x,a)$, but also
- ☒ *indirectly*, because a changes the optimal value of the decision variables $x_1^*(a),\ldots,x_n^*(a)$ and these in their turn affect the goal function,

but only the direct effect matters: the influence a has on the goal function via the partial derivative with respect to the final coordinate a.

CASE 2: $x^*(a)$ satisfies the KKT conditions.

We still have equation (121), but the partial derivatives no longer cancel out. To simplify notation, let's (i) omit arguments of functions, (ii) write

$$y=\left(\frac{\mathrm{d}}{\mathrm{d}a}x_1^*(a),\ldots,\frac{\mathrm{d}}{\mathrm{d}a}x_n^*(a)\right),$$

and (iii) have $\nabla f,\nabla h_1,\ldots,\nabla h_p$ denote the gradient/partial derivatives w.r.t. $x_1,\ldots,x_n$. So (121) becomes

$$m'(a)=\nabla f\cdot y+\frac{\partial}{\partial a}f. \tag{122}$$

By the KKT conditions, there are nonnegative multipliers $\mu_1, \ldots, \mu_p$ so that the gradient condition $\nabla f = \sum_{j=1}^{p} \mu_j \nabla h_j$ holds. If constraint j is nonbinding in the optimum, then $\mu_j = 0$ by complementary slackness. Or it is binding: $h_j(x^*(a), a) = 0$. Differentiating this expression with respect to a it follows that

$$\nabla h_j \cdot y + \frac{\partial}{\partial a} h_j = 0 \qquad \Longleftrightarrow \qquad \nabla h_j \cdot y = -\frac{\partial}{\partial a} h_j.$$

Substitute all this into (122):

$$\begin{aligned} m'(a) &= \nabla f \cdot y + \frac{\partial}{\partial a} f \\ &= \sum_{j=1}^{p} \mu_j \nabla h_j \cdot y + \frac{\partial}{\partial a} f \\ &= -\sum_{j=1}^{p} \mu_j \frac{\partial}{\partial a} h_j + \frac{\partial}{\partial a} f. \end{aligned}$$

In its full glory, with function arguments reinstated, the envelope theorem in this second case becomes

$$m'(a) = -\sum_{j=1}^{p} \mu_j \frac{\partial}{\partial a} h_j(x^*(a), a) + \frac{\partial}{\partial a} f(x^*(a), a). \tag{123}$$

Again, only the *direct* effect of the parameter a, now on f and the constraints $h_1, \ldots, h_p$, matters.

In an important special case we can assign a clearer economic meaning to the multipliers. Suppose the parameter a appears only once in the optimization problem (120), in some (say, the first) constraint, which is of the form

$$h_1(x, a) = g(x) - a \leq 0.$$

This may mean that $g(x)$ is the amount of a certain resource that is required to obtain x and the number a is the resource's available amount. The partial derivative of h_1 with respect to a is -1. By assumption, neither the goal function f, nor the other constraints $h_2, \ldots, h_p$ vary with a, so their partial derivatives with respect to a are zero. Substitution into (123) gives

$$m'(a) = \mu_1.$$

So the multiplier $\mu_1 \geq 0$ gives us the rate of change (slope) of the goal function if we slightly relax the resource constraint a: it tells us what such a change is worth. The multiplier is therefore sometimes called the *shadow price* of the resource: the price in units of the goal function (like profit or utility) the agent is willing to pay for a small increase in said resource.

Exercises section 25

25.1 Show that:

(a) If function $f : X \to Y$ between two metric spaces is an injection, then it is also a local injection at each point in its domain.

(b) The function $f : \mathbb{R} \to \mathbb{R}$ with $f(0) = 3$ and $f(x) = x$ if $x \neq 0$ is a local injection at each point in its domain, but not an injection. HINT: Draw the graph of f. For $x = 0$, consider neighborhoods $A = (-1, 1)$ of x and $B = (2, 4)$ of $f(0) = 3$ and draw $A \times B$ as well.

25.2 Give an example of a function $f : \mathbb{R} \to \mathbb{R}$ that is:

(a) a surjection, but at some point in its domain not a local surjection;

(b) a local surjection at each point in its domain, but not a surjection.

26 Correspondences

26.1 Motivation and definition

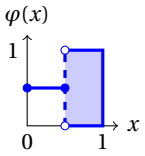
We often consider, given variables outside an economic agent's control, what this agent *can* do (e.g., budget set, strategies) or finds it *optimal* to do (e.g., demand, best responses). And those feasible or optimal choices may very well constitute a set of multiple elements. So it would be handy to have a generalization of functions (which map each point x in the domain to a single point $f(x)$ in its range) where points in the domain are mapped to *sets* instead of points. This generalization is a correspondence.[7]

Definition 26.1 A ***correspondence*** φ from a set X to a set Y assigns to each $x \in X$ a subset $\varphi(x)$ of Y. It is denoted by $\varphi : X \twoheadrightarrow Y$ (or $\varphi : X \rightrightarrows Y$) with multiple arrowheads to remind you that $\varphi(x)$ can consist of multiple points in Y. Its ***graph*** is the set

$$\operatorname{graph}(\varphi) = \{(x, y) \in X \times Y : y \in \varphi(x)\}.$$

Correspondences are also called 'multifunctions', 'multivalued functions', or 'point-to-set functions' and you can alternatively treat them as functions from X to the power set of Y. Correspondence $\varphi : X \twoheadrightarrow Y$ is called ***nonempty-valued*** if $\varphi(x)$ is a nonempty set for each $x \in X$. Other properties (convex-valued, closed-valued, compact-valued, ...) are defined likewise. Some authors define correspondences to be nonempty-valued. We won't need that here.

Example 26.1 The graph of the correspondence $\varphi : [0,1] \twoheadrightarrow [0,1]$ with $\varphi(x) = \left\{\frac{1}{2}\right\}$ if $x \in \left[0, \frac{1}{2}\right]$ and $\varphi(x) = [0,1]$ otherwise is drawn below.



◁

Example 26.2 (Matching pennies) Two players (player 1 choosing rows, player 2 choosing columns) simultaneously and independently choose heads (H) or tails (T). If their choices are the same, player 1 wins (utility 1) and player 2 loses (utility -1); if their choices are different, it's the other way around. This is summarized in the table below, where the numbers separated by a comma are the utilities of the first and second player, respectively.

		q	$1-q$
		H	T
p	H	$1,-1$	$-1,1$
$1-p$	T	$-1,1$	$1,-1$

Players aim to maximize their expected utility. If player 1 chooses H with probability $p \in [0,1]$ and consequently T with probability $1-p$ and player 2 does so with probability $q \in [0,1]$, then player 1's expected utility can be written as

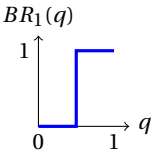
$$2p(2q-1) - 2q + 1.$$

[7] Most results in this section are stated without proof due to their length or complexity. If you want to read more, *Equilibrium Analysis* by Hildenbrand and Kirman (easier) and *Infinite Dimensional Analysis* by Aliprantis and Border (harder) go in substantial detail.

From this, we see that if $2q-1<0$, it is optimal to choose $p=0$; if $2q-1=0$, each $p\in[0,1]$ is optimal; and if $2q-1>0$, it is optimal to choose $p=1$. So the first player's best-response correspondence $BR_1:[0,1]\twoheadrightarrow[0,1]$ summarizing the set of optimal strategies in response to each strategy q of the opponent, is

$$BR_1(q)=\begin{cases}\{0\} & \text{if } q\in\left[0,\frac{1}{2}\right),\\ [0,1] & \text{if } q=\frac{1}{2},\\ \{1\} & \text{if } q\in\left(\frac{1}{2},1\right].\end{cases}$$

Its graph is drawn below.



◁

26.2 Continuity properties of correspondences

For a correspondence $\varphi:X\twoheadrightarrow Y$ between two metric spaces, we will define two continuity properties. Intuitively, upper-hemicontinuity at a point $x\in X$ says that φ does not 'explode' (become much larger) near x: if you allow some room $B\supseteq\varphi(x)$ for variation, then points x' near x have images $\varphi(x')$ that stay inside B. Lower-hemicontinuity at x says that φ does not 'implode' (become much smaller) near x: each point in $\varphi(x)$ can be approximated by points in $\varphi(x')$ whenever x' is near x.

Definition 26.2 Fix a correspondence $\varphi:X\twoheadrightarrow Y$ between two metric spaces.

- ☒ φ is ***upper-hemicontinuous (uhc) at*** $x\in X$ if for each open set B with $\varphi(x)\subseteq B$ there is a neighborhood A of x with $\varphi(x')\subseteq B$ for all $x'\in A$.
- ☒ φ is ***lower-hemicontinuous (lhc) at*** $x\in X$ if for each open set B with $\varphi(x)\cap B\neq\emptyset$ there is a neighborhood A of x with $\varphi(x')\cap B\neq\emptyset$ for all $x'\in A$.
- ☒ φ is ***continuous at*** $x\in X$ if it is both uhc and lhc at x.
- ☒ φ is ***(upper/lower-hemi)continuous*** if it is (upper/lower-hemi)continuous at each point in its domain.

Example 26.3 Correspondence φ in Example 26.1 'explodes' at $x=\frac{1}{2}$: it is not uhc there. For instance, $B=\left(\frac{1}{4},\frac{3}{4}\right)$ satisfies $\varphi\left(\frac{1}{2}\right)=\left\{\frac{1}{2}\right\}\subseteq B$, but there is no neighborhood A of $x=\frac{1}{2}$ with $\varphi(x')\subseteq B$ for each $x'\in A$ because each $x'>\frac{1}{2}$ has image $\varphi(x')=[0,1]$ with points outside B.

Correspondence BR_1 in Example 26.2 'implodes' at $q=\frac{1}{2}$: it is not lhc there. For instance, $B=\left(\frac{1}{4},\frac{3}{4}\right)$ satisfies $BR_1\left(\frac{1}{2}\right)\cap B\neq\emptyset$, but there is no neighborhood A of $x=\frac{1}{2}$ with $BR_1(x')\cap B\neq\emptyset$ for each $x'\in A$ because each $x'\neq\frac{1}{2}$ has image $BR_1(x')=\{0\}$ or $BR_1(x')=\{1\}$, which has no points in common with B.◁

Under some additional assumptions you can see whether a correspondence is upper-hemicontinuous by simply inspecting its graph:

Theorem 26.1 (Upper-hemicontinuity and closed graphs)

Let $\varphi:X\twoheadrightarrow Y$ be a correspondence between two metric spaces. Assume that φ is closed-valued and Y is compact. Then φ is uhc if and only if its graph is a closed set.

In Exercise 26.1 we show that the equivalence breaks down without the closed-valuedness and compactness assumptions. According to this theorem, correspondence φ in Example 26.1 is not uhc: its graph is not closed because a point like $(x, y) = \left(\frac{1}{2}, \frac{1}{4}\right)$ is a boundary point of the graph, but not an element of the graph. The theorem does assure that correspondence BR_1 in Example 26.2 is uhc: its graph is closed.

This closed-graph property together with the sequential characterization of closed sets in Theorem 13.4 proves that under the assumptions in Theorem 26.1, the correspondence is uhc if and only if for each convergent sequence in the graph, also its limit belongs to the graph.

Analogously, lower-hemicontinuous correspondences can be characterized using convergent sequences. Remember the intuition: φ is lhc at x if each point in $\varphi(x)$ can be approximated by points in $\varphi(x')$ whenever x' is near x. Formally:

Theorem 26.2

Correspondence $\varphi : X \twoheadrightarrow Y$ between two metric spaces is lhc at $x \in X$ if and only if for each sequence $(x_k)_{k\in\mathbb{N}}$ in X converging to x and each $y \in \varphi(x)$, there is a sequence $(y_k)_{k\in\mathbb{N}}$ with $y_k \in \varphi(x_k)$ for each $k \in \mathbb{N}$ that converges to y.

26.3 Berge's maximum theorem

In economics, upper-hemicontinuity pops up more often than lower-hemicontinuity. This is because the optimal choices of an economic agent, by force of Berge's maximum theorem below, often depend in an upper-hemicontinuous way on the model's parameters. In matching pennies, finding the row player's best-response correspondence is an example of such a parametric optimization problem where you strive to optimize a goal function over a set of feasible options, both of which may depend on parameters that are outside your control: the player wants to maximize the expected payoff, but cannot affect what the opponent does. Analogously, consumers are often modeled as maximizing their utility function over the budget set of commodity bundles they can afford given commodity prices and wealth.

To formalize a parametric optimization problem, we denote the parameter(s) by elements x in a set X, the choices by elements y in a set Y, have a goal function $f : X \times Y \to \mathbb{R}$, and specify the feasible choices as a correspondence $\varphi : X \twoheadrightarrow Y$ of the parameters. Given parameter $x \in X$, the aim is to find the maximal value

$$m(x) = \max_{y\in\varphi(x)} f(x, y)$$

of the goal function over the set of feasible choices $\varphi(x)$ given the parameter. Likewise, we want to know the set of optimal choices,

$$\mu(x) = \arg\max_{y\in\varphi(x)} f(x, y) = \{y^* \in \varphi(x) : f(x, y^*) \geq f(x, y) \text{ for all } y \in \varphi(x)\}.$$

Berge's maximum theorem is about the behavior of the maximal value m and the maximizers μ as the parameter changes.

Theorem 26.3 (Berge's maximum theorem)

Assume that

- ☒ the goal function $f : X \times Y \to \mathbb{R}$ is continuous;
- ☒ the constraint correspondence $\varphi : X \twoheadrightarrow Y$ is nonempty-valued, compact-valued and continuous.

Then

- ☒ the optimal value function $m : X \to \mathbb{R}$ is continuous and
- ☒ the correspondence $\mu : X \twoheadrightarrow Y$ of optimal choices is nonempty-valued, compact-valued, and upper-hemicontinuous.

The following example shows that both upper- and lower-hemicontinuity of the constraint correspondence φ are required to make the correspondence μ of maximizers upper-hemicontinuous.

Example 26.4 Maximize y with $y \in \varphi(x)$ where $\varphi : \mathbb{R} \twoheadrightarrow \mathbb{R}$ has $\varphi(x) = \{0\}$ if $x < 0$ and $\varphi(x) = [0,1]$ otherwise. All assumptions in Theorem 26.3 are satisfied except for lower-hemicontinuity of φ at $x = 0$. The maximizers are $\mu(x) = \{0\}$ if $x < 0$ and $\mu(x) = \{1\}$ otherwise, which is not uhc at $x = 0$.

If you change $\varphi(0)$ to $\{0\}$, all assumptions in Theorem 26.3 are satisfied except for upper-hemicontinuity of φ at $x = 0$ and again the correspondence of maximizers turns out not to be uhc at $x = 0$. ◁

In 'microfounded' models an equilibrium is often defined as a scenario that is (1) *feasible* and where all agents simultaneously choose something (2) *optimal* given whatever is outside their control. Berge says something about optimal behavior of individuals. We now want to aggregate this to all agents. This typically involves looking at products (the best-response correspondence of all players simultaneously) or sums (aggregate demand of consumers, aggregate supply of producers) of correspondences. The next theorems tell when such aggregates remain upper-hemicontinuous.

Theorem 26.4 (Upper-hemicontinuity of products)

Let $\varphi_1 : X_1 \twoheadrightarrow Y_1, \ldots, \varphi_n : X_n \twoheadrightarrow Y_n$ be correspondences between metric spaces. If each φ_i is uhc and compact-valued, then so is the product correspondence $\varphi : \times_{i=1}^n X_i \twoheadrightarrow \times_{i=1}^n Y_i$ with

$$\varphi(x_1, \ldots, x_n) = \varphi_1(x_1) \times \cdots \times \varphi_n(x_n) \qquad \text{for each } (x_1, \ldots, x_n) \in X_1 \times \cdots \times X_n.$$

Theorem 26.5 (Upper-hemicontinuity of sums)

Let $\varphi_1 : X \twoheadrightarrow \mathbb{R}^m, \ldots, \varphi_n : X \twoheadrightarrow \mathbb{R}^m$ be correspondences from a metric space X to $\mathbb{R}^m$. If each φ_i is uhc and compact-valued, then so is the sum correspondence $\varphi : X \twoheadrightarrow \mathbb{R}^m$ defined for each $x \in X$ by

$$\varphi(x) = \varphi_1(x) + \cdots + \varphi_n(x) = \left\{ y_1 + \cdots + y_n : y_1 \in \varphi_1(x), \ldots, y_n \in \varphi_n(x) \right\}.$$

26.4 The fixed-point theorems of Brouwer and Kakutani

In Example 16.8 we used the Intermediate Value Theorem to show that each continuous function $f : [0,1] \to [0,1]$ has a fixed point: $f(x^*) = x^*$ for some x^* in its domain. Brouwer's fixed-point theorem is a substantial generalization.

Theorem 26.6 (Brouwer's fixed-point theorem)

A continuous function $f : X \to X$ on a nonempty, convex, compact domain $X \subseteq \mathbb{R}^n$ has a fixed point: $f(x) = x$ for some $x \in X$.

There is a similar result for correspondences:

Theorem 26.7 (Kakutani's fixed-point theorem)

Assume that

- ☒ $X \subseteq \mathbb{R}^n$ is nonempty, convex, compact;
- ☒ $\varphi : X \twoheadrightarrow X$ is nonempty-, convex-, compact-valued and upper-hemicontinuous.

Then φ has at least one fixed point: $x \in \varphi(x)$ for some $x \in X$.

To summarize, classical existence results for economic equilibria often combine the results above:

- ☒ Berge assures that individual agents' behavior changes 'nicely' with the model's parameters;
- ☒ Theorems 26.4 and 26.5 help to show the same at the aggregate level;
- ☒ A fixed-point theorem shows that among their aggregate behavior, there is a fixed point, something that no agent wants to deviate from given whatever parameters happen to be outside their control.

26.5 Correspondences defined by inequalities

In applications, the values of a correspondence are often defined using (in)equalities. Here we provide some results on the hemicontinuity of such correspondences. Formally, we consider correspondences $\varphi : X \twoheadrightarrow Y$ of the form

$$\varphi(x) = \{y \in Y : g(x, y) \leq \mathbf{0}\} \tag{124}$$

for some function $g : X \times Y \to \mathbb{R}^\ell$. In other words, given $x \in X$, its image $\varphi(x)$ consists of those elements $y \in Y$ that satisfy some finite number ℓ of inequalities

$$g_1(x, y) \leq 0, \ldots, g_\ell(x, y) \leq 0.$$

(Analogously, we write $g(x, y) < \mathbf{0}$ when $g_i(x, y) < 0$ for each $i = 1, \ldots, \ell$.)

In (124) we can replace $\mathbf{0}$ with any other vector b in $\mathbb{R}^\ell$, because we can write $g(x, y) \leq b$ as $g(x, y) - b \leq \mathbf{0}$ and apply the results to the function $(x, y) \mapsto g(x, y) - b$. Also the direction of the inequality is irrelevant: $g(x, y) \geq \mathbf{0}$ is equivalent to $-g(x, y) \leq \mathbf{0}$, so we can apply our results to $-g$. It also covers equality restrictions: $g(x, y) = \mathbf{0}$ can be written as two inequalities, $g(x, y) \leq \mathbf{0}$ and $g(x, y) \geq \mathbf{0}$.

Example 26.5 (Budget sets) A consumer has an amount $w > 0$ (w for 'wealth') to spend on consumption of $\ell \in \mathbb{N}$ goods. Each good $i \in \{1, \ldots, \ell\}$ can be consumed in a nonnegative amount x_i and one unit costs $p_i > 0$. Given prices $p = (p_1, \ldots, p_\ell)$ and wealth w, the consumer can afford all consumption bundles $x = (x_1, \ldots, x_\ell)$ in the 'budget set'

$$\varphi(p, w) = \{x \in \mathbb{R}^\ell_+ : \underbrace{p_1 x_1 + \cdots + p_\ell x_\ell}_{= p^\top x} \leq w\} = \{x \in \mathbb{R}^\ell_+ : p^\top x - w \leq 0\}.$$

This budget correspondence, mapping each price-wealth combination $(p, w) \in \mathbb{R}^{\ell+1}_{++}$ (note: ℓ goods, one wealth, so (p, w) has $\ell + 1$ coordinates) to the set $\varphi(p, w)$ of affordable consumption bundles $x \in \mathbb{R}^\ell_+$ is of the desired form (124) with $X = \mathbb{R}^{\ell+1}_{++}$, $Y = \mathbb{R}^\ell_+$, and $g(p, w, x) = p^\top x - w$. ◁

To state the theorem about continuity properties of correspondences defined by inequalities, we need one more definition. A correspondence is locally bounded at a point x in its domain if all points near x have images inside the same compact set:

Definition 26.3 Correspondence $\varphi : X \twoheadrightarrow Y$ is ***locally bounded*** at a point $x \in X$ if there is a compact set K in Y such that $\varphi(x') \subseteq K$ for all x' in a neighborhood of x. The correspondence is ***locally bounded*** if it is locally bounded at each point in its domain.

This is often easy to check. It holds by default if Y itself is compact. And if Y is some Euclidean space $\mathbb{R}^n$, then local boundedness reduces to a simpler condition:

$$\text{there is a } \textit{bounded} \text{ set } W \text{ in } Y \text{ such that } \varphi(x') \in W \text{ for all } x' \text{ in a neighborhood of } x. \tag{125}$$

(Why? Well, by exercise 17.5 the closure of W satisfies definition 26.3.) This also explains where 'local *bounded*ness' got its name.

Theorem 26.8 (Continuity properties of correspondences defined by inequalities)

Let X and Y be metric spaces and $g : X \times Y \to \mathbb{R}^\ell$ a function. Define correspondence $\varphi : X \twoheadrightarrow Y$ for each $x \in X$ by

$$\varphi(x) = \{y \in Y : g(x,y) \leq \mathbf{0}\}.$$

(a) If g is continuous, then φ has a closed graph.

(b) If g is continuous and φ is locally bounded, then φ is upper-hemicontinuous.

(c) If

(i) for each (x,y) with $g(x,y) \leq \mathbf{0}$ there is a y' near y where the inequality is strict: formally, each neighborhood of y contains an element y' with $g(x,y') < \mathbf{0}$;

(ii) for each $y \in Y$, $g(\cdot, y)$ is a continuous function on X;

then φ is lower-hemicontinuous.

Proof: **(a)** If g is continuous, then φ's graph, the set

$$\text{graph}(\varphi) = \{(x,y) \in X \times Y : g(x,y) \leq \mathbf{0}\}$$

is the pre-image of the closed set $\{z \in \mathbb{R}^\ell : z \leq \mathbf{0}\}$, hence closed by Theorem 11.2.
(b) Let $x \in X$. We prove that φ is uhc at x. Let V be an open set with $\varphi(x) \subseteq V$. We must show that there is a neighborhood of x with $\varphi(x') \subseteq V$ for all x' in this neighborhood. Suppose that this is not the case. Then for each $k \in \mathbb{N}$, the neighborhood $B(x, 1/k)$ of x contains a point x_k with $\varphi(x_k) \nsubseteq V$. Pick any point $y_k \in \varphi(x_k)$ with $y_k \notin V$.

Since φ is locally bounded at x and the x_k's converge to x, the images $\varphi(x_k)$ and in particular their elements y_k belong to some compact set for large enough k. By Theorem 17.4, they have a convergent subsequence whose limit we call y. Each y_k lies in the closed set V^c, hence so does the limit y. In particular, $y \notin \varphi(x)$.

So the (x_k, y_k) along this subsequence are elements of the graph of φ, but their limit (x,y) is not. This contradicts that the graph of φ, by the previous part of the theorem, is a closed set.
(c) Let $x \in X$. To see that φ is lhc at x, let $B \subseteq Y$ be an open set with $\varphi(x) \cap B \neq \emptyset$. Pick any y in this intersection: $g(x,y) \leq \mathbf{0}$ and $y \in B$. By (i), neighborhood B of y contains an element y' with $g(x,y') < \mathbf{0}$. By (ii), $g(\cdot, y')$ is continuous, so the pre-image $A = \{x' \in X : g(x', y') < \mathbf{0}\}$ is open. It contains x and by definition of A, $y' \in \varphi(x') \cap B$ for all $x' \in A$, making φ lower-hemicontinuous at x. □

Example 26.6 (Budget sets continued) The budget correspondence φ from Example 26.5 has a closed graph, because the constraint function g with $g(p,w,x) = p^\top x - w$ is continuous.

It is also locally bounded. Intuitively, each budget set $\varphi(p,w)$ is itself bounded and it does not change much if you only make small changes in prices and wealth. You are asked to make this precise in Exercise 26.2. Therefore, the budget correspondence is upper-hemicontinuous.

Finally, part (c) of the theorem above assures that the budget correspondence is lower-hemicontinuous. The function g with $g(p,w,x) = p^\top x - w$ is continuous and for each (p,w,x) with $p^\top x - w \leq 0$ there is an x' arbitrarily near x with a strict inequality $p^\top x' - w < 0$: if x itself satisfies $p^\top x - w < 0$, just take $x' = x$; and if $p^\top x - w = 0$, take $x' = \alpha x$ where α is a real number in $(0,1)$. Commodity bundle x costs w, so x' only costs $\alpha w < w$, so $p^\top x' - w = \alpha w - w < 0$. Picking α sufficiently close to 1, there is such an x' in any neighborhood around x. ◁

Exercises section 26

26.1 For each correspondence below, answer the following: (1) Draw its graph. (2) Is this graph closed? (3) Is the correspondence uhc? (4) Which of the assumptions in Theorem 26.1 is violated?

(a) $\varphi : \mathbb{R} \twoheadrightarrow \mathbb{R}$ with $\varphi(0) = \{0\}$ and $\varphi(x) = \left\{\frac{1}{x}\right\}$ for all other x.

(b) $\varphi : \mathbb{R} \twoheadrightarrow [0,3]$ with $\varphi(x) = (1,2)$ for all x.

26.2 To see that the budget correspondence in Example 26.5 is locally bounded, we use condition (125). Let $(p,w) \in \mathbb{R}^{\ell+1}_{++}$ be a price-wealth combination. Show that

$$\left\{(p',w') \in \mathbb{R}^{\ell+1}_{++} : p'_1 > \tfrac{1}{2}p_1, \ldots, p'_\ell > \tfrac{1}{2}p_\ell, w' < 2w\right\}$$

is a neighborhood of (p,w) and that for all (p',w') in this neighborhood, $\varphi(p',w')$ is a subset of the bounded set

$$W = \left[0, \frac{4w}{p_1}\right] \times \cdots \times \left[0, \frac{4w}{p_\ell}\right].$$

26.3 Let $\varphi : X \twoheadrightarrow Y$ be a correspondence. Suppose $\varphi(x_0) = \emptyset$ for some $x_0 \in X$.

(a) Show: φ is lower hemicontinuous at x_0.

(b) Show: φ is upper hemicontinuous at x_0 if and only if $\varphi(x) = \emptyset$ for all x in a neighborhood of x_0.

26.4 Draw the graph of the correspondence $\varphi : \mathbb{R} \twoheadrightarrow \mathbb{R}$ with $\varphi(x) = \{y \in \mathbb{R} : x^2 + y^2 \leq 0\}$. Show that it is not lower hemicontinuous. Which condition in Theorem 26.8(c) is violated? Is it upper hemicontinuous?

26.5 For each $x \in \mathbb{R}$, let $\lfloor x \rfloor = \max\{z \in \mathbb{Z} : z \leq x\}$ be the largest integer less than or equal to x; it is obtained by rounding off x downwards to the nearest integer. For instance, $\lfloor \pi \rfloor = \lfloor 3.99 \rfloor = \lfloor 3 \rfloor = 3$. Draw the graph of the correspondence $\varphi : [0,3] \twoheadrightarrow \mathbb{R}$ with $\varphi(x) = \{y \in \mathbb{R} : y \leq \lfloor x \rfloor\}$. Is this correspondence (a) locally bounded? (b) upper hemicontinuous? (c) lower hemicontinuous?

27 More on orthogonality and projections

Orthogonal vectors are really easy to work with. Let set of vectors W be such that $\mathbf{0} \notin W$ and each pair of distinct vectors $v, w \in W$ is orthogonal. Firstly, W is linearly independent. Suppose some linear combination $\alpha_1 w_1 + \cdots + \alpha_n w_n$ of distinct vectors in W equals the zero vector. Then all α_i must be zero, because taking its inner product with an arbitrary w_i gives

$$0 = \langle \mathbf{0}, w_i \rangle = \langle \alpha_1 w_1 + \cdots + \alpha_n w_n, w_i \rangle = \alpha_i \langle w_i, w_i \rangle + \sum_{j \neq i} \alpha_j \underbrace{\langle w_j, w_i \rangle}_{=0},$$

so division by $\langle w_i, w_i \rangle \neq 0$ gives $\alpha_i = 0$. Secondly, if v is a linear combination of vectors in W,

$$v = \beta_1 w_1 + \cdots + \beta_m w_m,$$

the scalars β_i are easily found: use the same trick of taking the inner product with w_i and rearrange terms to find

$$\beta_i = \frac{\langle v, w_i \rangle}{\langle w_i, w_i \rangle} = \frac{\langle v, w_i \rangle}{\|w_i\|^2} \qquad \text{for each } i = 1, \ldots, m.$$

The set W is ***orthonormal*** if distinct elements are orthogonal and each has length one. In that case, the coefficients are even simpler to compute: by assumption each w_i has length $\|w_i\| = 1$, so $\beta_i = \langle v, w_i \rangle$.

Gram-Schmidt orthogonalization is a procedure to transform linearly independent vectors into orthogonal/-normal ones spanning the same space:

Theorem 27.1 (Gram-Schmidt orthogonalization)

Let $v_1, v_2, \ldots$ be linearly independent vectors in an inner product space V. Define $w_1 = v_1 / \|v_1\|$ and recursively, for larger $k \in \mathbb{N}$,

$$w_k = \frac{v_k - \langle v_k, w_1 \rangle w_1 - \cdots - \langle v_k, w_{k-1} \rangle w_{k-1}}{\|v_k - \langle v_k, w_1 \rangle w_1 - \cdots - \langle v_k, w_{k-1} \rangle w_{k-1}\|}.$$

Then $w_1, w_2, \ldots$ are orthonormal and for each $k \in \mathbb{N}$ the sets

$$\{v_1, \ldots, v_k\} \qquad \text{and} \qquad \{w_1, \ldots, w_k\}$$

span the same subspace of V.

Proof: We prove by induction on k that each w_k is well-defined, set $\{w_1, \ldots, w_k\}$ is orthonormal, and $\text{span}\{v_1, \ldots, v_k\} = \text{span}\{w_1, \ldots, w_k\}$. For $k = 1$, this is trivial: $v_1 \neq \mathbf{0}$ by linear independence and w_1 simply rescales v_1 to length one. Next, let k be any larger integer and assume the claim is true for $k-1$. Then:

- ☒ w_k is well-defined. By the induction hypothesis, $\text{span}\{v_1, \ldots, v_{k-1}\} = \text{span}\{w_1, \ldots, w_{k-1}\}$. By linear independence of the v_i's, v_k does not lie in this span. Hence,

  $$z_k := v_k - \langle v_k, w_1 \rangle w_1 - \cdots - \langle v_k, w_{k-1} \rangle w_{k-1}$$

  cannot equal $\mathbf{0}$, making $w_k = z_k / \|z_k\|$ a well-defined vector of length one.
- ☒ $\{w_1, \ldots, w_k\}$ is orthonormal. By the induction hypothesis, $\{w_1, \ldots, w_{k-1}\}$ is orthonormal, so we only need to show that $\langle w_k, w_\ell \rangle = 0$ for all $\ell < k$. Take such an ℓ. Orthonormality of $\{w_1, \ldots, w_{k-1}\}$ gives that $\langle w_i, w_\ell \rangle = 0$ for all $i = 1, \ldots, k-1$ distinct from ℓ and $\langle w_\ell, w_\ell \rangle = 1$. So

  $$\langle z_k, w_\ell \rangle = \langle v_k, w_\ell \rangle - \sum_{i=1}^{k-1} \langle v_k, w_i \rangle \langle w_i, w_\ell \rangle = \langle v_k, w_\ell \rangle - \langle v_k, w_\ell \rangle \langle w_\ell, w_\ell \rangle = 0.$$

Hence, also $\langle w_k, w_\ell \rangle = \frac{1}{\|z_k\|}\langle z_k, w_\ell \rangle = 0$.

☒ $\operatorname{span}\{v_1,\ldots,v_k\} = \operatorname{span}\{w_1,\ldots,w_k\}$. By the induction hypothesis, $\{v_1,\ldots,v_{k-1}\}$ and $\{w_1,\ldots,w_{k-1}\}$ have the same span. So z_k and consequently w_k can be written as a linear combination of $\{v_1,\ldots,v_k\}$, showing that $\operatorname{span}\{v_1,\ldots,v_{k-1}\} \supseteq \operatorname{span}\{w_1,\ldots,w_{k-1}\}$. A similar argument for

$$v_k = z_k + \sum_{i=1}^{k-1} \langle v_k, w_i \rangle w_i = \|z_k\| w_k + \sum_{i=1}^{k-1} \langle v_k, w_i \rangle w_i$$

gives the opposite inclusion. □

Definition 27.1 An inner product space is a ***Hilbert space*** if it is complete under the induced metric

$$d(x,y) = \|x-y\| = \sqrt{\langle x-y, x-y\rangle},$$

i.e, if every Cauchy sequence converges.

The prime example of a Hilbert space, of course, is $\mathbb{R}^n$ with its Euclidean distance. But so are:

☒ the space ℓ_2 of real sequences $x = (x_1, x_2, x_3, \ldots)$ for which $\sum_{i=1}^{\infty} x_i^2$ exists, with inner product

$$\langle x, y \rangle = \sum_{i=1}^{\infty} x_i y_i.$$

☒ the space $L_2[a,b]$ of square-integrable functions $f : [a,b] \to \mathbb{R}$, i.e., those for which

$$\int_a^b f(x)^2 \,\mathrm{d}x$$

exists, with inner product

$$\langle f, g \rangle = \int_a^b f(x) g(x) \,\mathrm{d}x.$$

The proofs, however, are outside the scope of this course (the former isn't too difficult, the latter cannot be done without a good formal knowledge of integrals). The first example is common in time-series analysis, the second is common in data compression, for instance for music files, where songs correspond with sound waves.

Theorem 27.2 (Nearest point theorem)

Let C be a nonempty, closed, convex subset of a real Hilbert space H. For each $x \in H$ there is a unique point $c^* \in C$ with minimal distance to x. It satisfies

$$\text{for all } c \in C: \qquad \langle x - c^*, c - c^* \rangle \le 0. \tag{126}$$

If, in addition, C is a subspace of vector space H, then $x - c^*$ is orthogonal to each vector in C.

Proof: EXISTENCE OF c^*: The set of distances $\{\|c - x\| : c \in C\}$ to points in C is nonempty and bounded from below by zero, so it has an infimum δ. Take a sequence $c_1, c_2, \ldots$ in C with distances $\|c_n - x\|$ converging to δ. Sequence $c_1, c_2, \ldots$ is Cauchy: for all m and n, convexity of C gives $(c_m + c_n)/2 \in C$, so

$$\|(c_m - x) + (c_n - x)\|^2 = \left\| 2\left(\frac{c_m + c_n}{2} - x\right)\right\|^2 = 4\left\|\frac{c_m + c_n}{2} - x\right\|^2 \ge 4\delta^2.$$

The parallelogram law and the previous expression give

$$\begin{aligned}\|c_m - c_n\|^2 = \|(c_m - x) - (c_n - x)\|^2 &= 2\|c_m - x\|^2 + 2\|c_n - x\|^2 - \|(c_m - x) + (c_n - x)\|^2 \\ &\leq 2\|c_m - x\|^2 + 2\|c_n - x\|^2 - 4\delta^2.\end{aligned}$$

Since $\|c_k - x\| \to \delta$ as $k \to \infty$, the final expression can be made arbitrarily close to zero by choosing m and n sufficiently large. By completeness of H, this Cauchy sequence has a limit $c^* \in H$. Since C is closed and the sequence lies in C, so does its limit c^*. And since $c_n \to c^*$, continuity of the norm gives that $\|c^* - x\| = \lim_n \|c_n - x\| = \delta$: c^* is a point in C closest to x.
UNIQUE c^*: If also $c \in C$ satisfies $\|c - x\| = \delta$, our parallelogram argument once more gives

$$0 \leq \|c^* - c\|^2 = \|(c^* - x) - (c - x)\|^2 = 2\|c^* - x\|^2 + 2\|c - x\|^2 - 4\left\|\frac{c^* + c}{2} - x\right\|^2 \leq 2\delta^2 + 2\delta^2 - 4\delta^2 = 0,$$

so $\|c^* - c\| = 0$, which implies that $c^* = c$.
PROPERTY (126): Let $c \in C$ and $\lambda \in (0,1)$. By convexity of C, $\lambda c + (1-\lambda)c^* \in C$. So by definition of c^*:

$$\begin{aligned}\|x - c^*\|^2 &\leq \left\|x - \left(\lambda c + (1-\lambda)c^*\right)\right\|^2 \\ &= \|(x - c^*) - \lambda(c - c^*)\|^2 \\ &= \|x - c^*\|^2 - 2\lambda\langle x - c^*, c - c^*\rangle + \lambda^2\|c - c^*\|^2.\end{aligned}$$

Simplifying this expression and dividing by λ gives

$$2\langle x - c^*, c - c^*\rangle \leq \lambda\|c - c^*\|^2.$$

Since $\lambda \in (0,1)$ was arbitrary, letting it go down to zero gives (126).
ORTHOGONALITY: Take any vector $d \in C$. If C is a subspace, also $d + c^*$ and $-d + c^*$ lie in C. Substituting those for c in (126) give $\langle x - c^*, d\rangle \leq 0$ and $\langle x - c^*, -d\rangle = -\langle x - c^*, d\rangle \leq 0$, so $\langle x - c^*, d\rangle = 0$. □

Definition 27.2 Let W be a nonempty subset of an inner product space V. Its ***orthogonal complement***

$$W^\perp = \{v \in V : \langle v, w\rangle = 0 \text{ for all } w \in W\}$$

is the set of vectors orthogonal to each element of W.

Example 27.1 (Orthogonality of kernel and row space) If A is an $m \times n$ real matrix, the function $T : \mathbb{R}^n \to \mathbb{R}^m$ with $T(x) = Ax$ is linear and it is a common, minor abuse of notation to write $\ker(A)$ for $\ker(T)$, the kernel/null space of this transformation. The ***row space*** of A is the subspace of $\mathbb{R}^n$ spanned by the rows of A. Denote the rows of A by $r_1, \ldots, r_m$. Then

$$\begin{aligned}x \in \ker(A) &\iff Ax = \mathbf{0} \\ &\iff \langle r_i, x\rangle = 0 \text{ for each row } i = 1, \ldots, m \\ &\iff x \text{ is orthogonal to } \operatorname{span}\{r_1, \ldots, r_m\}, \text{ the row space of } A.\end{aligned}$$

So the kernel of A is the orthogonal complement of the row space of A: $\ker(A) = \left(\operatorname{span}\{r_1, \ldots, r_m\}\right)^\perp$. ◁

Theorem 27.3 (Properties of orthogonal complements)

Let W be a nonempty subset of an inner product space V.

(a) $W^\perp$ is a subspace of V.

(b) $W^\perp$ is a closed set.

In the remainder of the theorem, assume W is not just a subset, but a subspace of V.

(c) $W \cap W^\perp = \{\mathbf{0}\}$.

(d) If V is finite-dimensional, then $V = W + W^\perp$ and each element of V can be written in exactly one way as the sum of a vector in W and a vector in $W^\perp$.

Proof: You are asked to prove (a), (b), and (c) in Exercise 27.1. For (d), use Gram-Schmidt to select an orthogonal basis $\{w_1, \ldots, w_k\}$ of W. Extend it to an orthogonal basis $\{w_1, \ldots, w_k, v_1, \ldots, v_m\}$ of V. By definition, $v_i \in W^\perp$ for all $i = 1, \ldots, m$. Given this basis, for each $v \in V$ there are scalars α_i and β_i such that

$$v = \underbrace{\alpha_1 w_1 + \cdots + \alpha_k w_k}_{\in W} + \underbrace{\beta_1 v_1 + \cdots + \beta_n v_n}_{\in W^\perp},$$

showing that it is the sum of a vector in W and a vector in $W^\perp$, i.e., $V = W + W^\perp$.

To see that the decomposition of v into elements of W and $W^\perp$ is unique, suppose that $v = u + u' = w + w'$ for some $u, w \in W$ and $u', w' \in W^\perp$. Then $(u - w) + (u' - w') = \mathbf{0}$, so $u - w = w' - u'$. But the left difference lies in W and the right lies in $W^\perp$. By (c), they only have the zero vector in common: $u - w = w' - u' = \mathbf{0}$. So $u = w$ and $u' = w'$: the decomposition is unique. □

Part (d) does not extend to infinite-dimensional inner product spaces:

Example 27.2 Consider $C[0,1]$ with inner product $\langle f, g\rangle = \int_0^1 f(x)g(x)\,\mathrm{d}x$. The set $W = \{f : f(0) = 0\}$ is a subspace. Its orthogonal complement $W^\perp$ is $\{\mathbf{0}\}$. It obviously contains the zero function. Suppose $W^\perp$ also contains some function g. Then h with $h(x) = xg(x)$ satisfies $h(0) = 0$, so h lies in W. Consequently, g and h are orthogonal: $\langle g, h\rangle = \int_0^1 xg(x)^2\,\mathrm{d}x = 0$. Since $x \mapsto xg(x)^2$ is nonnegative on $[0,1]$ and the area between it and the horizontal axis must be zero, we must have $xg(x)^2 = 0$ for all $x \in [0,1]$. This implies that $g(x) = 0$ for all $x \in (0,1]$. By continuity, also $g(0) = 0$. So $g = \mathbf{0}$. ◁

Exercises section 27

27.1 Prove parts (a) to (c) of Theorem 27.3.

28 The determinant

28.1 Axiomatic definition of the determinant

The determinant of a square matrix is a convenient tool in lots of applications (Is a matrix invertible? What are its eigenvalues? ...); we will introduce it by telling exactly what properties it satisfies. Throughout this section, we only talk about square matrices of real or complex numbers.

Theorem 28.1 (Axioms of the determinant)

There is exactly one function, called the ***determinant*** and denoted det, that assigns a scalar $\det(A)$ to each $n \times n$ matrix A and satisfies:

(DET1) The determinant is a linear function of each of its columns, when you keep all other columns fixed.

(DET2) If a matrix has two identical columns, its determinant is zero.

(DET3) The determinant of the identity matrix is one.

Instead of $\det(A)$, also $\det A$ and $|A|$ are common. It is important to keep in mind that (DET1) is a linearity property for each column *separately*. It says, for each column j, that if the j-th column of matrix A can be written as a linear combination $\alpha x + \beta y$ of two vectors x and y, then

$$\det(\ldots, \alpha x + \beta y, \ldots) = \alpha \det(\ldots, x, \ldots) + \beta \det(\ldots, y, \ldots).$$

The dots here hide the columns before and after the j-th and we keep them fixed. The determinant is sometimes called ***multilinear*** because of (DET1) and ***alternating*** because (DET1) and (DET2) imply that the determinant changes sign whenever two columns are interchanged:

(DET4) If matrix B is obtained from matrix A by exchanging two columns, then $\det(B) = -\det(A)$.

Proof: Denote A's columns by $a^1, \ldots, a^n$. Suppose we switch a^i and a^j to obtain B. Then

$$\begin{aligned}
\det(A) + \det(B) &= \det(\cdots, a^i, \cdots, a^j, \cdots) + \det(\cdots, a^j, \cdots, a^i, \cdots) \\
&= \det(\cdots, a^i, \cdots, a^j, \cdots) + \det(\cdots, a^j, \cdots, a^i, \cdots) \\
&+ \underbrace{\det(\cdots, a^i, \cdots, a^i, \cdots)}_{=0 \text{ by (DET2)}} + \underbrace{\det(\cdots, a^j, \cdots, a^j, \cdots)}_{=0 \text{ by (DET2)}} \\
&= \det(\cdots, a^i, \cdots, a^i + a^j, \cdots) + \det(\cdots, a^j, \cdots, a^i + a^j, \cdots) && \text{by (DET1)} \\
&= \det(\cdots, a^i + a^j, \cdots, a^i + a^j, \cdots) && \text{by (DET1)} \\
&= 0, && \text{by (DET2)}
\end{aligned}$$

so $\det(B) = -\det(A)$. □

28.2 Intuition behind the proof via the two by two case

The proof of Theorem 28.1 is postponed to Section 28.5. To get a feeling for how it works, let's see what these properties tell us about the determinant of some arbitrary 2×2 matrix $\left[\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right]$. We can write its first column as $a\left[\begin{smallmatrix} 1 \\ 0 \end{smallmatrix}\right] + c\left[\begin{smallmatrix} 0 \\ 1 \end{smallmatrix}\right]$, so

$$\det\begin{bmatrix} a & b \\ c & d \end{bmatrix} = a \det\begin{bmatrix} 1 & b \\ 0 & d \end{bmatrix} + c \det\begin{bmatrix} 0 & b \\ 1 & d \end{bmatrix}.$$

Likewise, the second column can be written as $b\left[\begin{smallmatrix}1\\0\end{smallmatrix}\right]+d\left[\begin{smallmatrix}0\\1\end{smallmatrix}\right]$, so using linearity of the determinant in the second column, we can split up each of the two matrices on its righthand side and simplify:

$$\det\begin{bmatrix} a & b \\ c & d \end{bmatrix} = ab\underbrace{\det\begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix}}_{=0 \text{ by (DET2)}} + ad\det\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + bc\det\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} + cd\underbrace{\det\begin{bmatrix} 0 & 0 \\ 1 & 1 \end{bmatrix}}_{=0 \text{ by (DET2)}}.$$

So only two terms are left. The first one involves the determinant of the matrix with the standard basis vectors in its usual order (which equals 1 by (DET3)) multiplied with the scalars a and d that used to be in the location of the ones. The second one involves the determinant of the matrix with the standard basis vectors in a different order, again multiplied with precisely those scalars b and c that used to be in the same place as the ones. By (DET4),

$$\det\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} = -\det\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \overset{\text{(DET3)}}{=} -1.$$

Conclude:

$$\det\begin{bmatrix} a & b \\ c & d \end{bmatrix} = ad - bc.$$

This is the crucial general insight:

> Since the determinant is zero if two columns are the same, the process of splitting up each column into standard basis vectors gives a lot of terms that disappear. The only ones that are left have the standard basis vectors — the columns of the identity matrix — arranged in a particular order and multiplied with exactly those entries in the original matrix corresponding with the location of the ones.

Such a rearrangement of n columns is summarized by a ***permutation*** $\sigma : \{1,\ldots,n\} \to \{1,\ldots,n\}$ of the set $\{1,\ldots,n\}$, which is required to be one-to-one and onto. Write $\mathrm{Perm}(n)$ for the set of all these permutations. With this new notation, each remaining term is the product of the determinant

$$\det(e_{\sigma(1)},\ldots,e_{\sigma(n)}), \tag{127}$$

of a matrix obtained by permuting the columns of the identity matrix, multiplied with the product of the entries $a_{\sigma(1),1}$ until $a_{\sigma(n),n}$ in the location of the ones in the matrix above.

The matrix in (127) can be obtained from the identity matrix I by repeatedly exchanging two of its columns. If you need an even number of column switches, (DET4) says that its determinant is the same as that of I; if it takes an odd number of switches, its determinant is $-\det(I)$. In the first case, we say that permutation σ has ***sign*** $\mathrm{sign}(\sigma) = 1$; in the second case it has sign $\mathrm{sign}(\sigma) = -1$. Part of the general proof consists in showing that this is independent of the order in which you switch columns.

Now we know everything to compute the determinant: the determinant in (127) equals $\det(I)\,\mathrm{sign}(\sigma)$ and is multiplied with $\prod_{i=1}^{n} a_{\sigma(i),i}$. Summing over all permutations we find that if a function f satisfies (DET1) and (DET2), it must be of the form

$$f(A) = f(I)\sum_{\sigma\in\mathrm{Perm}(n)}\left(\prod_{i=1}^{n} a_{\sigma(i),i}\right)\mathrm{sign}(\sigma). \tag{128}$$

Once you know the determinant of the identity matrix, you know the determinant of every other matrix as well. Property (DET3) is just an arbitrary normalization. It pins down a unique function satisfying properties (DET1) to (DET3) and that's the one we call the determinant:

$$\det(A) = \sum_{\sigma\in\mathrm{Perm}(n)}\left(\prod_{i=1}^{n} a_{\sigma(i),i}\right)\mathrm{sign}(\sigma).$$

This expression is known as the ***Leibniz formula*** for the determinant.

28.3 Further properties of the determinant

(DET5) For any two matrices A and B, $\det(AB) = \det(A)\det(B)$.

Proof: Fix $n \times n$ matrix A and define the function f that assigns to each $n \times n$ matrix B the value

$$f(B) = \det(AB) = \det(Ab^1, Ab^2, \ldots, Ab^n),$$

where $b^1, \ldots, b^n$ are the columns of B. Since the determinant satisfies (DET1) and (DET2), so does f. Moreover, $f(I) = \det(AI) = \det(A)$. So by (128),

$$\det(AB) = f(B) = \det(A) \sum_{\sigma \in \mathrm{Perm}(n)} \left(\prod_{i=1}^{n} b_{\sigma(i),i} \right) \mathrm{sign}(\sigma) = \det(A)\det(B). \qquad \square$$

(DET6) A matrix and its transpose have the same determinant: $\det(A) = \det(A^\top)$.

Proof: The determinant of $n \times n$ matrix $A^\top$ is

$$\det(A^\top) = \sum_{\sigma \in \mathrm{Perm}(n)} \left(\prod_{i=1}^{n} a_{i,\sigma(i)} \right) \mathrm{sign}(\sigma) = \sum_{\sigma \in \mathrm{Perm}(n)} \left(\prod_{i=1}^{n} a_{\sigma^{-1}(i),i} \right) \mathrm{sign}(\sigma).$$

Permutation σ and its inverse σ^{-1} have the same sign: you can invert/undo whatever rearrangement σ achieves in the same number of steps by simply doing them in opposite order. Moreover, as σ varies over all permutations, so does σ^{-1}. So we can rewrite:

$$\det(A^\top) = \sum_{\sigma \in \mathrm{Perm}(n)} \left(\prod_{i=1}^{n} a_{\sigma^{-1}(i),i} \right) \mathrm{sign}(\sigma^{-1}) = \det(A). \qquad \square$$

Many properties of the determinant are statements about columns. By (DET6), the same properties hold for rows. Some textbooks state the properties of determinants in terms of rows instead of columns.

(DET7) The determinant of a matrix is zero if and only if its columns are linearly dependent.

Proof: Consider an $n \times n$ matrix A. If its columns are linearly dependent, then there are scalars $\alpha_1, \ldots, \alpha_n$, not all zero, such that

$$\alpha_1 a^1 + \cdots + \alpha_n a^n = \mathbf{0}.$$

Without loss of generality (it only simplifies our notation), assume that $\alpha_1 \neq 0$. Then we can write the first column as a linear combination of the others:

$$a^1 = -\frac{\alpha_2}{\alpha_1} a^2 - \cdots - \frac{\alpha_n}{\alpha_1} a^n.$$

Substitute this as the first column and use linearity of the determinant in the first column:

$$\det(A) = \det(a^1, a^2, \ldots, a^n) = \det\left(-\frac{\alpha_2}{\alpha_1} a^2 - \cdots - \frac{\alpha_n}{\alpha_1} a^n, a^2, \ldots, a^n \right) = \sum_{j=2}^{n} -\frac{\alpha_j}{\alpha_1} \det(a^j, a^2, \ldots, a^n).$$

In the final sum, all determinants are zero by (DET2), because the first column is the same as the j-th.

If its n columns are linearly independent, they are a basis of the n-dimensional space. In particular, each standard basis vector is a linear combination of the columns of A, which means that there is a matrix X with $AX = I$. By (DET5), $\det(A)\det(X) = \det(AX) = \det(I) = 1$. So $\det(A) \neq 0$. $\square$

For triangular matrices the determinant is particularly easy to compute.

Definition 28.1 A square matrix A is

- ☒ ***upper triangular*** if all entries below the diagonal are zero: $a_{ij} = 0$ if $i > j$;
- ☒ ***lower triangular*** if all entries above the diagonal are zero: $a_{ij} = 0$ if $i < j$;
- ☒ ***triangular*** if it is upper or lower triangular;
- ☒ a ***diagonal matrix*** if it is both upper and lower triangular: $a_{ij} = 0$ if $i \neq j$.

(DET8) The determinant of a triangular matrix is the product of its diagonal entries.

Proof: Since the proof for lower triangular matrices is similar, I only prove it for upper triangular matrices. Look at the Leibniz formula for the determinant and pick some permutation σ:

- ☒ If σ is the identity permutation with $\sigma(i) = i$ for all $i \in \{1, \ldots, n\}$, then its sign is one, so

$$\left(\prod_{i=1}^{n} a_{\sigma(i),i}\right) \operatorname{sign}(\sigma) = \prod_{i=1}^{n} a_{i,i}$$

  is exactly the product of the diagonal entries.
- ☒ If σ is not the identity permutation, then there is a $k \in \{1, \ldots, n\}$ with $\sigma(k) > k$. (Why? Since $\sum_i \sigma(i) = \sum_i i$ and $\sigma(j) \neq j$ for some j, we cannot have $\sigma(k) \leq k$ for all k.) But then entry $a_{\sigma(k),k}$ equals zero, because it lies below the diagonal. So that permutation contributes zero to the determinant, since

$$\left(\prod_{i=1}^{n} a_{\sigma(i),i}\right) \operatorname{sign}(\sigma) = 0.$$

□

The following is an immediate consequence of multilinearity assumption (DET1):

(DET9) If the $n \times n$ matrix B is obtained from A by multiplying all entries in a specific column of A by a constant α, then $\det(B) = \alpha \det(A)$.

Taking $\alpha = 0$, we find:

(DET10) If A has a column with only zeroes, then $\det(A) = 0$.

And using (DET9) repeatedly for each of the n columns, we find:

(DET11) If the $n \times n$ matrix B is obtained from A by multiplying all entries of A by a constant α, then $\det(B) = \alpha^n \det(A)$.

(DET12) If two columns of A are proportional, then $\det(A) = 0$.

The latter property follows from (DET7). Finally,

(DET13) The determinant is unchanged if a multiple of one column of A is added to a different column.

Proof: Suppose the $n \times n$ matrix B is obtained from A by adding α times the i-th column of A to the j-th column of A (where $i \neq j$). By linearity (DET1), $\det(B)$ is equal to the determinant of A plus α times the determinant of the matrix obtained from A by replacing the j-th column with the i-th column. But the latter matrix has two identical columns i and j, so its determinant is zero by (DET2). □

Example 28.1 (Computing the determinant using Gaussian elimination) Property (DET6) together with (DET4), (DET9), and (DET13) tells how the determinant of a matrix is affected by the three operations in Gaussian elimination:

- ☒ exchanging the order of two rows changes the determinant by a factor -1;
- ☒ multiplying all entries in a row by a scalar $\alpha \neq 0$ multiplies the determinant by α;
- ☒ adding a multiple of one row to another does not affect its determinant.

And (DET8) says that the determinant of an upper triangular matrix is the product of the diagonal entries. So you can mechanically compute the determinant using Gaussian elimination:

$$\begin{aligned}
\det\begin{bmatrix} 0 & -2 & 1 \\ 4 & 1 & -1 \\ 4 & -3 & 4 \end{bmatrix} &= -\det\begin{bmatrix} 4 & 1 & -1 \\ 0 & -2 & 1 \\ 4 & -3 & 4 \end{bmatrix} \\
&= -\det\begin{bmatrix} 4 & 1 & -1 \\ 0 & -2 & 1 \\ 0 & -4 & 5 \end{bmatrix} \\
&= -\det\begin{bmatrix} 4 & 1 & -1 \\ 0 & -2 & 1 \\ 0 & 0 & 3 \end{bmatrix} \\
&= 24. \qquad \triangleleft
\end{aligned}$$

28.4 Expansion by cofactors

We already saw two ways of computing determinants: the Leibniz formula and Gaussian elimination. Here comes a third. Using linearity of the determinant in each of its rows or columns helps to create simpler matrices with a lot of entries equal to zero. This even helps us to write the determinant of a large matrix in terms of those of smaller matrices. That is the idea behind the method called ***expansion by cofactors***. The crucial step how a determinant of a $(k+1)\times(k+1)$ matrix with lots of zeros reduces to a determinant of a smaller $k\times k$ matrix is illustrated in the following example.

Example 28.2 If A is a $k\times k$ matrix, then it has the same determinant as $(k+1)\times(k+1)$ matrices

$$B = \begin{bmatrix} a_{11} & \cdots & a_{1k} & 0 \\ \vdots & & \vdots & \vdots \\ a_{k1} & \cdots & a_{kk} & 0 \\ 0 & \cdots & 0 & 1 \end{bmatrix} = \left[\begin{array}{ccc|c} & & & 0 \\ & A & & \vdots \\ & & & 0 \\ \hline 0 & \cdots & 0 & 1 \end{array}\right] \text{ and } C = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & a_{11} & \cdots & a_{1k} \\ \vdots & \vdots & & \vdots \\ 0 & a_{k1} & \cdots & a_{kk} \end{bmatrix} = \left[\begin{array}{c|ccc} 1 & 0 & \cdots & 0 \\ \hline 0 & & & \\ \vdots & & A & \\ 0 & & & \end{array}\right]$$

Since C is obtained from B by k row and k column interchanges, $\det(B) = \det(C)$. It remains to argue why $\det(B) = \det(A)$. The determinant of $(k+1)\times(k+1)$ matrix B is

$$\sum_{\sigma\in\mathrm{Perm}(k+1)} \left(\prod_{i=1}^{k+1} b_{\sigma(i),i}\right) \mathrm{sign}(\sigma).$$

The product in brackets is zero whenever $\sigma(k+1) \neq k+1$, so we can restrict attention to permutations σ with $\sigma(k+1) = k+1$. In that case, the product equals

$$b_{k+1,k+1} \prod_{i=1}^{k} b_{\sigma(i),i} = 1 \cdot \prod_{i=1}^{k} a_{\sigma(i),i}.$$

Here we use the restriction of σ to $\{1,\ldots,k\}$ which by construction is a permutation of $\{1,\ldots,k\}$ with the same sign as its 'sibling' σ on all of $\{1,\ldots,k,k+1\}$. So the determinant of B can be rewritten as

$$\sum_{\sigma\in\mathrm{Perm}(k)} \left(\prod_{i=1}^{k} a_{\sigma(i),i}\right) \mathrm{sign}(\sigma),$$

which is the determinant of A. ◁

Theorem 28.2 (Expansion by cofactors)

Let A be an $n \times n$ matrix. For each column j,

$$\det(A) = \sum_{i=1}^{n} a_{ij}(-1)^{i+j} \det(A_{ij}) \tag{129}$$

and for each row i,

$$\det(A) = \sum_{j=1}^{n} a_{ij}(-1)^{i+j} \det(A_{ij}),$$

where A_{ij} is the $(n-1) \times (n-1)$ matrix obtained from A by omitting its i-th row and j-th column.

The number $(-1)^{i+j} \det(A_{ij})$ is called the (i,j)-th ***cofactor*** of A.

Proof: We do the proof for columns; for rows it follows from (DET6) applied to the transpose of A. By linearity of the determinant in the j-th column,

$$\det(A) = \sum_{i=1}^{n} a_{ij} \det(a^1, \ldots, a^{j-1}, e_j, a^{j+1}, \ldots, a^n).$$

Fix $i \in \{1, \ldots, n\}$. We will find an expression for

$$\det(a^1, \ldots, a^{j-1}, e_j, a^{j+1}, \ldots, a^n) = \det \begin{bmatrix} a_{1,1} & \cdots & a_{1,j-1} & 0 & a_{1,j+1} & \cdots & a_{1,n} \\ \vdots & & \vdots & \vdots & \vdots & & \vdots \\ a_{i-1,1} & \cdots & a_{i-1,j-1} & 0 & a_{i-1,j+1} & \cdots & a_{i-1,n} \\ a_{i,1} & \cdots & a_{i,j-1} & 1 & a_{i,j+1} & \cdots & a_{i,n} \\ a_{i+1,1} & \cdots & a_{i+1,j-1} & 0 & a_{i+1,j+1} & \cdots & a_{i+1,n} \\ \vdots & & \vdots & \vdots & \vdots & & \vdots \\ a_{n,1} & \cdots & a_{n,j-1} & 0 & a_{n,j+1} & \cdots & a_{n,n} \end{bmatrix}. \tag{130}$$

Adding $-a_{i,k}$ times column j to column k does not affect the determinant by (DET13). Doing this for all columns $k \neq j$ shows that (130) equals

$$\det \begin{bmatrix} a_{1,1} & \cdots & a_{1,j-1} & 0 & a_{1,j+1} & \cdots & a_{1,n} \\ \vdots & & \vdots & \vdots & \vdots & & \vdots \\ a_{i-1,1} & \cdots & a_{i-1,j-1} & 0 & a_{i-1,j+1} & \cdots & a_{i-1,n} \\ 0 & \cdots & 0 & 1 & 0 & \cdots & 0 \\ a_{i+1,1} & \cdots & a_{i+1,j-1} & 0 & a_{i+1,j+1} & \cdots & a_{i+1,n} \\ \vdots & & \vdots & \vdots & \vdots & & \vdots \\ a_{n,1} & \cdots & a_{n,j-1} & 0 & a_{n,j+1} & \cdots & a_{n,n} \end{bmatrix}.$$

Moving row i to the top one interchange at a time takes $i-1$ steps. By (DET4), (130) equals

$$(-1)^{i-1}\det\begin{bmatrix} 0 & \cdots & 0 & 1 & 0 & \cdots & 0 \\ a_{1,1} & \cdots & a_{1,j-1} & 0 & a_{1,j+1} & \cdots & a_{1,n} \\ \vdots & & \vdots & \vdots & \vdots & & \vdots \\ a_{i-1,1} & \cdots & a_{i-1,j-1} & 0 & a_{i-1,j+1} & \cdots & a_{i-1,n} \\ a_{i+1,1} & \cdots & a_{i+1,j-1} & 0 & a_{i+1,j+1} & \cdots & a_{i+1,n} \\ \vdots & & \vdots & \vdots & \vdots & & \vdots \\ a_{n,1} & \cdots & a_{n,j-1} & 0 & a_{n,j+1} & \cdots & a_{n,n} \end{bmatrix}.$$

Moving column j to the left one interchange at a time takes $j-1$ steps. Using (DET4) and $(-1)^{i-1+j-1} = (-1)^{i+j}$, (130) equals

$$(-1)^{i+j}\det\begin{bmatrix} 1 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & a_{1,1} & \cdots & a_{1,j-1} & a_{1,j+1} & \cdots & a_{1,n} \\ \vdots & \vdots & & \vdots & \vdots & & \vdots \\ 0 & a_{i-1,1} & \cdots & a_{i-1,j-1} & a_{i-1,j+1} & \cdots & a_{i-1,n} \\ 0 & a_{i+1,1} & \cdots & a_{i+1,j-1} & a_{i+1,j+1} & \cdots & a_{i+1,n} \\ \vdots & \vdots & & \vdots & \vdots & & \vdots \\ 0 & a_{n,1} & \cdots & a_{n,j-1} & a_{n,j+1} & \cdots & a_{n,n} \end{bmatrix} = (-1)^{i+j}\det\left[\begin{array}{c|ccc} 1 & 0 & \cdots & 0 \\ \hline 0 & & & \\ \vdots & & A_{ij} & \\ 0 & & & \end{array}\right].$$

By Example 28.2, this is just $(-1)^{i+j}\det(A_{ij})$. Substituting this back into (130) proves (129). □

Example 28.3 Consider an $n \times n$ matrix in ***block form***

$$\begin{bmatrix} A & B \\ O & C \end{bmatrix}$$

where A is a $k \times k$ matrix and C an $\ell \times \ell$ matrix for positive integers with $k+\ell = n$, B is a $k \times \ell$ matrix, and O is the $\ell \times k$ matrix of zeroes. Then its determinant is $\det(A)\det(C)$.

To see this, define for arbitrary matrices of these dimensions the function f with values

$$f(A,B,C) = \det\begin{bmatrix} A & B \\ O & C \end{bmatrix}.$$

Keeping B and C fixed, f is multilinear (DET1) and alternating (DET4), so by (128), it satisfies

$$f(A,B,C) = \sum_{\sigma\in\mathrm{Perm}(k)}\left(\prod_{i=1}^{n} a_{\sigma(i),i}\right)\mathrm{sign}(\sigma)\,f(I,B,C) = \det(A)\,f(I,B,C).$$

Subtracting multiples of the columns of I from the columns of B does not affect the determinant by (DET13), so

$$f(I,B,C) = f(I,O,C).$$

Now either repeating the argument above for the columns of A, but now for the columns of C, or invoking Example 28.2 gives $f(I,O,C) = \det(C)$. Combining all this,

$$\det\begin{bmatrix} A & B \\ O & C \end{bmatrix} = f(A,B,C) = \det(A)f(I,B,C) = \det(A)f(I,O,C) = \det(A)\det(C).$$

◁

28.5 Postponed proofs

28.5.1 The proof of Theorem 28.1

Let A be an $n \times n$ matrix with columns $a^1, \ldots, a^n$. Write its first column as a linear combination of the standard basis vectors: $a^1 = a_{1,1}e_1 + \cdots + a_{n,1}e_n = \sum_{j_1=1}^{n} a_{j_1,1}e_{j_1}$. Since the determinant is linear in the first column,

$$\det(A) = \sum_{j_1=1}^{n} a_{j_1,1} \det(e_{j_1}, a^2, \cdots, a^n).$$

It is linear in the other columns as well. So we can repeat this and obtain

$$\det(A) = \sum_{j_1=1}^{n} \sum_{j_2=1}^{n} \cdots \sum_{j_n=1}^{n} a_{j_1,1} a_{j_2,2} \cdots a_{j_n,n} \det(e_{j_1}, e_{j_2}, \ldots, e_{j_n}).$$

Just as above, where we used summation index j_1 for the terms in the first column, we used summation index j_2 for the second column, j_3 for the third, and so on. These n summation indices range from 1 to n, so the sum consists of n^n terms. Most terms are zero: if any two of the indices $j_1, \ldots, j_n$ are the same, the determinant $\det(e_{j_1}, e_{j_2}, \ldots, e_{j_n})$ is zero, because it has two identical columns.

The only remaining terms are those where the n summation indices $j_1, \ldots, j_n$ have n distinct values: they are a rearrangement of the numbers $1, 2, \ldots, n$. Such a rearrangement is formally defined by a ***permutation*** $\sigma : \{1, \ldots, n\} \to \{1, \ldots, n\}$. A permutation is required to be both one-to-one (distinct arguments have distinct function values) and onto (its range is all of $\{1, \ldots, n\}$), i.e., it is a bijection. It rearranges $1, 2, \ldots, n$ in a new order $\sigma(1), \sigma(2), \ldots, \sigma(n)$. Writing $\mathrm{Perm}(n)$ for the set of permutations of $\{1, \ldots, n\}$, the formula for the determinant simplifies to

$$\det(A) = \sum_{\sigma \in \mathrm{Perm}(n)} a_{\sigma(1),1} a_{\sigma(2),2} \cdots a_{\sigma(n),n} \det(e_{\sigma(1)}, e_{\sigma(2)}, \ldots, e_{\sigma(n)}). \tag{131}$$

The matrix with columns $e_{\sigma(1)}, e_{\sigma(2)}, \ldots, e_{\sigma(n)}$ is obtained from the identity matrix by interchanging/rearranging its columns. By (DET3), its determinant is 1 if $\sigma(1), \sigma(2), \ldots, \sigma(n)$ can be obtained from $1, 2, \ldots, n$ after an even number of interchanges and -1 otherwise.

To formalize this, define the ***sign*** of a permutation σ to be $\mathrm{sign}(\sigma) = 1$ if it takes an even number of interchanges to rearrange $(1, 2, \ldots, n)$ into $(\sigma(1), \sigma(2), \ldots, \sigma(n))$ and $\mathrm{sign}(\sigma) = -1$ if the number of exchanges is odd. This is well-defined: although there are infinitely many ways to achieve this rearrangement, the miraculous thing is that the number of interchanges is either always even or always odd. Why? We will compare the sign function with another function, the signum, which *is* evidently uniquely defined and show they are the same.

Given permutation σ, an inversion is a pair of numbers that the permutation puts out of order: a pair $i, j \in \{1, \ldots, n\}$ with $i < j$ but $\sigma(i) > \sigma(j)$. Count the number I of inversions. Define the signum of σ to be $(-1)^I$. Now observe:

(a) The identity permutation with $\sigma(k) = k$ for all $k = 1, \ldots, n$ has zero inversions: its signum is 1.

(b) Starting with a permutation and interchanging two neighbors changes the order of only that pair of entries. So the number of inversions goes up by one if the neighbors were in the right order and now they're inverted; it goes down by one if they were in the wrong order and are now put right. Either way, the signum changes.

(c) So going from one permutation to another always requires an even number of neighbor switches if they have the same signum and an odd number if they have different signum.

(d) Switching two entries can be done via an odd number of neighbor switches. If the entries are k places apart, it takes k neighbor switches to move the entry in the back to the front and then $k-1$ more to put the front one to the back, a total of $2k-1$ neighbor switches. If you don't see this immediately, try it out in a specific example.

(e) So the signum changes if two entries are interchanged.

In summary: if you start from $1,2,\ldots,n$ (with signum 1) and do an even number of interchanges, the sign (by definition) and signum (by the previous step) are 1; after an odd number of interchanges, both sign and signum are -1. So sign and signum coincide, making the sign of a permutation well-defined.

So we now know that the determinant $\det(e_{\sigma(1)}, e_{\sigma(2)}, \ldots, e_{\sigma(n)})$ equals the sign of permutation σ. Substituting this into (131) we see that if we want the determinant to satisfy axioms (DET1) to (DET3), then we must define it as

$$\det(A) = \sum_{\sigma\in\mathrm{Perm}(n)} a_{\sigma(1),1}\, a_{\sigma(2),2} \cdots a_{\sigma(n),n}\, \mathrm{sign}(\sigma) = \sum_{\sigma\in\mathrm{Perm}(n)} \left(\prod_{i=1}^{n} a_{\sigma(i),i}\right) \mathrm{sign}(\sigma).$$

It remains to verify that this candidate actually satisfies the axioms:

- ☒ Linearity in each column follows from the fact that for each permutation σ, the product $\prod_{i=1}^{n} a_{\sigma(i),i}$ contains exactly one term from the column under consideration.
- ☒ The identity matrix has determinant one: if $\sigma \in \mathrm{Perm}(n)$ is not the identity, then $\sigma(i) \neq i$ for some i. The corresponding entry $a_{\sigma(i),i}$ of the identity matrix is off-diagonal and consequently equal to zero. So also $\prod_{i=1}^{n} a_{\sigma(i),i} = 0$: those permutations contribute zero to the determinant. If σ is the identity, then its sign is one and $\prod_{i=1}^{n} a_{\sigma(i),i} = \prod_{i=1}^{n} a_{i,i} = 1$. So the determinant is one.
- ☒ The determinant is zero if two columns, let's say columns k and ℓ, are the same. We divide the permutations into pairs: match permutation σ to permutation σ' that is the same as σ, except that k and ℓ switch places:

  $$\sigma'(k) = \sigma(\ell), \quad \sigma'(\ell) = \sigma(k), \quad \sigma'(m) = \sigma(m) \text{ for all other } m \in \{1,\ldots,n\}.$$

  By (e), these permutations have opposite sign. Since column k and ℓ are the same, also products $\prod_{i=1}^{n} a_{\sigma(i),i}$ and $\prod_{i=1}^{n} a_{\sigma'(i),i}$ are the same. So these terms cancel out and the determinant is zero.

29 Eigenvalues and eigenvectors

29.1 What are they and do they exist?

Given a linear function T from and to a vector space V, computing its function values $T(x)$ can be tedious. But sometimes life is simple. If there is a nonzero vector x such that $T(x) = \lambda x$ for some scalar λ, computing $T(x)$ reduces to a simple scalar multiplication. Since a lot of linear algebra is about applying linear functions, such easy cases are sufficiently important to have deserved an own name: x is an eigenvector and the rescaling factor λ an eigenvalue.

Before moving to the general definition, recall (see Remark 3.1) that scalar multiplication uses scalars from a field F. For most economic applications, *real* vector spaces ($F = \mathbb{R}$) are enough. But for eigenvalues and eigenvectors it is often convenient to also consider *complex* vector spaces ($F = \mathbb{C}$). Appendix B reviews all you need to know about complex numbers for this course.

Definition 29.1 Let V be a vector space over a field F.

- ☒ A ***linear operator*** is a linear function $T : V \to V$ from and to V.
- ☒ A scalar $\lambda \in F$ is an ***eigenvalue*** of T if there is a vector $x \neq \mathbf{0}$ in V with $T(x) = \lambda x$.
- ☒ This x is an ***eigenvector*** corresponding with eigenvalue λ.

The ***identity operator*** $I : V \to V$ is defined by $I(x) = x$ for all $x \in V$. Using it, we can rearrange the equation $T(x) = \lambda x$ to $(T - \lambda I)(x) = \mathbf{0}$. So the eigenvectors x corresponding to eigenvalue λ lie in the null space/kernel of the linear function $T - \lambda I$. This subspace is called the ***eigenspace*** of eigenvalue λ.

For an $n \times n$ matrix A, its eigenvalues/-vectors are those of the associated linear operator T with $T(x) = Ax$, i.e., scalars λ and vectors $x \neq \mathbf{0}$ with $Ax = \lambda x$. As above, we can rewrite this as $(A - \lambda I)x = \mathbf{0}$ if we introduce the $n \times n$ ***identity matrix*** I that satisfies $Ix = x$ for all vectors x. This is the matrix with ones on the diagonal ($a_{11} = a_{22} = \cdots = a_{nn} = 1$) and zeros everywhere else ($a_{ij} = 0$ if $i \neq j$). For example, the 2×2 and 3×3 identity matrices are

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

The obvious first question is if eigenvalues/-vectors always exist. Not if only real numbers are used:

Example 29.1 (Real eigenvalues do not always exist) The linear operator $T : \mathbb{R}^2 \to \mathbb{R}^2$ with $T(x_1, x_2) = (-x_2, x_1)$ or, in matrix form,

$$T(x) = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} x,$$

has no real eigenvalues. If x is a nonzero vector, then $T(x)$ is orthogonal to x: it rotates x by 90° counterclockwise (draw!). So this cannot result in a rescaling λx of x. Formally: suppose, to the contrary, that there are a scalar $\lambda \in \mathbb{R}$ and a vector $x \neq \mathbf{0}$ in $\mathbb{R}^2$ with $T(x) = \lambda x$:

$$(-x_2, x_1) = \lambda(x_1, x_2).$$

So $-x_2 = \lambda x_1$ and $x_1 = \lambda x_2$. Substitute the latter into the former to find that $(\lambda^2 + 1)x_2 = 0$. Since $\lambda^2 + 1 > 0$, this implies $x_2 = 0$. But then $x_1 = \lambda x_2 = 0$ as well, contradicting that $x \neq \mathbf{0}$. ◁

Verify by substitution in $T(x) = \lambda x$ that complex eigenvalues and eigenvectors do exist: $\lambda = i$ is an eigenvalue with eigenvector $(i, 1)$ and $\lambda = -i$ is an eigenvalue with eigenvector $(-i, 1)$. So is it enough to just move to complex vector spaces? No, some linear operators on infinite-dimensional vector spaces don't have eigenvalues at all, not even complex ones:

Example 29.2 (Eigenvalues do not always exist) The linear operator T on the complex vector space of sequences in $\mathbb{C}$ defined for each such sequence $(x_1, x_2, x_3, x_4, \ldots)$ by

$$T(x_1, x_2, x_3, x_4, \ldots) = (0, x_1, x_2, x_3, \ldots)$$

has no eigenvalues. Otherwise, there are a scalar $\lambda \in \mathbb{C}$ and sequence $x = (x_1, x_2, \ldots) \neq \mathbf{0}$ with $T(x) = \lambda x$:

$$(0, x_1, x_2, x_3, \ldots) = \lambda(x_1, x_2, x_3, x_4, \ldots).$$

If $\lambda = 0$, the right side is the zero vector and consequently $x = \mathbf{0}$, a contradiction. So $\lambda \neq 0$. Equating the first terms, $0 = \lambda x_1$, we see that $x_1 = 0$. The other terms say that $x_n = \lambda x_{n+1}$ for all $n \in \mathbb{N}$. By induction: $x = \mathbf{0}$, a contradiction. So no such λ and x exist. ◁

Now the good news: if complex numbers are allowed and you steer clear of infinite-dimensional vector spaces, eigenvalues/-vectors do exist:

Theorem 29.1 (Existence of eigenvalues)

Each linear operator on a finite-dimensional complex vector space other than $\{\mathbf{0}\}$ has an eigenvalue.

There is an elegant proof[8] without determinants; I provide it in section 29.4. But the following proof for square matrices gives a handy way to compute eigenvalues in small examples. An $n \times n$ matrix A of real or complex numbers has eigenvalue λ if $(A - \lambda I)x = \mathbf{0}$ for some vector $x \neq \mathbf{0}$: the columns of $A - \lambda I$ are linearly dependent. By (DET7), this happens exactly when $\det(A - \lambda I) = 0$:

$$\lambda \text{ is an eigenvalue of } A \iff \det(A - \lambda I) = 0.$$

By Leibniz's formula, the determinant $\det(A - \lambda I)$ is an n-th degree polynomial function of λ, the ***characteristic polynomial*** of A. So if we compute this characteristic polynomial and set it equal to zero, its roots are the desired eigenvalues. The Fundamental Theorem of Algebra (Theorem B.1) assures that this polynomial indeed has n not necessarily distinct roots in the set of complex numbers.

Example 29.3 For the 2×2 matrix

$$A = \begin{bmatrix} -3 & 1 \\ -4 & 2 \end{bmatrix},$$

the characteristic polynomial $\det(A - \lambda I)$ is the determinant of the matrix

$$A - \lambda I = \begin{bmatrix} -3 & 1 \\ -4 & 2 \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} -3 - \lambda & 1 \\ -4 & 2 - \lambda \end{bmatrix},$$

i.e., the polynomial

$$(-3 - \lambda)(2 - \lambda) - 1(-4) = \lambda^2 + \lambda - 2 = (\lambda - 1)(\lambda + 2).$$

Setting it equal to zero gives its two eigenvalues $\lambda = 1$ and $\lambda = -2$. ◁

Once we know the eigenvalues λ, the corresponding eigenvectors can be found by solving the system of linear equations $(A - \lambda I)x = \mathbf{0}$, for instance using Gaussian elimination. In our previous example, you should verify that each eigenvector corresponding with eigenvalue 1 is a multiple of $x = (1, 4)$ and each eigenvector corresponding with eigenvalue -2 is a multiple of $x = (1, 1)$.

Sometimes eigenvalues are easy to find:

[8] From Sheldon Axler (1995), Down with determinants! *American Mathematical Monthly* 102, 139–154.

Example 29.4 (Eigenvalues of triangular matrices) If A is a triangular matrix, then the same holds for $A - \lambda I$. By (DET8), its determinant is the product of the diagonal entries:

$$\det(A - \lambda I) = (a_{11} - \lambda)(a_{22} - \lambda) \cdots (a_{nn} - \lambda).$$

Setting it equal to zero we see that *the eigenvalues of a triangular matrix A are the numbers $a_{11}, \ldots, a_{nn}$ on its diagonal.* For example, the eigenvalues of

$$\begin{bmatrix} 1 & 0 \\ -2 & 3 \end{bmatrix}$$

are 1 and 3. ◁

The next result tells roughly where to find the eigenvalues of a square matrix.

Theorem 29.2 (Gerschgorin's disc theorem)

Each eigenvalue of an $n \times n$ (real or complex) matrix A lies in at least one disc $D_1, \ldots, D_n$, where

$$D_i = \left\{ \lambda : |\lambda - a_{ii}| \leq \sum_{j \neq i} |a_{ij}| \right\}$$

has center a_{ii} and radius $\sum_{j \neq i} |a_{ij}|$.

Proof: If λ is an eigenvalue with eigenvector x, then $Ax = \lambda x$. Since $x \neq \mathbf{0}$, there is a coordinate i with largest absolute value $|x_i| > 0$. We argue that λ lies in disk D_i. Rewrite coordinate i of $Ax = \lambda x$:

$$\sum_{j=1}^{n} a_{ij} x_j = \lambda x_i \qquad \text{so} \qquad \sum_{j \neq i} a_{ij} \frac{x_j}{x_i} = \lambda - a_{ii}.$$

Using the triangle inequality and $\frac{|x_j|}{|x_i|} \leq 1$ gives

$$|\lambda - a_{ii}| = \left| \sum_{j \neq i} a_{ij} \frac{x_j}{x_i} \right| \leq \sum_{j \neq i} |a_{ij}| \frac{|x_j|}{|x_i|} \leq \sum_{j \neq i} |a_{ij}|.$$

□

Observe that this proof works for both real and complex numbers.

Example 29.5 Let's apply it to find where the real eigenvalues of the following matrix might be:

$$A = \begin{bmatrix} 1 & 0 & -2 \\ 3 & -2 & -5 \\ -2 & 1 & 4 \end{bmatrix}$$

The disc D_1 has center $a_{11} = 1$ and its radius is the sum of the absolute values of the remaining entries in the first row: $0 + |-2| = 2$. So it is the set of real numbers whose distance to 1 is less than or equal to 2, i.e., $D_1 = [-1, 3]$. Likewise,

$$D_2 = \{r \in \mathbb{R} : |r - (-2)| \leq 3 + |-5|\} = \{r \in \mathbb{R} : |r - (-2)| \leq 8\} = [-10, 6],$$
$$D_3 = \{r \in \mathbb{R} : |r - 4| \leq |-2| + 1|\} = \{r \in \mathbb{R} : |r - 4| \leq 3\} = [1, 7].$$

So the real eigenvalues must lie in their union, $[-10, 7]$. Your favorite math software will tell you that its eigenvalues are roughly 4.1, -1.3, and 0.2, which indeed lie in the required interval. ◁

Example 29.6 (Stationary distributions of Markov chains) Let $n \in \mathbb{N}$. The ***unit simplex***

$$\Delta_n = \{x \in \mathbb{R}^n : x \geq \mathbf{0}, x_1 + \cdots + x_n = 1\}$$

consists of all probability vectors in $\mathbb{R}^n$: they are nonnegative and add up to one. A square matrix $A \in \mathbb{R}^{n \times n}$ is a ***stochastic matrix*** if each column (or each row, depending on an arbitrary choice of direction) is a probability vector. In the theory on Markov chains, stochastic matrices are used to describe transition probabilities: $a_{ij} \in [0,1]$ indicates (notice the order) that if you're in state j today, you're in state i tomorrow. For instance, suppose there are two states, state 1 being good weather and state 2 being bad, and the transition probability matrix is

$$\begin{bmatrix} \frac{4}{5} & \frac{1}{3} \\ \frac{1}{5} & \frac{2}{3} \end{bmatrix}.$$

The first column says that if the weather is good today, the probability of the weather being good tomorrow is $4/5$ and the probability of the weather being bad is $1/5$. The second column is interpreted likewise.

If a probability vector $x \in \Delta_n$ specifies the probability of being in any of the n states today, then Ax is the probability distribution over the states tomorrow. Observe that Ax indeed lies in Δ_n: it is a convex combination of the columns on A. Since each column of A lies in the convex set Δ_n, so does their convex combination.

We call $x \in \Delta_n$ a ***stationary distribution*** if the distribution over states remains unchanged over time: $Ax = x$. So a stationary distribution is an eigenvector corresponding with eigenvalue $\lambda = 1$. Equivalently, it is a fixed point of the function $f : \Delta_n \to \Delta_n$ with $f(x) = Ax$. This function is linear, hence continuous. And Δ_n is nonempty, convex, and compact. So a stationary distribution exists by Brouwer's fixed-point theorem.

Gerschgorin's disc theorem implies that all other eigenvalues are less than or equal to one in absolute value. First notice that A and $A^\top$ have the same eigenvalues (Exc. 29.1). So each eigenvalue λ lies in a disc D_i corresponding with some row i of $A^\top$. This disc has center a_{ii} and radius $1 - a_{ii}$, since its entries are nonnegative and sum to one. So

$$D_i = \{\lambda : |\lambda - a_{ii}| \leq 1 - a_{ii}\}.$$

And by the triangle inequality, λ in this disc satisfies

$$|\lambda| = |\lambda - a_{ii} + a_{ii}| \leq |\lambda - a_{ii}| + a_{ii} \leq (1 - a_{ii}) + a_{ii} = 1. \qquad \triangleleft$$

Theorem 29.3 (Perron-Frobenius)

Let A be a square matrix with positive real entries ($a_{ij} > 0$ for all i, j). Then A has a positive real eigenvalue. The largest such eigenvalue, call it λ^*, has a corresponding positive eigenvector. All other (real/complex) eigenvalues λ satisfy $|\lambda| \leq \lambda^*$.

29.2 Bases of eigenvectors: diagonalization

We motivated the importance of eigenvalues and eigenvectors in terms of the resulting simplicity of applying linear maps: they just rescale the eigenvectors. So imagine how easy life would be if, given

some linear operator $T : V \to V$, we could find an entire basis $x_1, \ldots, x_n$ of V that consists of eigenvectors of T. In that case, each vector $v \in V$ can be written uniquely as a linear combination

$$v = \alpha_1 x_1 + \cdots + \alpha_n x_n$$

of these eigenvectors and plugging this into the linear operator T we find

$$T(v) = \alpha_1 T(x_1) + \cdots + \alpha_n T(x_n) = \alpha_1(\lambda_1 x_1) + \cdots + \alpha_n(\lambda_n x_n), \tag{132}$$

where λ_i is the eigenvalue corresponding with eigenvector x_i. And doing this iteratively, we can also easily find higher powers $T^k(v)$. Remember that

$$T^k = \underbrace{T \circ \cdots \circ T}_{k \text{ terms}}$$

is the k-fold composition of T with itself: $T^2(v) = T(T(v))$, $T^3(v) = T(T(T(v)))$, and so on. By (132) and induction:

$$T^k(v) = \alpha_1 \lambda_1^k x_1 + \cdots + \alpha_n \lambda_n^k x_n.$$

Similarly, if $T(v) = Av$ for some square matrix A,

$$A^k v = \alpha_1 \lambda_1^k x_1 + \cdots + \alpha_n \lambda_n^k x_n. \tag{133}$$

Example 29.7 Consider the matrix

$$A = \begin{bmatrix} -3 & 1 \\ -4 & 2 \end{bmatrix},$$

from Example 29.3 and vector $v = (-1, 5)$. What is $A^7 v$?

This matrix had eigenvalues $\lambda_1 = 1$ and $\lambda_2 = -2$ with corresponding eigenvectors $x_1 = (1, 4)$ and $x_2 = (1, 1)$. These eigenvectors are linearly independent, so they are a basis of $\mathbb{R}^2$. In particular, we can write v as a linear combination $v = 2x_1 - 3x_2$. By (133),

$$A^7 v = 2\lambda_1^7 x_1 - 3\lambda_2^7 x_2 = 2 \begin{bmatrix} 1 \\ 4 \end{bmatrix} - 3(-2)^7 \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

That, of course, is a lot easier than computing A^7 explicitly. ◁

Since computing higher powers of matrices is hugely important in applications (in the Math II course, for instance, this is crucial for solving linear difference and differential equations), such a basis of eigenvectors is the holy grail of matrix multiplication. But such a basis does not always exist:

Example 29.8 (No basis of eigenvectors) The triangular matrix

$$A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$$

has a single eigenvalue $\lambda = 0$. The corresponding eigenspace is the null space of $A - \lambda I = A$: it is the span of $\{e_1\}$. Since this is only one-dimensional, there is no basis of eigenvectors. ◁

Are there interesting special cases where a basis of eigenvectors does exist? The following is useful:

Theorem 29.4 (Eigenvectors corresponding with distinct eigenvalues are linearly independent)

If operator T has distinct eigenvalues $\lambda_1, \ldots, \lambda_m$ and corresponding eigenvectors $v_1, \ldots, v_m$, then $v_1, \ldots, v_m$ are linearly independent.

Proof: I show by induction that for each $k = 1, \ldots, m$, vectors $\{v_1, \ldots, v_k\}$ are linearly independent. This is true if $k = 1$: $v_1 \neq \mathbf{0}$, so $\{v_1\}$ is linearly independent. Now let $k \in \{1, \ldots, m-1\}$ and assume $\{v_1, \ldots, v_k\}$ is linearly independent. Why is the same true for $\{v_1, \ldots, v_k, v_{k+1}\}$? Consider scalars α_i such that

$$\alpha_1 v_1 + \cdots + \alpha_k v_k + \alpha_{k+1} v_{k+1} = \mathbf{0}. \tag{134}$$

Apply linear operator $T - \lambda_{k+1} I$ to both sides of the equation. Since $(T - \lambda_{k+1} I) v_i = (\lambda_i - \lambda_{k+1}) v_i$, its final term with $i = k+1$ is $\mathbf{0}$ and drops out:

$$(T - \lambda_{k+1} I)(\alpha_1 v_1 + \cdots + \alpha_k v_k + \alpha_{k+1} v_{k+1}) = \alpha_1 (\lambda_1 - \lambda_{k+1}) v_1 + \cdots + \alpha_k (\lambda_k - \lambda_{k+1}) v_k = \mathbf{0}.$$

By assumption, vectors $\{v_1, \ldots, v_k\}$ are linearly independent, so for each $i = 1, \ldots, k$, scalar $\alpha_i (\lambda_i - \lambda_{k+1})$ must be zero. The eigenvalues are distinct: $\lambda_i - \lambda_{k+1} \neq 0$. So $\alpha_i = 0$ for all $i = 1, \ldots, k$. Substituting this in (134) and using $v_{k+1} \neq \mathbf{0}$ gives $\alpha_{k+1} = 0$. So vectors $\{v_1, \ldots, v_k, v_{k+1}\}$ are linearly independent. □

This gives us our first important special case: if an $n \times n$ matrix of real or complex numbers has n distinct eigenvalues, then any n corresponding eigenvectors are a basis of the n-dimensional space $\mathbb{R}^n$ (or $\mathbb{C}^n$ if some eigenvalues/-vectors involve nonreal numbers). Of course, complex numbers are a bit of a nuisance and don't always make sense when discussing economic applications (what are $3 + 7\sqrt{-1}$ apples?), so it would be nice if we could assure that all eigenvalues/-vectors of a real matrix are real. Example 29.1 already showed that, in general, we are out of luck. But symmetric matrices or operators $T : \mathbb{R}^n \to \mathbb{R}^n$ are particularly well-behaved: our second special case.

Definition 29.2 A linear operator $T : \mathbb{R}^n \to \mathbb{R}^n$ is ***symmetric*** if for all vectors $x, y \in \mathbb{R}^n$:

$$\langle T(x), y \rangle = \langle x, T(y) \rangle.$$

Here, $\langle \cdot, \cdot \rangle$ is the usual inner product on $\mathbb{R}^n$. Some authors use 'self-adjoint' or 'Hermitian' instead of 'symmetric'. The word 'symmetric' is more informative: if we write T as a matrix multiplication $T(x) = Ax$, then T is symmetric if and only if A is a ***symmetric matrix***: $A = A^\top$. To see this, notice that

$$\langle T(x), y \rangle = \langle x, T(y) \rangle \iff (Ax)^\top y = x^\top (Ay) \iff x^\top A^\top y = x^\top A y.$$

If we take $x = e_i$ and $y = e_j$, we find that $a_{ji} = e_i^\top A^\top e_j = e_i^\top A e_j = a_{ij}$, so $A = A^\top$.

Theorem 29.5 (Spectral theorem)

If $T : \mathbb{R}^n \to \mathbb{R}^n$ is a symmetric operator or A is a symmetric $n \times n$ matrix of real numbers, then all eigenvalues are real and there is an orthonormal basis of eigenvectors.

Proof: STEP 1: T has at least one real eigenvalue and corresponding real eigenvector.

Writing T in its matrix representation and realizing that each real number is a special case of a complex number with zero imaginary part, we can pretend, just for a short while, that T is defined on a complex vector space and consequently has an eigenvalue/-vector by Theorem 29.1: there is a possibly complex scalar λ and vector $x \neq \mathbf{0}$ with $T(x) = \lambda x$. By definition of the inner product and symmetry,

$$0 \leq \langle T(x), T(x) \rangle = \langle \lambda x, T(x) \rangle = \lambda \langle x, T(x) \rangle = \lambda \langle T(x), x \rangle = \lambda \langle \lambda x, x \rangle = \lambda^2 \langle x, x \rangle = \lambda^2 \|x\|^2.$$

Dividing by $\|x\|^2 > 0$ shows that λ^2 is nonnegative and real, so eigenvalue λ is real as well (exc. B.3). And since $x \neq \mathbf{0}$, either its vector of real parts or its vector of imaginary parts is distinct from $\mathbf{0}$ and will be an eigenvector of operator T with the same eigenvalue. So we can always find a real eigenvalue and a corresponding real eigenvector.

STEP 2: All eigenvalues of T are real and there is an orthonormal basis of eigenvectors.

By induction on the dimension of vector space V. If $\dim(V) = 1$, step 1 gives a real eigenvalue; the corresponding real eigenvector is a basis. If $\dim(V) = n > 1$, we may assume for the induction step that the claim is true for symmetric operators on spaces of dimension $n-1$. By step 1, T has a real eigenvector x and eigenvalue λ: $Tx = \lambda x$. Let $W = \{x\}^\perp$ be its orthogonal complement. By Theorem 27.3, W is an $(n-1)$-dimensional subspace of V. And if $w \in W$, then $T(w) \in W$:

$$w \in W \quad \Longrightarrow \quad \langle w, x\rangle = 0 \quad \Longrightarrow \quad \langle T(w), x\rangle = \langle w, T(x)\rangle = \langle w, \lambda x\rangle = \lambda\langle w, x\rangle = 0 \quad \Longrightarrow \quad T(w) \in W.$$

So $T : W \to W$ is a symmetric operator on a space of dimension $n-1$. By the induction hypothesis it has an orthonormal basis of eigenvectors $\{x_1, \ldots, x_{n-1}\}$ and corresponding real eigenvalues. Since x is orthogonal to all of these vectors, it is linearly independent from them, so adding it to this collection gives an orthogonal basis of eigenvectors of the full n-dimensional space V. Dividing x by its length gives an orthonormal basis of eigenvectors. □

Symmetric matrices play a crucial role in ***quadratic forms***, functions $f : \mathbb{R}^n \to \mathbb{R}$ of the form $f(x) = x^\top Ax$ for some $n \times n$ matrix A. Even though it doesn't look like it at first sight, you can always rewrite such a quadratic form in terms of a symmetric matrix:

there is a symmetric matrix B such that $x^\top Ax = x^\top Bx$ for all $x \in \mathbb{R}^n$.

This is easy: you are asked to verify in Exercise 29.2 that we can choose $B = \frac{1}{2}A + \frac{1}{2}A^\top$.

Example 29.9 Writing out the product, you can check that the quadratic form $f : \mathbb{R}^3 \to \mathbb{R}$ with

$$f(x_1, x_2, x_3) = 4x_1^2 - 2x_2^2 + 3x_3^2 - 6x_1x_2 + 5x_1x_3 - x_2x_3 \tag{135}$$

can be written in different forms, for instance as $f(x) = x^\top Ax$ and $f(x) = x^\top Bx$ with

$$A = \begin{bmatrix} 4 & -6 & 2 \\ 0 & -2 & -1 \\ 3 & 0 & 3 \end{bmatrix} \qquad \text{or} \qquad B = \tfrac{1}{2}A + \tfrac{1}{2}A^\top = \begin{bmatrix} 4 & -3 & \frac{5}{2} \\ -3 & -2 & -\frac{1}{2} \\ \frac{5}{2} & -\frac{1}{2} & 3 \end{bmatrix},$$

where B is symmetric. ◁

To understand how you find such coefficient matrices, write out $x^\top Ax$ for an $n \times n$ matrix A:

$$\begin{aligned} x^\top Ax = {} & a_{11}x_1x_1 + a_{12}x_1x_2 + \cdots + a_{1n}x_1x_n \\ & + a_{21}x_2x_1 + a_{22}x_2x_2 + \cdots + a_{2n}x_2x_n \\ & + \cdots \\ & + a_{n1}x_nx_1 + a_{n2}x_nx_2 + \cdots + a_{nn}x_nx_n \\ = {} & \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}x_ix_j. \end{aligned}$$

In this expression, terms of the form $x_i^2 = x_ix_i$ occur only once, with coefficient a_{ii}. So the coefficient in front of x_i^2 goes into row i and column i. In (135), for instance, x_3^2 has coefficient 3, so $a_{33} = 3$. But terms of the form x_ix_j with $i \neq j$ occur twice, once with coefficient a_{ij} and once with coefficient a_{ji}. So we can pick them freely, as long as we make sure that the coefficient in front of x_ix_j is equal to $a_{ij} + a_{ji}$. In (135), x_1x_3 has coefficient 5, so we must have $a_{13} + a_{31} = 5$. We chose $a_{13} = 2$ and $a_{31} = 3$. Of course, if we aim for symmetry, a_{ij} and a_{ji} must be the same, namely half the coefficient of x_ix_j.

Definition 29.3 A symmetric $n \times n$ matrix A of real numbers (or the quadratic form $f(x) = x^\top Ax$) is:

- ☒ ***positive definite*** if for each $x \in \mathbb{R}^n$ with $x \neq \mathbf{0}$: $x^\top Ax > 0$,
- ☒ ***positive semidefinite*** if for each $x \in \mathbb{R}^n$: $x^\top Ax \geq 0$,
- ☒ ***negative definite*** if for each $x \in \mathbb{R}^n$ with $x \neq \mathbf{0}$: $x^\top Ax < 0$,
- ☒ ***negative semidefinite*** if for each $x \in \mathbb{R}^n$: $x^\top Ax \leq 0$,
- ☒ ***indefinite*** if there are x and y in $\mathbb{R}^n$ with $x^\top Ax > 0$ and $y^\top Ay < 0$.

The quadratic form in Example 29.9 is indefinite, since $f(1,0,0) > 0$, but $f(0,1,0) < 0$.

Theorem 29.6 ((Semi)definite matrices and their eigenvalues)

A symmetric $n \times n$ matrix A of real numbers is

(a) positive definite if and only if all its eigenvalues are positive;

(b) positive semidefinite if and only if all its eigenvalues are nonnegative;

(c) negative definite if and only if all its eigenvalues are negative;

(d) negative semidefinite if and only if all its eigenvalues are nonpositive;

(e) indefinite if and only if it has both positive and negative eigenvalues.

Proof: **(a)** Since A is symmetric, the Spectral Theorem says that there are n not necessarily distinct real eigenvalues $\lambda_1, \ldots, \lambda_n$ with associated eigenvectors $v_1, \ldots, v_n$ that form an orthonormal basis of $\mathbb{R}^n$. Each x in $\mathbb{R}^n$ is a linear combination

$$x = \alpha_1 v_1 + \cdots + \alpha_n v_n$$

of these basis vectors and we find that

$$\begin{aligned} x^\top Ax &= (\alpha_1 v_1 + \cdots + \alpha_n v_n)^\top A(\alpha_1 v_1 + \cdots + \alpha_n v_n) \\ &= (\alpha_1 v_1 + \cdots + \alpha_n v_n)^\top (\alpha_1 \lambda_1 v_1 + \cdots + \alpha_n \lambda_n v_n) && (v_i\text{'s are eigenvectors}) \\ &= \alpha_1^2 \lambda_1 + \cdots + \alpha_n^2 \lambda_n. && (v_i\text{'s are orthonormal}) \end{aligned}$$

This is positive for each $x \neq \mathbf{0}$ if and only if all eigenvalues λ_i are positive. All other cases are similar.□

Square matrices (or linear operators) that give rise to a basis of eigenvectors are also called ***diagonalizable***. Here is why; assume there is a basis $v_1, \ldots, v_n$ of $\mathbb{R}^n$ (or $\mathbb{C}^n$) consisting of eigenvectors of a square matrix A with corresponding eigenvalues $\lambda_1, \ldots, \lambda_n$:

$$Av_1 = \lambda_1 v_1, \quad Av_2 = \lambda_2 v_2, \quad \cdots \quad Av_n = \lambda_n v_n.$$

In matrix notation, $AP = PD$, where $P = [v_1 \cdots v_n]$ is the matrix with the n eigenvectors as its columns and

$$D = \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{bmatrix}$$

is the diagonal matrix with numbers $\lambda_1, \ldots, \lambda_n$ on its diagonal and zeroes everywhere else. Since the n eigenvectors are a basis of V, P is invertible and we can rearrange $AP = PD$ to

$$A = PDP^{-1}.$$

This makes computing the powers of the matrix A really easy: by induction we have for each $k \in \mathbb{N}$, $A^k = PD^kP^{-1}$, where D^k is just the diagonal matrix with entries λ_i^k on the diagonal.

29.3 Bases of generalized eigenvectors: Jordan's theorem

Example 29.8 shows that some matrices A have too few eigenvectors to make a basis. The main message in this subsection is that we can instead find a special basis of so-called *generalized eigenvectors*. And such a basis remains sufficiently simple to yield nice expressions for powers A^k in applications like difference or differential equations. What is a generalized eigenvector?

Definition 29.4 Given a square (real or complex) matrix A and an eigenvalue λ, a ***generalized eigenvector*** of A is a nonzero vector x with $(A-\lambda I)^k x = \mathbf{0}$ for some positive integer k.

Ordinary eigenvectors x satisfy $(A-\lambda I)x$, so each eigenvector is also a generalized eigenvector. For linear operators T, the definition is analogous (replace A with T).

Example 29.10 The following matrix is triangular, so its only eigenvalue $\lambda = 2$ is on the diagonal.

$$A = \begin{bmatrix} 2 & 0 \\ 1 & 2 \end{bmatrix}$$

Solving $(A-2I)x = \mathbf{0}$, each eigenvector is a multiple of $v_1 = (0,1)$: not enough to be a basis of our two-dimensional space. So we look for a generalized eigenvector to complete the basis. Solving $(A-2I)^2 x = \mathbf{0}$ is hardly a challenge, since $(A-2I)^2$ is the zero matrix and all x are a solution. We can pick any x that is linearly independent from eigenvector v_1 to find a basis of generalized eigenvectors. For instance, $v_2 = (1,0)$ will do. You can verify that this particular choice satisfies $(A-2I)v_2 = v_1$. ◁

If x is a generalized eigenvector of matrix A with associated eigenvalue λ, we typically get other generalized eigenvectors for free. Let k be the smallest positive integer such that $(A-\lambda I)^k x = \mathbf{0}$. Then all vectors

$$x, \quad (A-\lambda I)x, \quad (A-\lambda I)^2 x, \quad \cdots, \quad (A-\lambda I)^{k-1} x \tag{136}$$

are generalized eigenvectors, because they are distinct from $\mathbf{0}$ by definition of k and for each $\ell = 1, \ldots, k-1$, the vector $(A-\lambda I)^\ell x$ satisfies

$$(A-\lambda I)^{k-\ell}(A-\lambda I)^\ell x = (A-\lambda I)^k x = \mathbf{0}.$$

We refer to the vectors in (136) as a ***(Jordan) chain*** of generalized eigenvectors (generated by x). The rightmost element $(A-\lambda I)^{k-1}x$ of a chain is always an ordinary eigenvector of A, because it is distinct from $\mathbf{0}$ and multiplying it with $(A-\lambda I)$ gives $\mathbf{0}$.

In (136) we started from the left, using x to generate a chain of generalized eigenvectors whose rightmost term was an ordinary eigenvector. Equivalently, we could have started from the right: relabeling the vectors as

$$v_k = x, \quad v_{k-1} = (A-\lambda I)x, \quad v_{k-2} = (A-\lambda I)^2 x, \quad \cdots, \quad v_1 = (A-\lambda I)^{k-1} x,$$

we begin with an ordinary eigenvector v_1 and recursively find v_{i+1} by solving

$$(A-\lambda I)v_{i+1} = v_i.$$

That is exactly how we found the vector v_2 in Example 29.10. This recursive process cannot go on indefinitely: the vectors in a chain are linearly independent (see Exercise 29.3), so if we start out with an $n \times n$ matrix (i.e., we want a basis for n-dimensional space), there cannot be more than n of them.

You can always find such chains of generalized eigenvectors that constitute a basis:

Theorem 29.7 (Jordan's theorem for complex matrices)

Each $n \times n$ matrix A of complex numbers has chains of generalized eigenvectors that constitute a basis.

The proof — in the later subsection of postponed proofs — isn't pretty; feel free to skip it. In practice, first find the ordinary eigenvalues and eigenvectors of A. If A has exactly r linearly independent eigenvectors, start a chain from each of them, i.e., start from such an eigenvector v_1 with corresponding eigenvalue λ and as long as possible — assuming you found v_i — try to find the next term v_{i+1} by solving

$$(A - \lambda I)v_{i+1} = v_i. \tag{137}$$

Example 29.11 Using expansion into cofactors with respect to the third row or column, which contain a lot of zeros, it follows that the characteristic polynomial of

$$A = \begin{bmatrix} 2 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

is $\det(A - \lambda I) = (1 - \lambda)^3$. So it has a single eigenvalue $\lambda = 1$. To find the eigenvectors, solve

$$\mathbf{0} = (A - \lambda I)x = \begin{bmatrix} 1 & 1 & 0 \\ -1 & -1 & 0 \\ 0 & 0 & 0 \end{bmatrix} x.$$

We can choose x_3 and x_2 freely and take $x_1 = -x_2$: the eigenspace is two-dimensional and spanned by linearly independent eigenvectors $v = (0,0,1)$ and $w = (-1,1,0)$. Since we have three dimensions, we need one more generalized eigenvector to form a basis. So let's try to find the chains starting from these eigenvectors. The chain from $v_1 = v$ has an abrupt end: by (137) we need to solve

$$(A - \lambda I)v_2 = v_1 \qquad \Longleftrightarrow \qquad \begin{bmatrix} 1 & 1 & 0 \\ -1 & -1 & 0 \\ 0 & 0 & 0 \end{bmatrix} v_2 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

which, considering the third coordinate, is impossible. So this chain has only one vector, v. We're better off with $v_1 = w$ as our starting point: we need to solve

$$(A - \lambda I)v_2 = v_1 \qquad \Longleftrightarrow \qquad \begin{bmatrix} 1 & 1 & 0 \\ -1 & -1 & 0 \\ 0 & 0 & 0 \end{bmatrix} v_2 = \begin{bmatrix} -1 \\ 1 \\ 0 \end{bmatrix}$$

and can choose, for instance, $v_2 = (-1,0,0)$ (any other solution would be fine as well). This gives us the desired basis

$$\{(0,0,1), (-1,1,0), (-1,0,0)\}$$

of generalized eigenvectors, the first one being a chain all by itself, the latter two in a chain of length two. You can check that the latter chain goes no further: there is no solution v_3 to $(A - \lambda I)v_3 = v_2$. ◁

Now that you know how to find a basis using chains of generalized eigenvectors, solving time-homogeneous systems of linear difference equations becomes trivial. Also in cases where the coefficient matrix doesn't have a basis of eigenvectors, cases that are not treated with care in most economics textbooks. I will illustrate the procedure with a 2×2 example.

Example 29.12 (Solving linear difference equations) Consider the system of linear difference equations where for each time $t = 0, 1, 2, \ldots$:

$$x(t+1) = Ax(t) \text{ with } A = \begin{bmatrix} 3 & -1 \\ 1 & 1 \end{bmatrix} \text{ and given initial state } x(0) \in \mathbb{R}^2. \tag{138}$$

Matrix A has characteristic polynomial $\det(A - \lambda I) = (\lambda - 2)^2$, so its only eigenvalue is $\lambda = 2$ and each eigenvector is a multiple of $v_1 = (1, 1)$. In a chain of generalized eigenvectors our next vector v_2 must satisfy $(A - \lambda I)v_2 = v_1$. Solve this to see that we can take, for instance, $v_2 = (2, 1)$. So v_1 and v_2 are a basis of $\mathbb{R}^2$ and the initial state $x(0)$ can be written uniquely as a linear combination $x(0) = \alpha v_1 + \beta v_2$. At any later time t, repeated application of (138) gives

$$x(t) = A^t x(0) = \alpha A^t v_1 + \beta A^t v_2. \tag{139}$$

And for powers $A^t v_i$ of vectors in a chain of generalized eigenvectors there are easy expressions that you can verify by induction: using that $Av_1 = \lambda v_1$ and $Av_2 = \lambda v_2 + v_1$, it follows that

$$A^t v_1 = \lambda^t v_1 \qquad \text{and} \qquad A^t v_2 = \lambda^t v_2 + t\lambda^{t-1} v_1.$$

Substituting this into (139) gives the solution to our difference equation:

$$\text{for each } t = 0, 1, 2, \ldots: \qquad x(t) = \alpha \lambda^t v_1 + \beta\left(\lambda^t v_2 + t\lambda^{t-1} v_1\right). \qquad \triangleleft$$

The general case works exactly the same. To solve difference equation $x(t+1) = Ax(t)$ for an $n \times n$ matrix A and initial state $x(0)$, write the initial state as a linear combination of the vectors in a basis of chains of generalized eigenvectors. Then use that there are explicit formulas for $A^t v_i$ given the vectors v_i in chains to find the solution for $x(t)$. These explicit formulas are provided in Exercise 29.4.

To understand why they look the way they do, it is good practice to try to find them from scratch by taking some chain $v_1, v_2, \ldots, v_m$ of generalized eigenvectors given eigenvalue λ. Vector v_1 is an ordinary eigenvector:

$$Av_1 = \lambda v_1.$$

Use this to find expressions for $A^t v_1$ for some low values $t = 1, 2, 3, \ldots$. That should give you a reasonable conjecture for the general formula. Likewise, by (137) we have for each $i = 1, \ldots, m-1$ that

$$Av_{i+1} = \lambda v_{i+1} + v_i.$$

So moving on to v_2, use $Av_2 = \lambda v_2 + v_1$ together with $Av_1 = \lambda v_1$ to find expressions for $A^t v_2$ for some low values of t to formulate a conjecture for the general formula. And you can repeat this for v_3, v_4, and so on, although the expressions get more tedious along the way.

29.4 Postponed proofs

29.4.1 Proof of Theorem 29.1: existence of eigenvalues

Let T be a linear operator on an n-dimensional complex vector space $V \neq \{\mathbf{0}\}$. Let $v \in V$, $v \neq \mathbf{0}$. Since V is n-dimensional, the $n+1$ vectors $v, Tv, T^2 v, \ldots, T^n v$ are linearly dependent, so there are scalars $a_0, \ldots, a_n$, not all zero, such that

$$a_0 v + a_1 Tv + \cdots + a_n T^n v = (a_0 I + a_1 T + \cdots + a_n T^n)v = \mathbf{0}. \tag{140}$$

Use these scalars as coefficients of a complex polynomial. The fundamental theorem of algebra says that it can be factorized:

$$a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n = c(x - r_1) \cdots (x - r_m).$$

Here $c \neq 0$ and the roots of the polynomial $r_1, \ldots, r_m$ are all complex. Number m is the degree of the polynomial, which may be less than n if $a_n = 0$. In combination with (140),

$$\mathbf{0} = (a_0 I + a_1 T + \cdots + a_n T^n) v = c(T - r_1 I) \cdots (T - r_m I) v.$$

So for at least one r_j, $T - r_j I$ is not injective, making r_j an eigenvalue of T.

29.4.2 Proof of Theorem 29.3: Perron-Frobenius

OBSERVATION: If A is an $n \times n$ matrix of strictly positive real numbers and v and w are nonnegative vectors with $v \geq w$, but $v \neq w$, then $Av > Aw$.

Proof: Since $Av = A(v - w + w) = A(v - w) + Aw$, we simply show that each coordinate of $A(v - w)$ is positive. By assumption, $v - w$ has at least one coordinate larger than zero, all others are nonnegative. For $i \in \{1, \ldots, n\}$, the i-th coordinate of $A(v - w)$ is the inner product of the positive i-th row of A and the nonnegative vector $v - w$ with at least one positive coordinate, i.e., a positive number. □

This leaves us properly prepared for the remainder of the proof. Look at all real numbers λ such that $Ax \geq \lambda x$ for some nonnegative vector x distinct from $\mathbf{0}$. The problem of finding the largest such λ over all these combinations (x, λ) has a solution (x^*, λ^*) with $\lambda^* > 0$. The argument isn't really linear algebra, so I defer it to Exercise 29.6.

In this solution, equality must hold: $Ax^* = \lambda^* x^*$. If, to the contrary, $Ax^* \geq \lambda^* x^*$ is not an equality, multiply both sides by A. Our observation with $v = Ax^*$ and $w = \lambda^* x^*$ gives a strict inequality $AAx^* > \lambda^* Ax^*$. So the positive vector $v = Ax^*$ satisfies $Av > \lambda^* v$. But then we could have increased λ^* a bit, contradicting that it is the largest one.

So $Ax^* = \lambda^* x^*$, making λ^* a positive eigenvalue and x^* a real eigenvector where all coordinates are positive (not just nonnegative) since the left side Ax^* of the equation is a positive vector.

To see that no other eigenvalue λ is larger than λ^* in absolute value, pick a corresponding eigenvector: $Ay = \lambda y$. Denoting, as usual, the entries of matrix A by a_{ij}, this equation says that for each coordinate $i \in \{1, \ldots, n\}$,

$$\lambda y_i = \sum_{j=1}^{n} a_{ij} y_j.$$

Taking absolute values and applying the triangle inequality gives

$$|\lambda| \, |y_i| = |\lambda y_i| = \left| \sum_{j=1}^{n} a_{ij} y_j \right| \leq \sum_{j=1}^{n} a_{ij} \, |y_i|.$$

If we define the vector $z = \left(|y_1|, \ldots, |y_n|\right)$, this means $Az \geq |\lambda| z$, so the pair $(z, |\lambda|)$ is feasible in the maximization problem we started with. Since (x^*, λ^*) was optimal, it follows that $|\lambda| \leq \lambda^*$.

29.4.3 Proof of Theorem 29.7: Jordan's theorem for matrices

We will do the proof for arbitrary linear operators T (replacing matrix A with operator T). Remember from (137) that we could generate chains starting with an eigenvector v_1 and eigenvalue λ and recursively picking

$$v_2 \text{ with } (T - \lambda I) v_2 = v_1, \qquad v_3 \text{ with } (T - \lambda I) v_3 = v_2,$$

and so on. It is sometimes convenient to represent such a chain graphically:

$$v_3 \xrightarrow{T-\lambda I} v_2 \xrightarrow{T-\lambda I} v_1 \xrightarrow{T-\lambda I} \mathbf{0},$$

where $w \xrightarrow{T-\lambda I} v$ means that applying $T - \lambda I$ to w gives v. So let us now prove:

Theorem 29.8 (Jordan's theorem for linear operators)

Each linear operator T on a complex vector space V of finite dimension $\dim(V) \geq 1$ has chains of generalized eigenvectors that constitute a basis of V.

Proof: If $\dim(V) = 1$, any nonzero $v \in V$ gives a basis. Now let $n \in \mathbb{N}, n \geq 2$, and assume the claim is true for linear operators on vector spaces of lower dimension than n.

T has an eigenvalue λ and corresponding eigenvector v. Since $v \neq \mathbf{0}$, the dimension r of $\ker(T-\lambda I)$ is larger than zero. By the rank-nullity theorem, the dimension of the range of $T-\lambda I$ is less than n. Call this range W.

STEP 1: $(T-\lambda I)(W) \subseteq (T-\lambda I)(V) = W$, so $T-\lambda I : W \to W$ is a well-defined linear operator. Since $\dim(W) = n - r < n$, the induction step assures that there are chains of generalized eigenvectors of $T-\lambda I$ and hence of T whose vectors $w_{i,j}$ are a basis of W.[9] Here we used that if μ is an eigenvalue of $T-\lambda I$, then $\mu+\lambda$ is an eigenvalue of T and the corresponding eigenvectors are the same.

$$\begin{array}{ccccccc} w_{1,n_1} & \xrightarrow{T-\lambda_1 I} & \cdots & \xrightarrow{T-\lambda_1 I} & w_{1,1} & \xrightarrow{T-\lambda_1 I} & \mathbf{0} \\ & & \vdots & & & & \\ w_{k,n_k} & \xrightarrow{T-\lambda_k I} & \cdots & \xrightarrow{T-\lambda_k I} & w_{k,1} & \xrightarrow{T-\lambda_k I} & \mathbf{0} \end{array}$$

The $w_{i,j}$'s are a basis of W.

STEP 2: Let q be the dimension of $W \cap \ker(T-\lambda I)$, the eigenspace of $T-\lambda I$ in W. So q chains in step 1 arose from this eigenvalue. Why? The $w_{i,j}$ span this eigenspace, so a subset of q of them is a basis. These must be q eigenvectors, so they lie on the rightmost side of q chains (the other $w_{i,j}$ are only *generalized* eigenvectors). We assume these are the first q chains: $\lambda_1 = \cdots = \lambda_q = \lambda$. At the other side of these chains we find w_{j,n_j} in W, the range of $T-\lambda I$. So there are y_j with

$$y_j \xrightarrow{T-\lambda I} w_{j,n_j} \qquad \text{for all } j = 1,\ldots,q.$$

STEP 3: Since $\ker(T-\lambda I)$ has dimension r and meets W in a q-dimensional subspace, some $(r-q)$-dimensional subspace Z of $\ker(T-\lambda I)$ meets W only at $\mathbf{0}$. Let $z_1,\ldots,z_{r-q}$ be a basis of Z. Together with the previous steps, this gives $q+(n-r)+(r-q) = n$ vectors in V in chains

$$\begin{array}{ccccccccc} y_1 & \xrightarrow{T-\lambda I} & w_{1,n_1} & \xrightarrow{T-\lambda I} & \cdots & \xrightarrow{T-\lambda I} & w_{1,1} & \xrightarrow{T-\lambda I} & \mathbf{0} \\ \vdots & & & & & & & & \\ y_q & \xrightarrow{T-\lambda I} & w_{q,n_q} & \xrightarrow{T-\lambda I} & \cdots & \xrightarrow{T-\lambda I} & w_{q,1} & \xrightarrow{T-\lambda I} & \mathbf{0} \\ & & w_{q+1,n_{q+1}} & \xrightarrow{T-\lambda_{q+1} I} & \cdots & \xrightarrow{T-\lambda_{q+1} I} & w_{q+1,1} & \xrightarrow{T-\lambda_{q+1} I} & \mathbf{0} \\ & & & & \vdots & & & & \\ & & w_{k,n_k} & \xrightarrow{T-\lambda_k I} & \cdots & \xrightarrow{T-\lambda_k I} & w_{q+1,1} & \xrightarrow{T-\lambda_k I} & \mathbf{0} \\ & & & & & & z_1 & \xrightarrow{T-\lambda I} & \mathbf{0} \\ & & & & & & \vdots & & \\ & & & & & & z_{r-q} & \xrightarrow{T-\lambda I} & \mathbf{0} \end{array}$$

[9] Vector $w_{i,j}$ is the j-th vector in the i-th chain; the length of that chain is denoted by n_i.

To show that they are a basis, it suffices to verify they are linearly independent. So consider scalars with

$$\sum a_i y_i + \sum_{i,j} b_{i,j} w_{i,j} + \sum c_i z_i = \mathbf{0}.$$

Apply $T - \lambda I$ to both sides of the equation. We see from the chains above that this gives a linear combination of only the $w_{i,j}$'s and that w_{i,n_i} gets coefficient a_i. By linear independence of the $w_{i,j}$, $a_i = 0$ for all i. So

$$\sum_{i,j} b_{i,j} w_{i,j} + \sum c_i z_i = \mathbf{0}.$$

But by construction Z and W only have $\mathbf{0}$ in common. So $\sum b_{i,j} w_{i,j} = \mathbf{0}$ and $\sum c_i z_i = \mathbf{0}$. By linear independence of the $w_{i,j}$ and the z_i, respectively: $b_{i,j} = 0$ and $c_i = 0$ for all i and j. This establishes linear independence. □

Exercises section 29

29.1 Let A be a square matrix of real or complex numbers. Show that A and its transpose have the same eigenvalues. What about their eigenvectors?

29.2 Consider a quadratic form $x \mapsto x^\top Ax$ for some $n \times n$ matrix A. Show that $B = \frac{1}{2}A + \frac{1}{2}A^\top$ is symmetric and satisfies $x^\top Ax = x^\top Bx$ for all $x \in \mathbb{R}^n$.

29.3 Show that the vectors in a chain of generalized eigenvectors are linearly independent.

29.4 Let square matrix A have eigenvalue λ and corresponding eigenvector v_1:

$$Av_1 = \lambda v_1.$$

Use v_1 to generate a chain of generalized eigenvectors $v_1, v_2, \ldots, v_m$: for each $i = 1, \ldots, m-1$,

$$(A - \lambda I)v_{i+1} = v_i \qquad \text{or, equivalently,} \qquad Av_{i+1} = \lambda v_{i+1} + v_i.$$

(a) Prove by induction that for all $t = 1, 2, 3, \ldots$:

$$A^t v_1 = \lambda^t v_1.$$

(b) Prove by induction that for all $t = 1, 2, 3, \ldots$:

$$A^t v_2 = \lambda^t v_2 + t\lambda^{t-1} v_1.$$

(c) Finally, look at the i-th vector v_i for an arbitrary $i \in \{1, \ldots, m\}$ and prove that for all $t = 1, 2, 3, \ldots$:

$$A^t v_i = \sum_{k=0}^{i-1} \binom{t}{k} \lambda^{t-k} v_{i-k} = \binom{t}{0}\lambda^t v_i + \binom{t}{1}\lambda^{t-1} v_{i-1} + \binom{t}{2}\lambda^{t-2} v_{i-2} + \cdots + \binom{t}{i-1}\lambda^{t-i+1} v_1.$$

REMEMBER: for nonnegative integers k, ℓ, the binomial coefficient $\binom{k}{\ell}$ is defined as

$$\binom{k}{\ell} = \begin{cases} \frac{k!}{\ell!(k-\ell)!} & \text{if } k \geq \ell, \\ 0 & \text{otherwise,} \end{cases}$$

where $0! = 1$ and for positive integers m,

$$m! = m(m-1)(m-2)\cdots 2 \cdot 1.$$

29.5 What is the solution to the difference equation $x(t+1) = Ax(t)$ for all $t = 0, 1, 2, \ldots$,

(a) For the matrix A from Example 29.10 and initial state $x(0) = (-5, 2)$?

(b) For the matrix A from Example 29.11 and initial state $x(0) = (1, -2, 3)$?

29.6 Fix a square matrix A with positive entries. Our proof of the Perron-Frobenius theorem uses that the problem of maximizing λ over all candidates (x,λ) where x is a nonnegative vector distinct from $\mathbf{0}$ and $Ax \geq \lambda x$ has a solution. We show that this is true by rewriting it to a problem where the Extreme Value Theorem applies.

(a) Show that there is a feasible (x,λ) with $\lambda > 0$. HINT: Take $x = (1,\ldots,1)$.

This assures that *if* there is a solution, λ is larger than zero.

(b) Show that if (x,λ) is feasible, then so is $(\alpha x,\lambda)$ for each number $\alpha > 0$.

Hence we can normalize the coordinates of x so that they sum to one: take $\alpha = \frac{1}{x_1+\cdots+x_n}$. This means that we can restrict our vectors x to lie in the unit simplex

$$\Delta_n = \left\{x \in \mathbb{R}^n_+ : x_1 + \cdots + x_n = 1\right\}.$$

Denote the largest entry a_{ij} in the positive matrix A by $a > 0$.

(c) Show that if (x,λ) with $x \in \Delta_n$ is feasible, then $\lambda \leq na$.

So the problem reduces to maximizing the continuous function $(x,\lambda) \mapsto \lambda$ with x in the compact set Δ_n and λ in the compact set $[0,na]$. By the Extreme Value Theorem, a maximum (x^*,λ^*) exists. By our first step, λ^* is positive.

A A reminder of common notation and terminology

A.1 Fields

You are, of course, familiar with four operations — addition, subtraction, multiplication, and division — on the set of real numbers. In mathematics, a set with these operations is referred to as a field. Its formal definition is:

Definition A.1 A ***field*** is a set F on which two operations + (addition) and $\cdot$ (multiplication) are defined such that

- ⊠ for each pair of elements $x, y \in F$ there is a unique element $x + y$, the sum of x and y, in F,
- ⊠ for each pair of elements $x, y \in F$, there is a unique element $x \cdot y$, the product of x and y, in F.

Referring to these two properties, it is sometimes said that F is "closed under addition" and "closed under multiplication", respectively. Moreover, the following conditions must hold, for all elements $x, y, z \in F$:

(F1) Commutativity of addition and multiplication: $x + y = y + x$ and $x \cdot y = y \cdot x$.

(F2) Associativity of addition and multiplication: $(x + y) + z = x + (y + z)$ and $(x \cdot y) \cdot z = x \cdot (y \cdot z)$.

(F3) Existence of identity elements for addition and multiplication: there are distinct elements 0 and 1 in F such that $x + 0 = x$ and $1 \cdot x = x$.

(F4) Existence of inverses for addition and multiplication: for each $x \in F$ there is a $y \in F$ with $x + y = 0$ and for each nonzero element $x \in F$ there is a $y \in F$ such that $x \cdot y = 1$.

(F5) Distributivity of multiplication over addition: $x \cdot (y + z) = x \cdot y + x \cdot z$.

For notational convenience, the product $x \cdot y$ of x and y is often written simply as xy.

So far, only addition and multiplication are defined. Subtraction and division are defined in terms of their inverses. By (F4), for each $x \in F$ there is an element $y \in F$ such that $x + y = 0$. It can be shown (along the lines of the proof of Theorem 3.1) that this element y is unique. It is denoted as $-x$. Subtraction is now defined as addition of the additive inverse:

$$x - y = x + (-y).$$

Similarly, by (F4), each nonzero (division by zero is not allowed!) $x \in F$, has a unique $y \in F$ with $xy = 1$. We denote this y by x^{-1} and define division as multiplication with the multiplicative inverse:

$$x/y = xy^{-1}.$$

Example A.1 The set $\mathbb{N} = \{1, 2, 3, \ldots\}$ of ***natural numbers*** or ***positive integers*** with the usual addition and multiplication is *not* a field: properties (F3) and (F4) do not hold: there is no zero element in $\mathbb{N}$ and there are no additive and multiplicative inverses. For instance, there is no $y \in \mathbb{N}$ such that $2 \cdot y = 1$. ◁

Example A.2 The set $\mathbb{Z} = \{\ldots, -3, -2, -1, 0, 1, 2, 3, \ldots\}$ of ***integers*** with the usual addition and multiplication is *not* a field: again, property (F4) does not hold. ◁

Example A.3 The set $\mathbb{Q} = \{p/q : p, q \in \mathbb{Z}, q \neq 0\}$ of ***rational numbers*** with the usual addition and multiplication is a field. ◁

Example A.4 The set $\mathbb{R}$ of ***real numbers*** with the usual addition and multiplication is a field.

The real numbers can be constructed as a completion of the rational numbers in such a way that a sequence defined by a decimal expansion like $(3, 3.1, 3.14, 3.141, 3.1415, \ldots)$ converges to a unique real number. ◁

Example A.5 The set $\mathbb{C}$ of ***complex numbers*** consists of all numbers of the form $a+bi$, where $a,b\in\mathbb{R}$, and $i^2=-1$. Equality in $\mathbb{C}$ is defined by $a+bi=c+di$ if and only if $a=c$ and $b=d$. This set is a field if addition is defined by

$$(a+bi)+(c+di)=(a+c)+(b+d)i$$

and multiplication is defined by

$$(a+bi)(c+di)=(ac-bd)+(ad+bc)i.$$ ◁

A.2 Sets

Sets are typically denoted by capital letters, elements by lower-case letters. Sets are sometimes indicated by curly braces { and } in one of the following ways:

- ☒ by explicitly listing its elements: $S=\{1,2,3,4,5,6\}$ or $S=\{1,\ldots,6\}$, where '…' is used if it ought to be clear from the context what the other elements are;
- ☒ by some characterizing property P, like $S=\{x : x \text{ satisfies } P\}$. For instance,

$$S=\{x : x \text{ is a positive integer}, x\leq 6\}.$$

We write $x\in S$ to denote that x belongs to/is an element of set S and $x\notin S$ to denote that x does not belong to S. Other standard notation we will use:

notation	name	characterizing property
$\emptyset$	the empty set	the set that contains no elements
$A\subseteq B$	A is a subset of B	each element of A belongs to B : $a\in A\Rightarrow a\in B$
$A=B$	sets A and B are equal	$A\subseteq B$ and $B\subseteq A$
$A\subset B$	A is a proper subset of B	$A\subseteq B$, but $A\neq B$
$A\cap B$	intersection of A and B	elements of both A and B: $A\cap B=\{x : x\in A, x\in B\}$
$A\cup B$	union of A and B	elements of A or B (or both): $A\cup B=\{x : x\in A \text{ or } x\in B\}$
$A\setminus B$		elements of A, but not of B: $A\setminus B=\{x : x\in A, x\notin B\}$

Some of these concepts are illustrated using the Venn diagrams below:

If A is a subset of some larger set X, we denote by A^c the complement of A w.r.t. X:

$$A^c=\{x\in X : x\notin A\}=X\setminus A.$$

The notation $X\setminus A$ explicitly tells with respect to which set the complement is taken; this is supposed to be clear from the context if we write A^c. If one considers the union or intersection of more than two sets, it is often convenient to use an index set. For instance, the union of three sets A_1, A_2, A_3 can be denoted as

$$A_1\cup A_2\cup A_3=\cup_{i=1}^3 A_i=\cup_{i\in\{1,2,3\}}A_i=\{x : x\in A_1 \text{ or } x\in A_2 \text{ or } x\in A_3\}$$

and their intersection as

$$A_1\cap A_2\cap A_3=\cap_{i=1}^3 A_i=\cap_{i\in\{1,2,3\}}A_i=\{x : x\in A_1 \text{ and } x\in A_2 \text{ and } x\in A_3\}.$$

Generally, suppose that for each index $i \in I$ from an index set I, you have defined a set A_i. Then their union is denoted by

$$\cup_{i\in I} A_i = \{x : x \in A_i \text{ for some } i \in I\}$$

and their intersection by

$$\cap_{i\in I} A_i = \{x : x \in A_i \text{ for all } i \in I\}.$$

If the index set is clear from the context, this is often abbreviated as $\cup_i A_i$ and $\cap_i A_i$.

It is easy to verify ***De Morgan's Laws***:

- ☒ the complement of a union of sets is the intersection of their complements: $(\cup_{i\in I} A_i)^c = \cap_{i\in I} A_i^c$.
- ☒ the complement of an intersection of sets is the union of their complements: $(\cap_{i\in I} A_i)^c = \cup_{i\in I} A_i^c$.

In measure theory and topology we often consider sets whose elements are also sets. For instance, the set of all subsets of $\{0,1\}$ is

$$\{\emptyset, \{0\}, \{1\}, \{0,1\}\}.$$

In such and related cases, it is common to speak of a collection (or family) of sets, rather than a set of sets.

We use the following common notation for sets of numbers:

notation	is the set of
$\mathbb{N}$	positive integers: 1, 2, 3, …
$\mathbb{Z}$	integers: …, -2, -1, 0, 1, 2, …
$\mathbb{Q}$	rational numbers: p/q with $p, q \in \mathbb{Z}, q \neq 0$
$\mathbb{R}$	real numbers
$\mathbb{R}_+$	nonnegative real numbers: $[0,\infty)$
$\mathbb{C}$	complex numbers

The important property that distinguishes real from rational numbers is the 'least upper bound property': every nonempty set of real numbers that is bounded from above has a smallest upper bound, its ***supremum***. Similarly, every nonempty set of real numbers that is bounded from below has a greatest lower bound, its ***infimum***. If the supremum and infimum belong to the sets under consideration, they are referred to as the set's maximum and minimum, respectively. For instance, the set $(0,1]$ has infimum 0 and supremum 1. Since $0 \notin (0,1]$, the set has no minimum. Since $1 \in (0,1]$, this is the set's maximum.

A.3 Functions

A ***function*** f from a set X to a set Y, denoted $f : X \to Y$, assigns to each ***argument*** $x \in X$ a single ***image*** or ***function value*** $f(x) \in Y$. Functions are also called maps, mappings, or transformations. The set X is the ***domain*** of the function, the set Y is its ***codomain***. The codomain should not be confused with the ***range*** of f, which is the set $\{f(x) : x \in X\}$ of all function values actually attained by the function.

Example A.6 The function $f : \mathbb{R} \to \mathbb{R}$ with $f(x) = x^2$ has the set of real numbers as its domain and codomain, but the square of a real number is nonnegative and the range of f is $[0,\infty)$. ◁

The ***graph*** of $f : X \to Y$ is the set $\{(x, f(x)) : x \in X\} \subseteq X \times Y$. If X and Y are sets of real numbers and f is sufficiently simple, it is common to draw the graph of f with the arguments x on the horizontal and values $y = f(x)$ on the vertical axis.

The symbol $\mapsto$ means 'maps to'. It is used to save on notation if the name f of a function is not important, but we only care how it transforms its arguments into function values: the function with $f(x) = x^2$ from our earlier example can be written as $x \mapsto x^2$.

A.4 Injective, surjective, and bijective functions

A function $f : X \to Y$ is:

- ☒ ***injective*** (or *one-to-one* or an *injection*) if different arguments give different function values: for all x and x' in X, if $x \neq x'$, then $f(x) \neq f(x')$;
- ☒ ***surjective*** (or *onto* or a *surjection*) if each element of Y is the image of some element of X: for each $y \in Y$ there is an $x \in X$ with $f(x) = y$;
- ☒ ***bijective*** (or *invertible* or a *bijection*) if it is both injective and surjective.

In other words, if a function is injective, each element of Y is the image of *at most* one element of X. And if it is surjective, each element of Y is the image of *at least* one element of X: the function's codomain and range coincide. So if it is bijective, each element of Y is the image of *exactly one* element of X; we can then define its ***inverse*** $f^{-1} : Y \to X$ that assigns to each $y \in Y$ the unique element $x \in X$ with $f(x) = y$. That is,

$$f^{-1}(y) = x \qquad \Longleftrightarrow \qquad f(x) = y.$$

Example A.7 Here are four functions from and to the real numbers. Sketching their graphs helps you see that:

- ☒ $x \mapsto 2x - 2$ is both injective and surjective; its inverse, obtained by solving $y = 2x - 2$ for x, is the function $y \mapsto \frac{1}{2}y + 1$;
- ☒ the strictly increasing function $x \mapsto e^x$ is injective, but not surjective;
- ☒ $x \mapsto x(x-1)(x-2)$ is surjective but not injective;
- ☒ the constant function $x \mapsto 1$ is neither injective nor surjective. ◁

A.5 Decimal representations

Recall from high school that each real number has a decimal representation, like

$$\pi = 3.1415926535\cdots$$

In general, each real number can be written as

$$z.d_1 d_2 d_3 d_4 \cdots = z + \sum_{i=1}^{\infty} d_i \cdot 10^{-i}$$

where z is an integer and $d_i \in \{0, \ldots, 9\}$ is the digit in the i-th place after the decimal point.

The proof below settles a pesky detail. Some real numbers have more than one — indeed, exactly two — decimal representations. These representations coincide up to some point, after which one continues with $a9999\cdots$ (an infinite string of nines) and the other with $b0000\cdots$ (an infinite string of zeros), where digit b is one unit higher than digit a. For example,

$$\frac{1}{4} = 0.250000\cdots = 0.249999\cdots \qquad \text{and} \qquad \frac{37}{100} = 0.370000\cdots = 0.369999\cdots$$

Adding or subtracting an integer, it suffices to show this for numbers between zero and one.

Theorem A.1

Each real number x with $0 \leq x < 1$ has at most two decimal representations.

Proof: Since $1 = 0.999\cdots$, multiplying by 10^{-n} shows that for all $n \in \mathbb{N}$,

$$10^{-n} = \sum_{i=n+1}^{\infty} 9 \cdot 10^{-i} \tag{141}$$

Suppose $x \in [0,1)$ can be written as $x = \sum_{i=1}^{\infty} a_i \cdot 10^{-i} = \sum_{i=1}^{\infty} b_i \cdot 10^{-i}$ for digits a_i, b_i in $\{0,\ldots,9\}$. Let n be the first term, if any, where $a_n \neq b_n$. Without loss of generality, $a_n < b_n$. Then

$$\begin{aligned}
x &= \sum_{i<n} a_i \cdot 10^{-i} + a_n \cdot 10^{-n} + \sum_{i>n} a_i \cdot 10^{-i} \\
&\leq \sum_{i<n} a_i \cdot 10^{-i} + a_n \cdot 10^{-n} + \sum_{i>n} 9 \cdot 10^{-i} && \text{(replacing all } a_i \text{ with } i > n \text{ by 9)} \\
&= \sum_{i<n} a_i \cdot 10^{-i} + (a_n + 1) \cdot 10^{-n} + 0 && \text{(since } \textstyle\sum_{i>n} 9 \cdot 10^{-i} = 10^{-n} \text{ by (141))} \\
&\leq \sum_{i<n} a_i \cdot 10^{-i} + b_n \cdot 10^{-n} + 0 && \text{(since } a_n < b_n\text{)} \\
&\leq \sum_{i<n} a_i \cdot 10^{-i} + b_n \cdot 10^{-n} + \sum_{i>n} b_i \cdot 10^{-i} && \text{(adding the decimals } b_i \text{ for } i > n\text{)} \\
&= \sum_{i<n} b_i \cdot 10^{-i} + b_n \cdot 10^{-n} + \sum_{i>n} b_i \cdot 10^{-i} && \text{(since } a_i = b_i \text{ if } i < n\text{)} \\
&= x.
\end{aligned}$$

The first inequality would be strict if $a_i \neq 9$ for any $i > n$, the second if $a_n + 1 \neq b_n$, and the third if $b_i \neq 0$ for any $i > n$. So these three things must be ruled out: there are at most two decimal representations for x where the first distinct digits differ by one unit, the smaller one is followed by a string of nines and the larger by a string of zeros. □

A.6 The Mean Value Theorem

In its simplest form, the Mean Value Theorem says that for a differentiable function $f : [a,b] \to \mathbb{R}$, there is a point c somewhere in the interval (a,b) with slope $f'(c)$ equal to the average/mean change $(f(b) - f(a))/(b-a)$ in the function value over this interval:

Theorem A.2 (Mean Value Theorem I)

If $f : [a,b] \to \mathbb{R}$ is continuous on $[a,b] \subseteq \mathbb{R}$ and differentiable on (a,b), then there is a point $c \in (a,b)$ with

$$f'(c) = \frac{f(b) - f(a)}{b-a}.$$

Proof: The function $h : [a,b] \to \mathbb{R}$ with $h(x) = f(x) - \frac{f(b)-f(a)}{b-a}x$ is continuous on $[a,b]$ and differentiable on (a,b), because f is. Moreover, $h(a) = h(b) = f(b) - f(a)$. If h is constant on $[a,b]$, its derivative is zero: $h'(c) = f'(c) - \frac{f(b)-f(a)}{b-a} = 0$ for each $c \in (a,b)$. If h is not constant, it has an extremum in the interior (a,b). The first order condition at such an internal extremum c is that $h'(c) = f'(c) - \frac{f(b)-f(a)}{b-a} = 0$. □

B Complex numbers

B.1 Why do we need them?

Already in high school you learned that a quadratic polynomial in x like

$$ax^2 + bx + c$$

with $a, b, c \in \mathbb{R}$ and $a \neq 0$ (otherwise it isn't quadratic) has roots that must be of the form

$$x = \frac{-b - \sqrt{b^2 - 4ac}}{2a} \qquad \text{and} \qquad x = \frac{-b + \sqrt{b^2 - 4ac}}{2a}. \tag{142}$$

Well, as long as the discriminant $b^2 - 4ac$ is nonnegative: *negative* real numbers don't have a square root in $\mathbb{R}$. For polynomial $\frac{1}{2}x^2 + 3x + 5$, for instance, mindless substitution into (142) would suggest roots $x = -3 - \sqrt{-1}$ and $x = -3 + \sqrt{-1}$, but that makes no sense: there is no real number whose square root is -1. So this quadratic polynomial has no real roots: if you draw its graph, it lies entirely above the x-axis.

Complex numbers were introduced in algebra precisely to overcome this problem and make sure that *every* quadratic polynomial or, for that matter, *any* nonconstant polynomial

$$a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n$$

has a root. But that requires us to make sense of 'numbers' like $-3 - \sqrt{-1}$ and $-3 + \sqrt{-1}$ above or, more generally, those of the form $a + b\sqrt{-1}$ with a and b in $\mathbb{R}$.

How? Some brute force is a good way to get a grip on this. Denote $\sqrt{-1}$ by i and call

$$z = a + b\sqrt{-1} = a + bi$$

with $a, b \in \mathbb{R}$ a ***complex number*** with ***real part*** $\mathrm{Re}(z) = a$ and ***imaginary part*** $\mathrm{Im}(z) = b$. To keep the mental strain to a minimum, it would be really nice if the two fundamental operations on numbers, addition and multiplication, simply boil down to expanding brackets. For instance, we would like the sum of complex numbers $a + bi$ and $c + di$ to be the number

$$(a + bi) + (c + di) = (a + c) + (b + d)i. \tag{143}$$

Analogously, we would like their product simply to be

$$(a + bi)(c + di) = ac + adi + bci + bdi^2$$

and remembering that i denotes $\sqrt{-1}$ we replace i^2 with -1 and simplify this expression to

$$(a + bi)(c + di) = (ac - bd) + (ad + bc)i. \tag{144}$$

This achieves exactly what we want. Firstly, the set $\mathbb{C}$ of complex numbers with addition and multiplication defined as in (143) and (144) satisfies all standard arithmetic rules: it is a *field*; see Appendix A.1. We often write $a + 0i$ as a and $0 + bi$ as bi. Under this convention, each real number a is a complex number $a + 0i$: the set of reals corresponds with the subset of complex numbers whose imaginary part is zero. If $z = a + bi$, its additive inverse is $-z = -a - bi$. Moreover, if $z = a + bi \neq 0$, its multiplicative inverse is a complex number w with $zw = 1$. Can we find it? Let's try $w = a - bi$. Since $(a + bi)(a - bi) = a^2 + b^2$, we're close: the righthand side should be 1, not $a^2 + b^2$. So just divide by $a^2 + b^2$ to find that

$$\frac{1}{z} = \frac{a - bi}{a^2 + b^2} = \frac{a}{a^2 + b^2} - \frac{b}{a^2 + b^2}\, i.$$

Example B.1 If $v = -2+3i$ and $w = 7-2i$, compute: (a) $v+w$, (b) $v-w$, (c) vw, (d) v/w.
SOLUTION:

(a) $v+w = (-2+3i)+(7-2i) = 5+i$

(b) $v-w = (-2+3i)-(7-2i) = -9+5i$

(c) $vw = (-2+3i)(7-2i) = -14+4i+21i-6i^2 = -8+25i$.

(d) $\frac{v}{w} = \frac{-2+3i}{7-2i} = \frac{-2+3i}{7-2i}\cdot\frac{7+2i}{7+2i} = \frac{-14-4i+21i+6i^2}{49+4} = -\frac{20}{53}+\frac{17}{53}i$. ◁

Secondly, and more importantly, it delivers what we set out to do: each polynomial $p : \mathbb{C} \to \mathbb{C}$ defined as

$$p(x) = a_0 + a_1 x + \cdots + a_n x^n$$

with complex coefficients has a root. Its proof is in subsection B.4.

Theorem B.1 (Fundamental theorem of algebra)

Every polynomial $p : \mathbb{C} \to \mathbb{C}$ with degree $n \geq 1$ has a complex root: $p(z) = 0$ for some $z \in \mathbb{C}$.

If z_1 is a root of the n-th degree polynomial p above, we can factor out the term $x - z_1$ and write

$$p(x) = (x - z_1)q(x)$$

where q is a polynomial of degree only $n-1$. As long as $n-1 \geq 1$, this polynomial q has a root as well. So repeatedly applying the fundamental theorem of algebra, the polynomial p can be ***factorized***: there are complex numbers $z_1, \ldots, z_n$ and $c \neq 0$ such that

$$p(x) = a_0 + a_1 x + \cdots + a_n x^n = c(x - z_1)(x - z_2)\cdots(x - z_n).$$

So each n-th degree polynomial over the complex numbers has n not necessarily distinct roots $z_1, \ldots, z_n$.

B.2 Polar coordinates

Complex number $z = a + bi$ is characterized by the pair of real numbers (a, b). Drawing its real part a on the horizontal and its imaginary part b on the vertical axis we can represent complex numbers graphically as points in the ***complex plane***. Draw a line piece between the origin and such an arbitrary point (a, b). See Figure 2. This point is completely described by (1) how far away it is from the origin, i.e., the length r of the line piece, and (2) what angle φ, by convention in radians, the line piece has with the horizontal axis. That is the idea behind the representation of complex numbers using ***polar coordinates***. Given r and φ we see that $\cos\varphi = a/r$ and $\sin\varphi = b/r$. Therefore,

$$z = a + bi = r\left(\cos\varphi + i\sin\varphi\right).$$

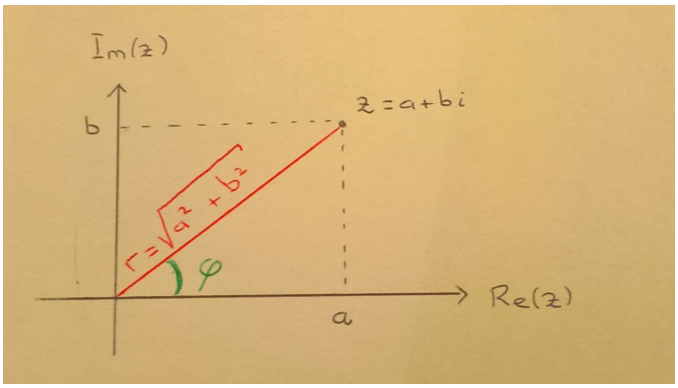


Figure 2: The complex plane

The right side is called the ***polar form*** of complex number z.

By Pythagoras' law, $r = \sqrt{a^2+b^2}$. This is called the ***absolute value*** or ***modulus*** of complex number $a+bi$, denoted $|z|$. If we define for each complex number $z = a+bi$ its ***(complex) conjugate*** $\overline{z}$ as

$$\overline{z} = \overline{a+bi} = a-bi,$$

we notice that

$$z\overline{z} = (a+bi)(a-bi) = a^2+b^2 \qquad \text{and consequently} \qquad |z| = \sqrt{a^2+b^2} = \sqrt{z\overline{z}}.$$

Straight from the definition of conjugates we have for any pair of complex numbers z_1 and z_2:

$$\mathrm{Re}(z_1) = \frac{z_1+\overline{z_1}}{2} \qquad \overline{\overline{z_1}} = z_1 \qquad \overline{z_1+z_2} = \overline{z_1}+\overline{z_2} \qquad \overline{z_1 z_2} = \overline{z_1}\,\overline{z_2}$$

$$\mathrm{Im}(z_1) = \frac{z_1-\overline{z_1}}{2} \qquad \overline{z_1}\overline{z_2} = \overline{z_1}z_2 \qquad (z_1 = \overline{z_1} \iff z_1 \text{ is real})$$

The properties of the absolute value on $\mathbb{C}$ are familiar from those on the real numbers. For all $z_1, z_2 \in \mathbb{C}$:

$$|z_1| \geq 0 \text{ with equality if and only if } z_1 = 0, \qquad |z_1 z_2| = |z_1||z_2|, \qquad |z_1+z_2| \leq |z_1|+|z_2|. \tag{145}$$

Formally, the absolute value $|\cdot|$ is a ***norm*** on the set $\mathbb{C}$ of complex numbers. (See exercise B.5).

B.3 Euler's identity

Leonhard Euler recognized that the polar form of a complex number could be rewritten as

$$r\left(\cos\varphi + i\sin\varphi\right) = re^{i\varphi}. \tag{146}$$

The special case where $r = 1$ and $\varphi = \pi$ together with $\cos\pi = -1$ and $\sin\pi = 0$ gives

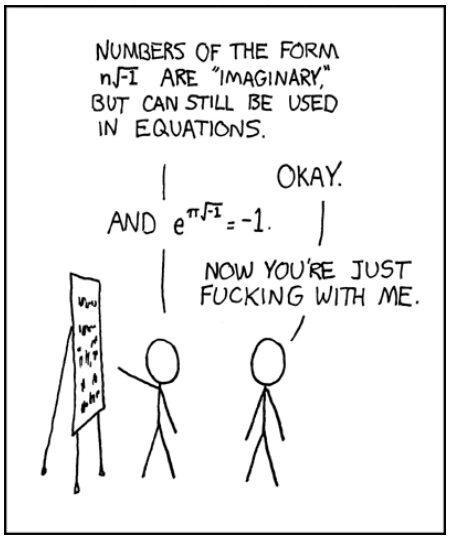
$$e^{i\pi} + 1 = 0.$$

This relation between $0, 1, e, \pi$, and i, arguably mathematics' most important constants, is called ***Euler's identity***. Mathematicians often rate it as one of the most beautiful theorems in our field. But why does (146) hold? It comes from definitions of the exponential, sine, and cosine functions as infinite series:

$$e^t = \sum_{k=0}^{\infty} \frac{t^k}{k!} = 1 + t + \frac{t^2}{2!} + \frac{t^3}{3!} + \cdots \tag{147}$$

$$\cos t = \sum_{k=0}^{\infty} \frac{(-1)^k t^{2k}}{(2k)!} = 1 - \frac{t^2}{2!} + \frac{t^4}{4!} - \frac{t^6}{6!} + \cdots \tag{148}$$

$$\sin t = \sum_{k=0}^{\infty} \frac{(-1)^k t^{2k+1}}{(2k+1)!} = t - \frac{t^3}{3!} + \frac{t^5}{5!} - \frac{t^7}{7!} + \cdots \tag{149}$$



e to the pi times *i*, xkcd.com/179/, CC BY-NC 2.5 license

Substitute $t = i\varphi$ in the exponential function to find

$$e^{i\varphi} = \sum_{k=0}^{\infty} \frac{(i\varphi)^k}{k!} = 1 + i\varphi + \frac{i^2\varphi^2}{2!} + \frac{i^3\varphi^3}{3!} + \cdots$$

Split this into two separate sums. The terms corresponding with even k are

$$1+\frac{i^2\varphi^2}{2!}+\frac{i^4\varphi^4}{4!}+\frac{i^6\varphi^6}{6!}+\cdots=1-\frac{\varphi^2}{2!}+\frac{\varphi^4}{4!}-\frac{\varphi^6}{6!}+\cdots=\cos\varphi,$$

where we used that $(i^2,i^4,i^6,i^8,\ldots)=(-1,1,-1,1,\ldots)$. And those corresponding with odd k are

$$\begin{aligned}i\varphi+\frac{i^3\varphi^3}{3!}+\frac{i^5\varphi^5}{5!}+\frac{i^7\varphi^7}{7!}+\cdots&=i\left(\varphi+\frac{i^2\varphi^3}{3!}+\frac{i^4\varphi^5}{5!}+\frac{i^6\varphi^7}{7!}+\cdots\right)\\&=i\left(\varphi-\frac{\varphi^3}{3!}+\frac{\varphi^5}{5!}-\frac{\varphi^7}{7!}+\cdots\right)\\&=i\sin\varphi.\end{aligned}$$

Putting them back together gives (146). We now have three ways of denoting complex numbers:

$$a+bi=r\left(\cos\varphi+i\sin\varphi\right)=re^{i\varphi},$$

with a, b, r, and φ related as in Figure 2. The identity $(e^x)^y=e^{xy}$ gives ***de Moivre's theorem***:

$$\text{for all } n=1,2,3,\ldots:\qquad (a+bi)^n=(re^{i\varphi})^n=r^ne^{i(n\varphi)}=r^n\left(\cos n\varphi+i\sin n\varphi\right).$$

B.4 Proof of the fundamental theorem of algebra

Consider a complex polynomial $p:\mathbb{C}\to\mathbb{C}$ defined as

$$p(x)=a_0+a_1x+\cdots+a_nx^n,\qquad a_n\neq 0.$$

To show: $p(z)=0$ for some $z\in\mathbb{C}$.[10] By the triangle inequality,

$$\left|a_nx^n\right|=\left|p(x)-a_0-a_1x-\cdots-a_{n-1}x^{n-1}\right|\leq\left|p(x)\right|+|a_0|+|a_1x|+\cdots+\left|a_{n-1}x^{n-1}\right|.$$

Rearranging terms and writing $\left|a_jx^j\right|=\left|a_j\right||x|^j$ gives

$$\begin{aligned}\left|p(x)\right|&\geq|a_n||x|^n-|a_0|-|a_1||x|-\cdots-|a_{n-1}||x|^{n-1}\\&=|x|^n\left(|a_n|-\frac{|a_0|}{|x|^n}-\frac{|a_1|}{|x|^{n-1}}-\cdots-\frac{|a_{n-1}|}{|x|}\right).\end{aligned}$$

Let $x_0\in\mathbb{C}$. In the previous line the term in brackets goes to $|a_n|>0$ as $|x|\to\infty$. So for $R>0$ large enough, $|x|>R$ implies that the term exceeds $|a_n|/2$. If R is also so large that $R^n\frac{|a_n|}{2}>\left|p(x_0)\right|$ then

$$|x|>R\qquad\Longrightarrow\qquad\left|p(x)\right|>R^n\frac{|a_n|}{2}>\left|p(x_0)\right|.$$

Therefore, any x with small $\left|p(x)\right|$ must lie in the closed ball around the origin with radius R. By the Extreme Value Theorem, the continuous function

$$(\alpha,\beta)\mapsto\left|p(\alpha+\beta i)\right|$$

has a minimum at some point (α_0,β_0) in this compact ball $\{(\alpha,\beta)\in\mathbb{R}^2:\alpha^2+\beta^2\leq R^2\}$. We argue that $z=\alpha_0+\beta_0 i$ has minimal value $\left|p(z)\right|=0$, making it the desired root of the polynomial.

[10] This proof is based on O. R. B. de Oliveira (2011), The Fundamental Theorem of Algebra: An elementary and direct proof, *The Mathematical Intelligencer* 33, 1–2.

Translating $x \mapsto x-z$ if necessary we may assume that $z=0$. Let $U=\{u\in\mathbb{C}:|u|=1\}$. By construction, for each $\alpha\geq 0$ and $u\in U$, $\big|p(\alpha u)\big|^2-\big|p(0)\big|^2\geq 0$ and $p(0)=a_0$. So $p(x)=p(0)+x^k q(x)$ for some $k\in\{1,\ldots,n\}$ and some polynomial q with $q(0)\neq 0$. With $x=\alpha u$ we find

$$0\leq \big|p(\alpha u)\big|^2-\big|p(0)\big|^2=\Big|p(0)+\alpha^k u^k q(\alpha u)\Big|^2-\big|p(0)\big|^2=2\alpha^k \operatorname{Re}\Big[\overline{p(0)}\,u^k q(\alpha u)\Big]+\alpha^{2k}\big|q(\alpha u)\big|^2. \quad (150)$$

Divide by $\alpha^k>0$ and let $\alpha\to 0$ to find that for all $u\in U$:

$$2\operatorname{Re}\Big[\overline{p(0)}\,u^k q(0)\Big]\geq 0. \quad (151)$$

Each $u\in U$ can be written in polar coordinates $u=\cos\varphi+i\sin\varphi$, since $r=|u|=1$. By de Moivre's theorem, $u^k=\cos k\varphi+i\sin k\varphi$. In particular, there are four elements $u\in U$ with u^k equal to 1, -1, i, and $-i$, respectively. For instance, for $u^k=1$ we want $\cos k\varphi=1$ and $\sin k\varphi=0$, so taking $\varphi=0$ works. Putting $u^k=1$ into (151) gives $\operatorname{Re}(\overline{p(0)}q(0))\geq 0$. And $u^k=-1$ gives the reverse inequality. So $\operatorname{Re}(\overline{p(0)}q(0))=0$. Likewise, plugging in $u^k=i$ and $u^k=-i$ and using that $\operatorname{Re}(wi)=-\operatorname{Im}(w)$ for each complex w gives $\operatorname{Im}(\overline{p(0)}q(0))=0$. So $\overline{p(0)}q(0)=0$. Dividing by $q(0)\neq 0$, gives $\overline{p(0)}=0$, so $p(0)=0$!

Exercises section B

B.1 Given $v=4-i$ and $w=-3+2i$, compute: (a) $v+w$, (b) $2w-iv$, (c) vw, (d) $\overline{v}$, (e) $\frac{v}{w}$, (f) $\frac{wi}{(1-i)v}$.

B.2 Write in polar coordinates: (a) $2\sqrt{3}-6i$, (b) -1, (c) $1+\sqrt{3}i$, (d) $-3-3i$

B.3 Let z be a complex number with the property that z^2 is real and nonnegative. Prove that z itself is real.

B.4 Assuming that $a_2+b_2 i\neq 0$, write $\frac{a_1+b_1 i}{a_2+b_2 i}$ as a complex number $a+bi$.

B.5 Verify the three properties in (145).

B.6 **(Alternative proof of de Moivre's theorem)** Multiply complex numbers $e^{i\varphi_k}=\cos\varphi_k+i\sin\varphi_k$ corresponding with two angles φ_1 and φ_2 to prove trigonometric identities

$$\sin(\varphi_1+\varphi_2)=\sin\varphi_1\cos\varphi_2+\cos\varphi_1\sin\varphi_2, \quad (152)$$
$$\cos(\varphi_1+\varphi_2)=\sin\varphi_1\sin\varphi_2-\cos\varphi_1\cos\varphi_2. \quad (153)$$

Use these equations to prove by induction:

$$\text{for all } n=1,2,3,\ldots: \qquad (a+bi)^n=r^n\big(\cos n\varphi+i\sin n\varphi\big). \quad (154)$$

B.7 Verify the final equality in (150).

C Some short solutions

1.3 $\Longrightarrow$: Assume R is negatively transitive: for all $x, y, z \in X$, if xRy, then xRz or zRy. Let $a, b, c \in X$ satisfy (not aRb) and (not bRc). We must show that (not aRc).

If, to the contrary, aRc, then negative transitivity implies that aRb or bRc, a contradiction.

$\Longleftarrow$: Assume R satisfies

$$\text{for all } x, y, z \in X, \text{ if (not } xRy\text{) and (not } yRz\text{), then (not } xRz\text{).} \tag{155}$$

Let $a, b, c \in X$ satisfy aRb. We must show that aRc or cRb.

If, to the contrary, both are false, then (155) with a, b, c instead of x, z, y implies that (not aRb), a contradiction.

3.1 Looking at Definition 3.1 you see that (a) and (d) are true. But (b) is false: that definition does not mention subtraction. Remember: we only introduced subtraction indirectly in terms of addition in the text following Theorem 3.1. Also (c) is false: the definition of a vector space involves scalar multiplication, i.e., multiplication of a vector with a number, not with another vector.

3.2 (a) No: W is not closed under scalar multiplication. For instance, vector $x = (1, \ldots, 1)$ with all coordinates equal to one belongs to W. Take scalar $\alpha = 1/2$. Then $\alpha x = (1/2, \ldots, 1/2)$ does *not* belong to W, since its coordinates are not integers. By Theorem 3.2, W is not a subspace of $\mathbb{R}^n$.

(b) Yes:

☒ The $n \times n$ zero matrix is symmetric:

$$\mathbf{0}^\top = \begin{pmatrix} 0 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & 0 \end{pmatrix}^\top = \begin{pmatrix} 0 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & 0 \end{pmatrix} = \mathbf{0}.$$

So W contains the zero matrix.

☒ W is closed under addition: if A and B are symmetric matrices, then so is their sum, because

$$(A+B)^\top = A^\top + B^\top = A + B.$$

☒ W is closed under scalar multiplication: if A is a symmetric matrix and α a scalar, then αA is symmetric, because

$$(\alpha A)^\top = \alpha A^\top = \alpha A.$$

☒ By Theorem 3.2, W is a subspace of $\mathbb{R}^{n \times n}$.

(c) Yes:

☒ The zero function $\mathbf{0} : [0,1] \to \mathbb{R}$ with $\mathbf{0}(x) = 0$ for all $x \in [0,1]$ definitely belongs to W, since it equals zero at both $x = 0$ and $x = 1$.

☒ W is closed under addition: if $f, g \in W$, then $f(0) = f(1)$ and $g(0) = g(1)$, so

$$(f+g)(0) = f(0) + g(0) = f(1) + g(1) = (f+g)(1).$$

☒ W is closed under scalar multiplication: if $f \in W$ and $\alpha \in \mathbb{R}$, then

$$(\alpha f)(0) = \alpha f(0) = \alpha f(1) = (\alpha f)(1).$$

☒ By Theorem 3.2, W is a subspace of $C[0,1]$.

(d) Yes:

☒ In the zero sequence $\mathbf{0} = (0, 0, 0, \ldots)$ there are no terms k with $x_k \neq 0$. That is certainly a finite number, so $\mathbf{0}$ belongs to W, making W nonempty.

- ☒ W is closed under addition: Let x and y lie in W. Note that if $x_i = 0$ and $y_i = 0$, then their sum is zero. So the only coordinates in which $x_i + y_i$ can possibly be different from zero are those where x_i or y_i is different from zero. And since x and y lie in W, there are only finitely many such coordinates: $x + y$ lies in W.
- ☒ W is closed under scalar multiplication: Let $x \in W$ and let α be a scalar. If $\alpha = 0$, then all coordinates of αx are zero, so $\alpha x \in W$. And if α is distinct from zero, the coordinates of αx that are equal to zero are the same as those where x is equal to zero, by assumption a finite number. So $\alpha x \in W$ also in this case.
- ☒ By Theorem 3.2, W is a subspace of $\mathbb{R}^{\mathbb{N}}$.

3.3 (a) Yes:

- ☒ Since $A\mathbf{0} = \mathbf{0}$, the set contains the zero vector.
- ☒ It is closed under addition: if x and y belong to the set ($Ax = Ay = \mathbf{0}$), then $A(x+y) = Ax+Ay = \mathbf{0}+\mathbf{0} = \mathbf{0}$, so also $x + y$ belongs to the set.
- ☒ It is closed under scalar multiplication: if x belongs to the set ($Ax = \mathbf{0}$) and α is a scalar, then $A(\alpha x) = \alpha(Ax) = \alpha\mathbf{0} = \mathbf{0}$, so also αx belongs to the set.
- ☒ By Theorem 3.2, the set $\{\mathbf{0}\}$ is a subspace of $\mathbb{R}^n$.

(b) No: since $A\mathbf{0} = \mathbf{0} \neq b$, the set does not contain the zero vector. By Theorem 3.2, it cannot be a subspace.

3.4

- ☒ $\{\mathbf{0}\}$ contains the zero vector $\mathbf{0}$.
- ☒ $\{\mathbf{0}\}$ is closed under addition: $\mathbf{0} + \mathbf{0} = \mathbf{0} \in \{\mathbf{0}\}$ by (V3).
- ☒ $\{\mathbf{0}\}$ is closed under scalar multiplication: for each scalar α, $\alpha\mathbf{0} = \mathbf{0} \in \{\mathbf{0}\}$ by Theorem 3.1(e).
- ☒ By Theorem 3.2, $\{\mathbf{0}\}$ is a subspace of V.

3.5 (a)

- ☒ $\times_{i\in I} V_i$ is closed under addition: Let $x = (x_i)_{i\in I}$ and $y = (y_i)_{i\in I}$ be elements of $\times_{i\in I} V_i$. For each $i \in I$, vector space V_i is closed under addition, so $(x + y)_i = x_i + y_i \in V_i$. Hence, $x + y \in \times_{i\in I} V_i$.
- ☒ $\times_{i\in I} V_i$ is closed under scalar multiplication: Let $x = (x_i)_{i\in I}$ be an element of $\times_{i\in I} V_i$ and let α be a scalar. For each $i \in I$, vector space V_i is closed under scalar multiplication, so $(\alpha x)_i = \alpha x_i \in V_i$. Hence, $\alpha x \in \times_{i\in I} V_i$.
- ☒ (V1): Let $x = (x_i)_{i\in I}$ and $y = (y_i)_{i\in I}$ be elements of $\times_{i\in I} V_i$. For each $i \in I$, vector space V_i satisfies (V1), so
$$(x + y)_i = x_i + y_i \overset{\text{(V1)}}{=} y_i + x_i = (y + x)_i.$$
Hence, $x + y = y + x$.
- ☒ (V2): Let $x = (x_i)_{i\in I}$, $y = (y_i)_{i\in I}$, and $z = (z_i)_{i\in I}$ be elements of $\times_{i\in I} V_i$. For each $i \in I$, vector space V_i satisfies (V2), so
$$((x + y) + z)_i = (x + y)_i + z_i = (x_i + y_i) + z_i \overset{\text{(V2)}}{=} x_i + (y_i + z_i) = x_i + (y + z)_i = (x + (y + z))_i.$$
Hence, $(x + y) + z = x + (y + z)$.
- ☒ (V3): Let $x = (x_i)_{i\in I}$ be an element of $\times_{i\in I} V_i$. For each $i \in I$, vector space V_i satisfies (V3), so there is a $\mathbf{0}_i \in V_i$ with $x_i + \mathbf{0}_i = x_i$ for all $x_i \in V_i$. It follows that $\mathbf{0} = (\mathbf{0}_i)_{i\in I}$ satisfies $(x + \mathbf{0})_i = x_i + \mathbf{0}_i = x_i$, so $x + \mathbf{0} = x$.
- ☒ (V4): Let $x = (x_i)_{i\in I}$ be an element of $\times_{i\in I} V_i$. For each $i \in I$, vector space V_i satisfies (V4), so there is a $y_i \in V_i$ with $x_i + y_i = \mathbf{0}_i$. It follows that $y = (y_i)_{i\in I}$ satisfies $(x + y)_i = x_i + y_i = \mathbf{0}_i$, so $x + y = \mathbf{0}$.
- ☒ (V5): Let $x = (x_i)_{i\in I}$ be an element of $\times_{i\in I} V_i$ and let α and β be scalars. For each $i \in I$, vector space V_i satisfies (V5), so
$$((\alpha\beta)x)_i = (\alpha\beta)x_i \overset{\text{(V5)}}{=} \alpha(\beta x_i) = \alpha(\beta x)_i = (\alpha(\beta x))_i.$$
Hence, $(\alpha\beta)x = \alpha(\beta x)$.
- ☒ (V6): Let $x = (x_i)_{i\in I}$ be an element of $\times_{i\in I} V_i$. For each $i \in I$, vector space V_i satisfies (V6), so
$$(1x)_i = 1x_i \overset{\text{(V6)}}{=} x_i.$$
Hence, $1x = x$.

☒ (V7): Let $x = (x_i)_{i\in I}$ and $y = (y_i)_{i\in I}$ be elements of $\times_{i\in I} V_i$ and let α be a scalar. For each $i \in I$, vector space V_i satisfies (V7), so

$$(\alpha(x+y))_i = \alpha(x+y)_i = \alpha(x_i+y_i) \overset{\text{(V7)}}{=} \alpha x_i + \alpha y_i = (\alpha x)_i + (\alpha y)_i = (\alpha x + \alpha y)_i.$$

Hence, $\alpha(x+y) = \alpha x + \alpha y$.

☒ (V8): Let $x = (x_i)_{i\in I}$ be an element of $\times_{i\in I} V_i$ and let α and β be scalars. For each $i \in I$, vector space V_i satisfies (V8), so

$$((\alpha+\beta)x)_i = (\alpha+\beta)x_i \overset{\text{(V8)}}{=} \alpha x_i + \beta x_i = (\alpha x)_i + (\beta x)_i = (\alpha x + \beta x)_i.$$

Hence, $(\alpha+\beta)x = \alpha x + \beta x$.

(b)

Example	I	V_i
3.1	$\{(i,j) : i \in \{1,2\}, j \in \{1,2,3\}\}$	$\mathbb{R}$
3.2	$\mathbb{R}$	$\mathbb{R}$
3.3	$\{1,\ldots,n\}$	$\mathbb{R}$
3.4	$\{(i,j) : i \in \{1,\ldots,m\}, j \in \{1,\ldots,n\}\}$	$\mathbb{R}$
3.5	$\mathbb{N}$	$\mathbb{R}$

3.6 When adding polynomials, say p and q, we will often simplify expressions by writing both of them in generic form:

$$p(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0,$$
$$q(x) = b_n x^n + b_{n-1}x^{n-1} + \cdots + b_1 x + b_0,$$

even if the polynomials have different degrees. If p has degree n and q has degree $m < n$, we can simply take $b_n = b_{n-1} = \cdots = b_{m+1} = 0$. Throughout the solution, let

$$p(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0,$$
$$q(x) = b_n x^n + b_{n-1}x^{n-1} + \cdots + b_1 x + b_0,$$
$$r(x) = c_n x^n + c_{n-1}x^{n-1} + \cdots + c_1 x + c_0$$

be elements of $P(\mathbb{R})$ and let α, β be elements of $\mathbb{R}$.

☒ Expressions (1) and (2) show that $P(\mathbb{R})$ is closed under addition and scalar multiplication.

☒ (V1): Using commutativity of addition on $\mathbb{R}$ (field property (F1) in Appendix A.1), we find

$$\begin{aligned}(p+q)(x) &= (a_n+b_n)x^n + (a_{n-1}+b_{n-1})x^{n-1} + \cdots + (a_1+b_1)x + (a_0+b_0)\\ &= (b_n+a_n)x^n + (b_{n-1}+a_{n-1})x^{n-1} + \cdots + (b_1+a_1)x + (b_0+a_0)\\ &= (q+p)(x),\end{aligned}$$

showing that $p+q = q+p$.

☒ (V2): Using associativity of addition on $\mathbb{R}$ (field property (F2) in Appendix A.1), we find

$$\begin{aligned}((p+q)+r)(x) &= ((a_n+b_n)+c_n)x^n + ((a_{n-1}+b_{n-1})+c_{n-1})x^{n-1} + \cdots + ((a_1+b_1)+c_1)x + ((a_0+b_0)+c_0)\\ &= (a_n+(b_n+c_n))x^n + (a_{n-1}+(b_{n-1}+c_{n-1}))x^{n-1} + \cdots + (a_1+(b_1+c_1))x + (a_0+(b_0+c_0))\\ &= (p+(q+r))(x),\end{aligned}$$

showing that $(p+q)+r = p+(q+r)$.

☒ (V3): Using that $a_i + 0 = a_i$ for each real number a_i (field property (F3) in Appendix A.1), we see that the zero polynomial

$$\mathbf{0}(x) = 0x^n + 0x^{n-1} + \cdots + 0x + 0$$

satisfies

$$\begin{aligned}(p + \mathbf{0})(x) &= (a_n + 0)x^n + (a_{n-1} + 0)x^{n-1} + \cdots + (a_1 + 0)x + (a_0 + 0) \\ &= a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0 \\ &= p(x),\end{aligned}$$

showing that $p + \mathbf{0} = p$.

☒ (V4): Using the fact that each real number a_i has an additive inverse $-a_i$ satisfying $a_i + (-a_i) = 0$ (field property (F4) in Appendix A.1), we see that the polynomial

$$\hat{p}(x) = (-a_n)x^n + (-a_{n-1})x^{n-1} + \cdots + (-a_1)x + (-a_0)$$

satisfies

$$\begin{aligned}(p + \hat{p})(x) &= (a_n + (-a_n))x^n + (a_{n-1} + (-a_{n-1}))x^{n-1} + \cdots + (a_1 + (-a_1))x + (a_0 + (-a_0)) \\ &= 0x^n + 0x^{n-1} + \cdots + 0x + 0 \\ &= \mathbf{0}(x),\end{aligned}$$

showing that $p + \hat{p} = \mathbf{0}$.

☒ (V5): Using associativity of multiplication on $\mathbb{R}$ (field property (F2) in Appendix A.1), we see that

$$\begin{aligned}((\alpha\beta)p)(x) &= (\alpha\beta)a_n x^n + (\alpha\beta)a_{n-1}x^{n-1} + \cdots + (\alpha\beta)a_1 x + (\alpha\beta)a_0 \\ &= \alpha(\beta a_n)x^n + \alpha(\beta a_{n-1})x^{n-1} + \cdots + \alpha(\beta a_1)x + \alpha(\beta a_0) \\ &= (\alpha(\beta p))(x),\end{aligned}$$

showing that $(\alpha\beta)p = \alpha(\beta p)$.

☒ (V6): Since $1a_i = a_i$ for each real number a_i (field property (F3) in Appendix A.1), we see that

$$\begin{aligned}(1p)(x) &= (1a_n)x^n + (1a_{n-1})x^{n-1} + \cdots + (1a_1)x + (1a_0) \\ &= a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0 \\ &= p(x),\end{aligned}$$

showing that $1p = p$.

☒ (V7): Distributivity of multiplication over addition on $\mathbb{R}$ (field property (F5) in Appendix A.1) implies

$$\begin{aligned}(\alpha(p + q))(x) &= \alpha(a_n + b_n)x^n + \alpha(a_{n-1} + b_{n-1})x^{n-1} + \cdots + \alpha(a_1 + b_1)x + \alpha(a_0 + b_0) \\ &= (\alpha a_n + \alpha b_n)x^n + (\alpha a_{n-1} + \alpha b_{n-1})x^{n-1} + \cdots + (\alpha a_1 + \alpha b_1)x + (\alpha a_0 + \alpha b_0) \\ &= (\alpha p + \alpha q)(x),\end{aligned}$$

showing that $\alpha(p + q) = \alpha p + \alpha q$.

☒ (V8): Distributivity of multiplication over addition on $\mathbb{R}$ (field property (F5) in Appendix A.1) implies

$$\begin{aligned}((\alpha + \beta)p)(x) &= (\alpha + \beta)a_n x^n + (\alpha + \beta)a_{n-1}x^{n-1} + \cdots + (\alpha + \beta)a_1 x + (\alpha + \beta)a_0 \\ &= (\alpha a_n + \beta a_n)x^n + (\alpha a_{n-1} + \beta a_{n-1})x^{n-1} + \cdots + (\alpha a_1 + \beta a_1)x + (\alpha a_0 + \beta a_0) \\ &= (\alpha p + \beta p)(x),\end{aligned}$$

showing that $(\alpha + \beta)p = \alpha p + \beta p$.

3.7 (a) Let W be a subspace of V: W is a vector space under the operations of addition and multiplication defined on V. In particular, W is closed under addition and scalar multiplication: (ii) and (iii) hold. The only tricky part is that the zero vector of W must be the zero vector $\mathbf{0}$ of V. Since W is a vector space, (V3) states that there is a vector $\mathbf{0}_W$ such that $x+\mathbf{0}_W = x$ for all $x \in W$. But also $x+\mathbf{0} = x$ for all $x \in W$. Hence $\mathbf{0}_W = \mathbf{0}$ by the cancellation law, proving (i).

Conversely, assume that $W \subseteq V$ satisfies properties (i), (ii), and (iii) in Theorem 3.2. We need to prove that it is a vector space. By (ii) and (iii), it is closed under addition and scalar multiplication. Since properties (V1), (V2), (V5), (V6), (V7), (V8) hold for all elements of V, they automatically hold for all elements of $W \subseteq V$. (V3) holds by (i). It remains to prove that (V4) holds on W: for each $x \in W$ there is an $y \in W$ with $x+y = \mathbf{0}$.

If $x \in W$, then $(-1)x \in W$ by (iii). By Theorem 3.1(f), $-x = (-1)x \in W$: the additive inverse of x lies in W.

(b) Let W be a subspace of V. By Theorem 3.2, it contains the zero vector, so W is nonempty. Next, let $x, y \in W$ and $\alpha, \beta \in \mathbb{R}$. Since W is closed under scalar multiplication, αx and βy lie in W. And since W is closed under addition, $\alpha x + \beta y$ lies in W.

Conversely, assume that W satisfies the properties in (b). Since W is nonempty, let $w \in W$. It follows that $0w + 0w = (0+0)w = 0w = \mathbf{0} \in W$. Since $\alpha x + \beta y \in W$ whenever $x, y \in W$ and $\alpha, \beta \in \mathbb{R}$, it follows that W is closed under addition (take $\alpha = \beta = 1$) and scalar multiplication (take $\beta = 0$).

4.1 Recall from the text following Definition 4.2 that a finite set $W = \{w_1, \ldots, w_n\}$ of different vectors is linearly independent if and only if

$$\alpha_1 w_1 + \cdots + \alpha_n w_n = \mathbf{0} \tag{156}$$

implies that $\alpha_1 = \cdots = \alpha_n = 0$. So we solve equation (156): if the only solution is $\alpha_1 = \cdots = \alpha_n = 0$, set W is linearly independent; otherwise it is linearly dependent.

(a) W is linearly independent. The equation (156) becomes

$$\alpha_1(1,0) + \alpha_2(2,-1) = \mathbf{0} = (0,0).$$

Rewrite:

$$\alpha_1(1,0) + \alpha_2(2,-1) = (\alpha_1 + 2\alpha_2, -\alpha_2) = (0,0).$$

The second coordinate says $-\alpha_2 = 0$, so $\alpha_2 = 0$. Substitute this into the first coordinate: $\alpha_1 + 2 \cdot 0 = 0$, so $\alpha_1 = 0$. So the only solution is $\alpha_1 = \alpha_2 = 0$: W is linearly independent.

(b) W is linearly independent. The system of linear equations in (156) becomes

$$\alpha_1(1,2,3) + \alpha_2(0,2,3) + \alpha_3(-4,4,5) = \mathbf{0} = (0,0,0).$$

In matrix notation, the system can be written as

$$\begin{bmatrix} 1 & 0 & -4 \\ 2 & 2 & 4 \\ 3 & 3 & 5 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}.$$

By Gaussian elimination on the augmented matrix:

$$\left[\begin{array}{ccc|c} 1 & 0 & -4 & 0 \\ 2 & 2 & 4 & 0 \\ 3 & 3 & 5 & 0 \end{array}\right] \sim \left[\begin{array}{ccc|c} 1 & 0 & -4 & 0 \\ 0 & 2 & 12 & 0 \\ 0 & 3 & 17 & 0 \end{array}\right] \sim \left[\begin{array}{ccc|c} 1 & 0 & -4 & 0 \\ 0 & 1 & 6 & 0 \\ 0 & 3 & 17 & 0 \end{array}\right] \sim \left[\begin{array}{ccc|c} 1 & 0 & -4 & 0 \\ 0 & 1 & 6 & 0 \\ 0 & 0 & -1 & 0 \end{array}\right]$$

$$\sim \left[\begin{array}{ccc|c} 1 & 0 & -4 & 0 \\ 0 & 1 & 6 & 0 \\ 0 & 0 & 1 & 0 \end{array}\right] \sim \left[\begin{array}{ccc|c} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{array}\right]$$

So $\alpha_1 = \alpha_2 = \alpha_3 = 0$ is the only solution: W is linearly independent.

(c) W is linearly dependent: as in the previous case, the system of equations can be written in matrix notation

$$\begin{bmatrix} 1 & 0 & 1 \\ 2 & 2 & -2 \\ 3 & 3 & -3 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix},$$

By Gaussian elimination:

$$\left[\begin{array}{ccc|c} 1 & 0 & 1 & 0 \\ 2 & 2 & -2 & 0 \\ 3 & 3 & -3 & 0 \end{array}\right] \sim \left[\begin{array}{ccc|c} 1 & 0 & 1 & 0 \\ 0 & 2 & -4 & 0 \\ 0 & 3 & -6 & 0 \end{array}\right] \sim \left[\begin{array}{ccc|c} 1 & 0 & 1 & 0 \\ 0 & 1 & -2 & 0 \\ 0 & 3 & -6 & 0 \end{array}\right] \sim \left[\begin{array}{ccc|c} 1 & 0 & 1 & 0 \\ 0 & 1 & -2 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right].$$

So α_3 is a free variable: the set of solutions is $\{(-\alpha_3, 2\alpha_3, \alpha_3) : \alpha_3 \in \mathbb{R}\}$. Since there are nonzero solutions — for instance, if we choose $\alpha_3 = 1$, we find a solution $(\alpha_1, \alpha_2, \alpha_3) = (-1, 2, 1)$ — the set W is linearly dependent.

(d) W is linearly independent. Equation (156) becomes

$$\alpha_1 \cdot 3 + \alpha_2 \cdot x + \alpha_3 \cdot (2x^2 + x - 2) = 0.$$

Rewrite:

$$(2\alpha_3)x^2 + (\alpha_2 + \alpha_3)x + (3\alpha_1 - 2\alpha_3) = 0.$$

So all coefficients $2\alpha_3$, $\alpha_2 + \alpha_3$, and $3\alpha_1 - 2\alpha_3$ of the polynomial on the left-hand side must be zero. The first gives $\alpha_3 = 0$. Substituting this into the other two gives $\alpha_2 = 0$ and $\alpha_1 = 0$. So the only solution is $\alpha_1 = \alpha_2 = \alpha_3 = 0$: W is linearly independent.

(e) W is linearly independent. Equation (156) becomes $\alpha_1 f + \alpha_2 g = \mathbf{0}$. Rewrite:

$$\alpha_1 f(x) + \alpha_2 g(x) = \alpha_1 x + \alpha_2 \frac{1}{x+2} = 0 \qquad \text{for all } x \in [0,1].$$

If we substitute $x = 0$, it follows that $\alpha_1 \cdot 0 + \alpha_2 \cdot \frac{1}{0+2} = \frac{\alpha_2}{2} = 0$, so $\alpha_2 = 0$. Using this and substituting $x = 1$ then gives $\alpha_1 \cdot 1 + 0 \cdot \frac{1}{1+2} = \alpha_1 = 0$. So the only solution is $\alpha_1 = \alpha_2 = 0$: W is linearly independent.

4.2 (a) By definition, a subspace is closed under addition and scalar multiplication. Consequently, by induction, it contains all linear combinations of its elements. Formally, if U is a subspace containing W, it must contain $\operatorname{span}(W)$ as well: $\operatorname{span}(W) \subseteq U$. Since $\operatorname{span}(W)$ is a subspace, this establishes that $\operatorname{span}(W)$ is indeed the smallest subspace containing W.

(b) The intersection of all subspaces containing W trivially contains W and, by Theorem 3.2, is again a subspace. By (a), this intersection contains $\operatorname{span}(W)$. For the converse inclusion, simply observe that $\operatorname{span}(W)$ is one of the subspaces containing W, so that the intersection must be contained in $\operatorname{span}(W)$.

4.3 ☒ Assume W is linearly dependent. Then W is nonempty by Definition 4.2. If W has exactly one element w, it must be that $\alpha w = \mathbf{0}$ for some nonzero scalar α. Dividing by α gives $w = \alpha^{-1}\mathbf{0} = \mathbf{0}$, i.e., $W = \{\mathbf{0}\}$. If W has more than one element, there is a finite number $n \in \mathbb{N}$ of distinct vectors $w_1, \ldots, w_n$ in W and scalars $\alpha_1, \ldots, \alpha_n$, not all zero, such that

$$\alpha_1 w_1 + \cdots + \alpha_n w_n = \mathbf{0}.$$

Relabeling if necessary, we may assume that $\alpha_1 \neq 0$. It follows that

$$\alpha_1 w_1 = -\alpha_2 w_2 - \cdots - \alpha_n w_n,$$

and, after dividing by $\alpha_1 \neq 0$, that

$$w_1 = -\frac{\alpha_2}{\alpha_1} w_2 - \cdots - \frac{\alpha_n}{\alpha_1} w_n,$$

making w_1 a linear combination of $w_2, \ldots, w_n$.

☒ Conversely, assume that $W = \{\mathbf{0}\}$ or there exist distinct vectors $w, w_1, \ldots, w_n$ in W such that w is a linear combination of $w_1, \ldots, w_n$. In the first case, W is linearly dependent, since $\alpha\mathbf{0} = \mathbf{0}$ for each nonzero scalar α by Theorem 3.1(e). In the second case, there are scalars $\alpha_1, \ldots, \alpha_n$ with

$$w = \alpha_1 w_1 + \cdots + \alpha_n w_n.$$

Rearrange terms:

$$-1\, w + \alpha_1 w_1 + \cdots + \alpha_n w_n = \mathbf{0}.$$

The expression on the left is a nontrivial (since w has scalar $-1 \neq 0$) linear combination of distinct vectors resulting in the zero vector, so W is linearly dependent.

5.1 (a)
- ☒ The collection $\{\{1\},\ \{2,3\}\}$ consists of the two finite subsets $\{1\}$ and $\{2,3\}$ of $\mathbb{N}$. Since neither contains the other, this collection is not a chain.
- ☒ Since $\{1\} \subseteq \{1,2\} \subseteq \{1,2,3\} \subseteq \cdots$, the collection of all sets of the form $\{1, \ldots, n\}$ for $n \in \mathbb{N}$ is a chain.
- ☒ No, the chain in the answer above has no upper bound. Suppose it does: then there is an element of $A \in \mathscr{A}$ that contains all sets of the form $\{1, \ldots, n\}$. That is impossible: A has to be nonempty and a finite subset of $\mathbb{N}$, so it has a largest element, say k. But then it does not contain the set $\{1, \ldots, k, k+1\}$ belonging to the chain.
- ☒ No, $\mathscr{A}$ has no maximal element. Each element $A \in \mathscr{A}$ is a finite subset of $\mathbb{N}$. Therefore, we can choose an element $n \in \mathbb{N}, n \notin A$. Hence, $A \subseteq A \cup \{n\} \in \mathscr{A}$ shows that each element of $\mathscr{A}$ is contained in a strictly larger set in $\mathscr{A}$.

(b) The answers to the first three questions can be copied from (a). But $\mathscr{A}$ has one maximal element, namely $\{-37\}$. This is the only element of $\mathscr{A}$ containing the number -37, hence there is no set in $\mathscr{A}$ that properly contains $\{-37\}$.

(c) The answer to the first question can be copied from (a). Here are the others:
- ☒ Since $\emptyset \subseteq \{1\} \subseteq \{1,2\}$, the collection $\{\emptyset, \{1\}, \{1,2\}\}$ is a chain.
- ☒ Each chain in $\mathscr{A}$ has an upper bound: since each set in the chain has at most two elements, the chain must have a largest set (with at most two elements). Consequently, this set is an upper bound of the chain!
- ☒ $\mathscr{A}$ has infinitely many maximal elements, namely all subsets of $\mathbb{N}$ with two elements. By construction, there is no set in $\mathscr{A}$ that can properly contain a two-element set, since such a set has to have at least three elements and therefore does not belong to $\mathscr{A}$.

6.1 Let $x \in \mathbb{R}^2$. By linearity of T we have

$$T(x) = T(x_1 e_1 + x_2 e_2) = x_1 T(e_1) + x_2 T(e_2) = \begin{bmatrix} T(e_1) & T(e_2) \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = Ax,$$

where $A \in \mathbb{R}^{3\times 2}$ has $T(e_1)$ as its first and $T(e_2)$ as its second column. So we need to find $T(e_1)$ and $T(e_2)$. Since

$$e_1 = (1,0) = (1,1) - \frac{1}{2}(0,2) \qquad \text{and} \qquad e_2 = (0,1) = \frac{1}{2}(0,2),$$

linearity of T gives

$$T(e_1) = T(1,1) - \frac{1}{2}T(0,2) = (2,0,-3) - \frac{1}{2}(-1,4,2) = \left(\frac{5}{2}, -2, -4\right),$$
$$T(e_2) = \frac{1}{2}T(0,2) = \frac{1}{2}(-1,4,2) = \left(-\frac{1}{2}, 2, 1\right).$$

Conclude:

$$A = \begin{bmatrix} T(e_1) & T(e_2) \end{bmatrix} = \begin{bmatrix} \frac{5}{2} & -\frac{1}{2} \\ -2 & 2 \\ -4 & 1 \end{bmatrix}$$

6.2 (a) The null space consists of all vectors x with $T(x) = Ax = \mathbf{0}$. We solve the system $Ax = \mathbf{0}$ of linear equations by Gaussian elimination on the augmented matrix (or just on A, because the augmented matrix in the special case where $\mathbf{0}$ is the final column may be overdoing it a bit: if the final column is the zero vector, it isn't affected by the elementary row operations of the Gaussian elimination process.):

$$\left[\begin{array}{ccccc|c} 1 & 4 & 5 & 6 & 9 & 0 \\ 3 & -2 & 1 & 4 & -1 & 0 \\ 1 & 0 & -1 & -2 & -1 & 0 \\ 2 & 3 & 5 & 7 & 8 & 0 \end{array}\right] \sim \left[\begin{array}{ccccc|c} 1 & 4 & 5 & 6 & 9 & 0 \\ 0 & -14 & -14 & -14 & -28 & 0 \\ 0 & -4 & -6 & -8 & -10 & 0 \\ 0 & -5 & -5 & -5 & -10 & 0 \end{array}\right] \sim \left[\begin{array}{ccccc|c} 1 & 4 & 5 & 6 & 9 & 0 \\ 0 & 1 & 1 & 1 & 2 & 0 \\ 0 & 0 & -2 & -4 & -2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right]$$

$$\sim \left[\begin{array}{ccccc|c} 1 & 4 & 0 & -4 & 4 & 0 \\ 0 & 1 & 0 & -1 & 1 & 0 \\ 0 & 0 & 1 & 2 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right] \sim \left[\begin{array}{ccccc|c} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & -1 & 1 & 0 \\ 0 & 0 & 1 & 2 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right]$$

So the null space consists of all vectors x of the form

$$x = \begin{bmatrix} 0 \\ x_4 - x_5 \\ -2x_4 - x_5 \\ x_4 \\ x_5 \end{bmatrix} = x_4 \begin{bmatrix} 0 \\ 1 \\ -2 \\ 1 \\ 0 \end{bmatrix} + x_5 \begin{bmatrix} 0 \\ -1 \\ -1 \\ 0 \\ 1 \end{bmatrix} \text{ with } x_4, x_5 \in \mathbb{R} \text{ arbitrary real numbers.}$$

(b) From the previous expression, we see that the null space is spanned by the two vectors in the set

$$\left\{ \begin{bmatrix} 0 \\ 1 \\ -2 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ -1 \\ -1 \\ 0 \\ 1 \end{bmatrix} \right\}.$$

If follows from looking at their fourth and fifth coordinates that the vectors are linearly independent. So:

- ☒ These two vectors are linearly independent and span the null space: they are a basis for the null space.
- ☒ This basis for the null space has two elements, so the null space is two-dimensional.

(c) No. Since the null space is two-dimensional, each basis has two elements (Theorem 4.2). The set B_1 has three elements, so it can't be a basis.

(d) No. The elements of a basis must be linearly independent. The vectors in B_2 are linearly dependent (the first equals three times the second), so it can't be a basis.

(e) Yes. The vectors belong to the null space (verify) and are linearly independent (look at coordinates three and four). By Theorem 4.1, this two-element set B_3 can be extended to a basis B of the null space: B must satisfy $B_3 \subseteq B$ and have two elements (since the null space is two-dimensional). So $B = B_3$: the set B_3 is a basis.

6.5 ☒ To see that $L(V,W)$ is a subspace of the functions from V to W, we apply Theorem 3.2:

1. The zero function $\mathbf{0} : V \to W$ with $\mathbf{0}(x) = \mathbf{0}$ for all $x \in V$, belongs to $L(V,W)$: for all $x, y \in V$ and all scalars α:
$$\mathbf{0}(x+y) = \mathbf{0} = \mathbf{0} + \mathbf{0} = \mathbf{0}(x) + \mathbf{0}(y) \quad \text{and} \quad \mathbf{0}(\alpha x) = \mathbf{0} = \alpha\mathbf{0} = \alpha\mathbf{0}(x).$$
2. $L(V,W)$ is closed under addition: if $T_1, T_2 \in L(V,W)$, then $T_1 + T_2 \in L(V,W)$. Indeed, for all $x, y \in V$ and all scalars α:
$$\begin{aligned} (T_1+T_2)(x+y) &= T_1(x+y) + T_2(x+y) && \text{(by def. of } T_1+T_2) \\ &= T_1(x) + T_1(y) + T_2(x) + T_2(y) && \text{(by linearity of } T_1 \text{ and } T_2) \\ &= T_1(x) + T_2(x) + T_1(y) + T_2(y) && \text{(after rearranging terms)} \\ &= (T_1+T_2)(x) + (T_1+T_2)(y) && \text{(by def. of } T_1+T_2) \end{aligned}$$

and

$$\begin{aligned}(T_1+T_2)(\alpha x) &= T_1(\alpha x)+T_2(\alpha x) && \text{(by def. of } T_1+T_2)\\ &= \alpha T_1(x)+\alpha T_2(x) && \text{(by linearity of } T_1 \text{ and } T_2)\\ &= \alpha(T_1(x)+T_2(x)) && \text{(by distributivity)}\\ &= \alpha(T_1+T_2)(x). && \text{(by def. of } T_1+T_2)\end{aligned}$$

3. $L(V,W)$ is closed under scalar multiplication: if $T \in L(V,W)$ and $\alpha \in \mathbb{R}$, then $\alpha T \in L(V,W)$. Indeed, for all $x, y \in V$ and all scalars β, reasoning as above gives:

$$(\alpha T)(x+y) = \alpha(T(x+y)) = \alpha(T(x)+T(y)) = \alpha(T(x)) + \alpha(T(y)) = (\alpha T)(x) + (\alpha T)(y),$$

and

$$(\alpha T)(\beta x) = \alpha(T(\beta x)) = \alpha(\beta T(x)) = \beta(\alpha T(x)) = \beta(\alpha T)(x).$$

4. By Theorem 3.2, $L(V,W)$ is a subspace of the set of functions from V to W.

☒ Recall that $0x = \mathbf{0}$ for every vector x (Theorem 3.1(d)). Since T is linear:

$$T\mathbf{0} = T0\mathbf{0} = 0T\mathbf{0} = \mathbf{0}.$$

☒ We use Theorem 3.2 to show that the range of a linear function $T : V \to W$ is a subspace of W:

1. We showed before that $T(\mathbf{0}) = \mathbf{0}$, so $\text{range}(T) = \{T(v) : v \in V\}$ contains the zero vector.
2. The range is closed under addition: if x and y belong to the range of T, there exist v_1 and v_2 in V with $T(v_1) = x$ and $T(v_2) = y$. Since V is a vector space, $v_1 + v_2 \in V$ and by linearity of T, $T(v_1+v_2) = T(v_1) + T(v_2) = x + y$, showing that $x + y$ lies in the range of T.
3. The range is closed under scalar multiplication: if x lies in the range of T and α is a scalar, then there is a vector $v \in V$ with $T(v) = x$. Since V is a vector space: $\alpha v \in V$. Since T is linear, $T(\alpha v) = \alpha T(v) = \alpha x$, showing that αx lies in the range of T.

Conclude from Theorem 3.2 that $T(U)$ is a subspace of W.

☒ We use Theorem 3.2 to show that the kernel/null space of T is a subspace of V:

1. It contains the zero vector, since $\mathbf{0} \in V$ and $T(\mathbf{0}) = \mathbf{0}$.
2. The null space is closed under addition: if $x, y \in \ker(T)$, then $T(x) = T(y) = \mathbf{0}$. Since V is a vector space: $x + y \in V$. By linearity of T: $T(x+y) = T(x) + T(y) = \mathbf{0} + \mathbf{0} = \mathbf{0}$. Therefore, $x + y \in \ker(T)$.
3. The null space is closed under scalar multiplication: if $x \in \ker(T)$ and α is a scalar, then αx lies in V, since it is a vector space. Linearity of T gives $T(\alpha x) = \alpha T(x) = \alpha\mathbf{0} = \mathbf{0}$, so $\alpha x \in \ker(T)$.

Conclude from Theorem 3.2 that the null space of T is a subspace of V.

6.7 Throughout the solution, let $n \in \mathbb{N}$ denote the dimension of V.

(a) $\implies$ (b): A bijective function is both injective (and surjective).

(b) $\implies$ (c): Assume $T : V \to V$ is injective. By Exercise 6.4, its null space is $\{\mathbf{0}\}$ of dimension zero. By the rank-nullity theorem, $\dim(\text{range}(T)) = \dim(V) = n$.

So the range of T is a subspace of V with the same dimension as V. The only such subspace is V itself: $\text{range}(T) = V$, making T surjective.

A more explicit argument that $\text{range}(T) = V$: the range of T is an n-dimensional subspace of V. Let $\{b_1, \ldots, b_n\}$ be a basis of this range. These are n linearly independent vectors in V. By Theorem 4.3, they are a basis of V. So

$$\text{range}(T) = \text{span}(\{b_1, \ldots, b_n\}) = V.$$

(c) $\implies$ (a): Assume $T : V \to V$ is surjective: its range is V with dimension n. By the rank-nullity theorem, the null space of T has dimension zero, i.e., it consists only of the zero vector. By Exercise 6.4, T is injective. Being both injective and surjective, T is bijective.

7.1 (a) For each coordinate j:

$$\|x\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2} \geq \sqrt{x_j^2} = |x_j|.$$

Hence,

$$\|x\|_2 \geq \max\{|x_1|, \ldots, |x_n|\} = \|x\|_\infty.$$

Moreover,

$$\|x\|_2 = \sqrt{x_1^2 + \cdots + x_n^2} \leq \sqrt{\|x\|_\infty^2 + \cdots + \|x\|_\infty^2} = \sqrt{n\|x\|_\infty^2} = \sqrt{n}\|x\|_\infty$$

(b)

$$\|x\|_\infty = \max\{|x_1|, \ldots, |x_n|\} \leq \sum_{i=1}^{n} |x_i| = \|x\|_1,$$
$$\|x\|_1 = \sum_{i=1}^{n} |x_i| \leq \sum_{i=1}^{n} \max\{|x_1|, \ldots, |x_n|\} = n\|x\|_\infty.$$

(c) Write $x = \sum_{i=1}^{n} x_i e_i$ as a linear combination of standard basis vectors and use the triangle inequality:

$$\|x\|_2 = \left\| \sum_{i=1}^{n} x_i e_i \right\| \leq \sum_{i=1}^{n} \|x_i e_i\| = \sum_{i=1}^{n} |x_1| \underbrace{\|e_i\|}_{=1} = \sum_{i=1}^{n} |x_i| = \|x\|_1.$$

Write $\sum_{i=1}^{n} |x_i|$ as inner product of $(|x_1|, \ldots, |x_n|)$ and $(1, \ldots, 1)$ and use the Cauchy-Schwarz inequality:

$$\|x\|_1 = \sum_{i=1}^{n} |x_i| = \langle (|x_1|, \ldots, |x_n|), (1, \ldots, 1) \rangle \leq \|(|x_1|, \ldots, |x_n|)\| \|(1, \ldots, 1)\| = \sqrt{n}\|x\|_2.$$

7.2 Let $x, y \in V$. Rewriting (note: $|x| \leq \varepsilon \iff -\varepsilon \leq x \leq \varepsilon$), we need to prove

$$-\|x - y\| \leq \|x\| - \|y\| \leq \|x - y\|.$$

The first inequality follows from

$$\|y\| = \|x - (x - y)\| \overset{\text{(N4)}}{\leq} \|x\| + \|-(x - y)\| \overset{\text{(N3)}}{=} \|x\| + |-1|\|x - y\| = \|x\| + \|x - y\|.$$

Similarly, for the second inequality:

$$\|x\| = \|y + (x - y)\| \leq \|y\| + \|x - y\|.$$

7.3 PROOF THAT EXAMPLE 7.2 IS A NORMED VECTOR SPACE: the norm is generated by an inner product, so this follows from Theorem 7.1 and the discussion preceding it.
PROOF THAT EXAMPLE 7.3 IS A NORMED VECTOR SPACE: Let $x, y \in \mathbb{R}^n$ and $\alpha \in \mathbb{R}$.

(N1) $\|x\|_1 = \sum_{i=1}^{n} |x_i|$ is the sum of nonnegative terms, hence nonnegative.

(N2) $\|x\|_1 = \sum_{i=1}^{n} |x_i|$ is zero if and only if all of the nonnegative terms $|x_i|$ are zero, i.e. if and only if $x = \mathbf{0}$.

(N3) $\|\alpha x\|_1 = \sum_{i=1}^{n} |\alpha x_i| = \sum_{i=1}^{n} |\alpha||x_i| = |\alpha| \sum_{i=1}^{n} |x_i| = |\alpha|\|x\|_1$.

(N4) Using the triangle inequality for the absolute value (which in its turn is a consequence of the triangle inequality of the Euclidean norm on $\mathbb{R}^n$ with $n = 1$), we find

$$\|x + y\|_1 = \sum_{i=1}^{n} |x_i + y_i| \leq \sum_{i=1}^{n} \big(|x_i| + |y_i|\big) = \sum_{i=1}^{n} |x_i| + \sum_{i=1}^{n} |y_i| = \|x\|_1 + \|y\|_1.$$

PROOF THAT EXAMPLE 7.4 IS A NORMED VECTOR SPACE: Let $x, y \in \mathbb{R}^n$ and $\alpha \in \mathbb{R}$.

(N1) $\|x\|_\infty = \sup_i |x_i|$ is the supremum of n nonnegative terms, hence nonnegative.

(N2) $\|x\|_\infty = \sup_i |x_i|$ is zero if and only if $|x_i| = 0$ for all i, i.e. if and only if $x = \mathbf{0}$.

(N3) $\|\alpha x\|_\infty = \sup_i |\alpha x_i| = \sup_i |\alpha||x_i| = |\alpha| \sup_i |x_i| = |\alpha|\|x\|_\infty$.

(N4) For each i, $|x_i + y_i| \le |x_i| + |y_i| \le \sup_i |x_i| + \sup_i |y_i| = \|x\|_\infty + \|y\|_\infty$. Taking the supremum, $\|x + y\|_\infty \le \|x\|_\infty + \|y\|_\infty$.

PROOF THAT EXAMPLE 7.5 IS A NORMED VECTOR SPACE: Let $x, y \in \mathbb{R}^n$ and $\alpha \in \mathbb{R}$.

(N1) $\|x\|_p = \left(\sum_{i=1}^n |x_i|^p\right)^{1/p}$ is the p-th root of a nonnegative number, hence nonnegative.

(N2) $\|x\|_p = \left(\sum_{i=1}^n |x_i|^p\right)^{1/p} = 0$ if and only if each summand $|x_i|^p$ is zero, if and only if $x = \mathbf{0}$.

(N3) $\|\alpha x\|_p = \left(\sum_{i=1}^n |\alpha x_i|^p\right)^{1/p} = \left(\sum_{i=1}^n |\alpha|^p |x_i|^p\right)^{1/p} = \left(|\alpha|^p \sum_{i=1}^n |x_i|^p\right)^{1/p} = |\alpha|\|x\|_p$.

(N4) Exercise 7.6(c).

PROOF THAT EXAMPLE 7.6 IS A NORMED VECTOR SPACE: Let $x, y \in B(\mathbb{N})$ and let $\alpha \in \mathbb{R}$.

(N1) $x \in B(\mathbb{N})$ means that x is bounded: there is an $M \in \mathbb{R}$ with $|x_i| \le M$ for all $i \in \mathbb{N}$. So $\|x\|_\infty = \sup_{i\in\mathbb{N}} |x_i|$ is the supremum of a nonempty set of nonnegative numbers that is bounded from above by M. Hence, the supremum exists and is nonnegative.

Properties (N2) to (N4) follow as in Example 7.4.

PROOF THAT EXAMPLE 7.7 IS A NORMED VECTOR SPACE: Since the continuous function $x \mapsto |f(x)|$ achieves a maximum on the compact interval $[a, b]$, $\|f\|_\infty$ is well-defined for each $f \in C[a, b]$. (I am counting on you knowing this from earlier calculus courses. We'll prove it more generally in Theorem 17.3.) Let $f, g \in C[a, b]$ and $\alpha \in \mathbb{R}$.

(N1) $\|f\|_\infty = \max\{|f(x)| : x \in [a, b]\} \ge |f(a)| \ge 0$.

(N2) $\|f\|_\infty = \max\{|f(x)| : x \in [a, b]\} = 0$ if and only if $|f(x)| = 0$ for all $x \in [a, b]$. This means that $f(x) = 0$ for all $x \in [a, b]$, i.e., that $f = \mathbf{0}$ is the zero function.

(N3) $\|\alpha f\|_\infty = \max\{|\alpha f(x)| : x \in [a, b]\} = \max\{|\alpha||f(x)| : x \in [a, b]\} = |\alpha| \max\{|f(x)| : x \in [a, b]\} = |\alpha|\|f\|_\infty$.

(N4) Using the triangle inequality for the absolute value, we find for each $x \in [a, b]$:

$$|f(x) + g(x)| \le |f(x)| + |g(x)| \le \|f\|_\infty + \|g\|_\infty.$$

Taking the maximum over $x \in [a, b]$ gives $\|f + g\|_\infty \le \|f\|_\infty + \|g\|_\infty$.

PROOF THAT EXAMPLE 7.8 IS A NORMED VECTOR SPACE: Let $f, g \in C[a, b]$ and $\alpha \in \mathbb{R}$.

(N1) Since $|f(x)| \ge 0$ for all $x \in [a, b]$, integrating gives

$$\|f\|_1 = \int_a^b |f(x)|\,\mathrm{d}x \ge \int_a^b 0\,\mathrm{d}x = 0.$$

(N2) If $f = \mathbf{0}$, then $\|f\|_1 = \int_a^b 0\,\mathrm{d}x = 0$. If $f \ne \mathbf{0}$, then $|f(x)| > 0$ for some $x \in [a, b]$. By continuity, $|f(x)|$ is bounded away from zero on a subset of $[a, b]$, so $\|f\|_1 = \int_a^b |f(x)|\,\mathrm{d}x > 0$.

(N3) By linearity of the Riemann integral,

$$\|\alpha f\|_1 = \int_a^b |\alpha f(x)|\,\mathrm{d}x = \int_a^b |\alpha||f(x)|\,\mathrm{d}x = |\alpha| \int_a^b |f(x)|\,\mathrm{d}x = |\alpha|\|f\|_1.$$

(N4) Using the triangle inequality for the absolute value, we find for each $x \in [a, b]$ that

$$0 \le |f(x) + g(x)| \le |f(x)| + |g(x)|.$$

Integrating over $[a, b]$ and using linearity of the integral gives

$$\|f + g\|_1 = \int_a^b |f(x) + g(x)|\,\mathrm{d}x \le \int_a^b |f(x)| + |g(x)|\,\mathrm{d}x = \int_a^b |f(x)|\,\mathrm{d}x + \int_a^b |g(x)|\,\mathrm{d}x = \|f\|_1 + \|g\|_1.$$

7.4 We have

$$\|x+y\|^2 = \langle x+y, x+y\rangle = \langle x,x\rangle + \langle x,y\rangle + \langle y,x\rangle + \langle y,y\rangle$$
$$\|x-y\|^2 = \langle x-y, x-y\rangle = \langle x,x\rangle - \langle x,y\rangle - \langle y,x\rangle + \langle y,y\rangle.$$

Subtracting the second equation from the first gives the polarization identity; adding them gives the parallelogram law.

7.5 (a) Assume that one of x and y is a multiple of the other. Without loss of generality (the other case is similar), suppose that $x = cy$ for some scalar c. Then the left and right side of Cauchy-Schwarz become

$$|\langle x,y\rangle| = |\langle cy,y\rangle| = |c\langle y,y\rangle| = |c|\,\|y\|^2,$$
$$\|x\|\,\|y\| = \|cy\|\,\|y\| = |c|\,\|y\|\,\|y\| = |c|\,\|y\|^2,$$

so they are equal.

Conversely, assume that the Cauchy-Schwarz inequality holds with equality: $|\langle x,y\rangle| = \|x\|\,\|y\|$. If $y = \mathbf{0}$, this is certainly the case (both sides are zero) and $y = 0x$ shows that y is a multiple of x. So suppose that $y \neq \mathbf{0}$. Then taking squares and rewriting gives

$$\langle x,y\rangle^2 = \|x\|^2\|y\|^2 \qquad \Longrightarrow \qquad \|x\|^2 - \frac{\langle x,y\rangle^2}{\|y\|^2} = 0.$$

In the proof of the Cauchy-Schwarz inequality, we established the equality

$$\|x-\alpha y\|^2 = \|x\|^2 - \frac{\langle x,y\rangle^2}{\|y\|^2}$$

for a certain choice of α. And we just argued that the right term and consequently also $\|x-\alpha y\|^2$ equals zero. Hence $x - \alpha y = \mathbf{0}$, making x a multiple of y.

(b) Assume that one of x and y is a nonnegative multiple of the other. Without loss of generality (the other case is similar), suppose that $x = cy$ for some nonnegative scalar c. Then

$$\|x+y\| = \|cy+y\| = \|(c+1)y\| = (c+1)\|y\| = c\|y\| + \|y\| = \|x\| + \|y\|,$$

so the triangle inequality holds with equality.

Conversely, assume that the triangle inequality holds with equality: $\|x+y\| = \|x\| + \|y\|$. If $y = \mathbf{0}$, this is certainly the case (both sides equal $\|x\|$) and $y = 0x$ shows that y is a nonnegative multiple of x. So suppose that $y \neq \mathbf{0}$.

The proof of the triangle inequality used the chain of (in)equalities

$$\begin{aligned}\|x+y\|^2 &= \langle x+y, x+y\rangle = \langle x,x\rangle + \langle y,x\rangle + \langle x,y\rangle + \langle y,y\rangle \\ &\leq \langle x,x\rangle + 2|\langle x,y\rangle| + \langle y,y\rangle \leq \|x\|^2 + 2\|x\|\|y\| + \|y\|^2 = (\|x\| + \|y\|)^2.\end{aligned}$$

There are two weak inequalities in this derivation; equality holds if and only if both $\langle x,y\rangle = |\langle x,y\rangle|$ and Cauchy-Schwarz holds with equality. The first means that $\langle x,y\rangle \geq 0$. The second means that $x = \alpha y$ with $\alpha = \langle x,y\rangle / \langle y,y\rangle$ as in the proof of Cauchy-Schwarz. Combining both, we find that $x = \alpha y$ with $\alpha = \langle x,y\rangle / \langle y,y\rangle \geq 0$: x is a nonnegative multiple of y.

7.6 Our method of proof is from G.H. Woeginger (2009) "When Cauchy and Hölder met Minkowski: a tour through well-known inequalities", Math. Magazine 82, 202–207.

(a) For $n = 1$, the inequality becomes $f(x_i, y_i) \leq f(x_i, y_i)$ which is trivially true. Now let $n \in \mathbb{N}$ and assume the inequality holds for sums of n terms. Let's prove it is true for $n+1$ terms. The first inequality below is the induction hypothesis, the second comes from concavity of g using $\lambda = \sum_{i=1}^{n} y_i / \sum_{i=1}^{n+1} y_i \in (0,1)$:

$$\begin{aligned}
\sum_{i=1}^{n+1} f(x_i, y_i) &= \sum_{i=1}^{n} f(x_i, y_i) + f(x_{n+1}, y_{n+1}) \\
&\leq f\left(\sum_{i=1}^{n} x_i, \sum_{i+1}^{n} y_i\right) + f(x_{n+1}, y_{n+1}) \\
&= \sum_{i=1}^{n} y_i \cdot g\left(\frac{\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} y_i}\right) + y_{n+1} \cdot g\left(\frac{x_{n+1}}{y_{n+1}}\right) \\
&= \left(\sum_{i=1}^{n+1} y_i\right) \cdot \left[\lambda \cdot g\left(\frac{\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} y_i}\right) + (1-\lambda) \cdot g\left(\frac{x_{n+1}}{y_{n+1}}\right)\right] \\
&\leq \left(\sum_{i=1}^{n+1} y_i\right) \cdot \left[g\left(\lambda \frac{\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} y_i} + (1-\lambda)\frac{x_{n+1}}{y_{n+1}}\right)\right] \\
&= \left(\sum_{i=1}^{n+1} y_i\right) \cdot g\left(\frac{\sum_{i=1}^{n+1} x_i}{\sum_{i=1}^{n+1} y_i}\right) \\
&= f\left(\sum_{i=1}^{n+1} x_i, \sum_{i=1}^{n+1} y_i\right).
\end{aligned}$$

(b) If $g(x) = x^{1/p}$, then $f(x,y) = y \cdot \left(\frac{x}{y}\right)^{1/p} = x^{1/p} y^{1-1/p} = x^{1/p} y^{1/q}$ and (a) gives

$$\sum_{i=1}^{n} x_i^{1/p} y_i^{1/q} \leq \left(\sum_{i=1}^{n} x_i\right)^{1/p} \left(\sum_{i=1}^{n} y_i\right)^{1/q}$$

whenever $x_1, \ldots, x_n, y_1, \ldots, y_n > 0$. Now choose $x, y \in \mathbb{R}^n$ arbitrarily. We may ignore coordinates equal to zero without loss of generality and substitute $|x_i|^p > 0$ for x_i and $|y_i|^q > 0$ for y_i to find

$$\sum_{i=1}^{n} |x_i||y_i| \leq \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p} \left(\sum_{i=1}^{n} |y_i|^q\right)^{1/q} = \|x\|_p \|y\|_q.$$

(c) If $g(x) = (x^{1/p} + 1)^p$, then $f(x,y) = y \cdot \left(\left(\frac{x}{y}\right)^{1/p} + 1\right)^p = (x^{1/p} + y^{1/p})^p$ and (a) gives

$$\sum_{i=1}^{n} (x_i^{1/p} + y_i^{1/p})^p \leq \left(\left(\sum_{i=1}^{n} x_i\right)^{1/p} + \left(\sum_{i=1}^{n} y_i\right)^{1/p}\right)^p$$

whenever $x_1, \ldots, x_n, y_1, \ldots, y_n > 0$. Now choose $x, y \in \mathbb{R}^n$ arbitrarily. We may ignore coordinates equal to zero without loss of generality and substitute $|x_i|^p > 0$ for x_i and $|y_i|^p > 0$ for y_i to find

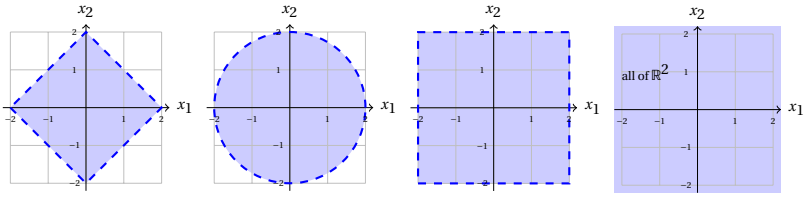
$$\sum_{i=1}^{n} (|x_i| + |y_i|)^p \leq \left(\left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p} + \left(\sum_{i=1}^{n} |y_i|^p\right)^{1/p}\right)^p.$$

According to the triangle inequality for the absolute value, $|x_i + y_i| \leq |x_i| + |y_i|$, so

$$\sum_{i=1}^{n} |x_i + y_i|^p \leq \left(\left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p} + \left(\sum_{i=1}^{n} |y_i|^p\right)^{1/p}\right)^p.$$

Taking p-th roots on both sides gives $\|x + y\|_p \leq \|x\|_p + \|y\|_p$.

8.1 The requested pictures are:



d_1 d_2 d_∞ discrete

8.2

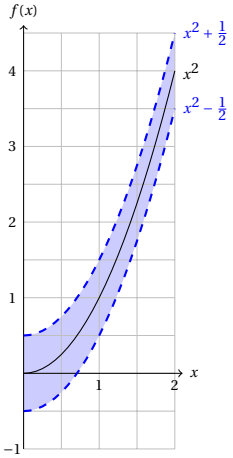
$$d_1(x,y) = |x_1 - y_1| + |x_2 - y_2| + |x_3 - y_3| = |1-2| + |0-6| + |4-2| = 1+6+2 = 9,$$

$$d_2(x,y) = \sqrt{(x_1-y_1)^2 + (x_2-y_2)^2 + (x_3-y_3)^2} = \sqrt{(1-2)^2 + (0-6)^2 + (4-2)^2} = \sqrt{1+36+4} = \sqrt{41},$$

$$d_\infty(x,y) = \max\{|x_1-y_1|, |x_2-y_2|, |x_3-y_3|\} = \max\{|1-2|, |0-6|, |4-2|\} = \max\{1,6,2\} = 6,$$

$$d_{\text{discrete}}(x,y) = 1 \text{ since } x \neq y.$$

8.3 All functions with a graph in the shaded area:



8.4 (a) $d_H(x,y) = 3$ because x and y differ in all three coordinates.

(b) The number of elements in a finite set S is often denoted by $|S|$. So $d_H(x,y) = |\{i \in \{1,\dots,n\} : x_i \neq y_i\}|$. Let $x, y, z \in \mathbb{R}^n$.

(D1) $d_H(x,y) = |\{i \in \{1,\dots,n\} : x_i \neq y_i\}| \in \{0,1,\dots,n\}$, so $d_H(x,y) \geq 0$.

(D2) $d_H(x,y) = |\{i \in \{1,\dots,n\} : x_i \neq y_i\}| = 0$ if and only if $x_i = y_i$ for all i, i.e., if and only if $x = y$.

(D3) $d_H(x,y) = |\{i \in \{1,\dots,n\} : x_i \neq y_i\}| = |\{i \in \{1,\dots,n\} : y_i \neq x_i\}| = d_H(y,x)$.

(D4) To verify that $d_H(x,z) \leq d_H(x,y) + d_H(y,z)$, note that for each i with $x_i \neq z_i$, it follows that $x_i \neq y_i$ or $y_i \neq z_i$, or both:

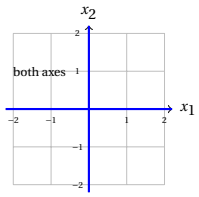
$$\{i : x_i \neq z_i\} \subseteq \{i : x_i \neq y_i\} \cup \{i : y_i \neq z_i\}.$$

Hence,

$$d_H(x,z) = |\{i : x_i \neq z_i\}| \leq |\{i : x_i \neq y_i\} \cup \{i : y_i \neq z_i\}| \leq |\{i : x_i \neq y_i\}| + |\{i : y_i \neq z_i\}| = d_H(x,y) + d_H(y,z).$$

(c) To see that d_H is not generated by a norm, note that $d_H(\alpha x, \alpha y) = d_H(x,y)$ for all $\alpha \neq 0$. But a distance d generated by a norm $\|\cdot\|$ satisfies $d(\alpha x, \alpha y) = \|\alpha x - \alpha y\| = \|\alpha(x-y)\| \overset{\text{(N3)}}{=} |\alpha|\|x-y\| = \alpha d(x,y)$.

(d)



Hamming

8.5 If a function $d : X \times X \to \mathbb{R}$ on a nonempty set X satisfies (D2) to (D4), it also satisfies (D1): for all $x, y \in X$,

$$0 \overset{(D2)}{=} d(x,x) \overset{(D4)}{\leq} d(x,y) + d(y,x) \overset{(D3)}{=} 2d(x,y).$$

And $2d(x,y) \geq 0$ gives $d(x,y) \geq 0$.

8.6 A violation of symmetry (D3) is probably easiest to imagine: in a city with some one-way streets, the distance you need to travel from x to reach y may well be different from the distance you need to travel from y to reach x. Think, for instance, of a circular bypass around a city on which you are only allowed to drive in clockwise direction.

For a violation of (D2), imagine a public transportation system where the buss stops in a city are divided into three zones (zone 1, 2, and 3); travel within a zone is free, travel between distinct zones costs one unit for each intermediate zone. The function that assigns to each pair of buss stops its travel costs — that is, $d(x,y) = |\text{zone}(x) - \text{zone}(y)|$ where $\text{zone}(x)$ denotes the zone to which buss stop x belongs — satisfies all properties of a distance function except (D2): the travel cost between distinct buss stops in the same zone is zero.

For a violation of the triangle inequality (D4), imagine three houses. There is a path of one kilometer between houses x and y and between houses y and z. But there happens to be a lake between house x and z so that the path between those houses is circuitous and takes five kilometers. If we define the distance between two houses to be the length of the path that connects them, the triangle inequality is violated: the path between x and z is five kilometers, which is longer than the length of the path from x to y and then from y to z (each one kilometer).

In each scenario I pointed out which property was violated. Try to convince yourself that the other properties are satisfied.

8.7 PROOF THAT d' IS A METRIC: Let $x, y, z \in V$.

(D1) $d(x,y) \geq 0$ by (D1) for metric d, and $1 \geq 0$, so $d'(x,y) = \min\{d(x,y), 1\} \geq 0$.

(D2) $d'(x,y) = \min\{d(x,y), 1\} = 0$ if and only if $d(x,y) = 0$, which by (D2) for metric d is true if and only if $x = y$.

(D3) By (D3) for metric d, $d'(x,y) = \min\{d(x,y), 1\} = \min\{d(y,x), 1\} = d'(y,x)$.

(D4) To show:

$$\underbrace{\min\{d(x,z), 1\}}_{=d'(x,z)} \leq \underbrace{\min\{d(x,y), 1\}}_{=d'(x,y)} + \underbrace{\min\{d(y,z), 1\}}_{=d'(y,z)}.$$

Case 1: if $\min\{d(x,y), 1\} + \min\{d(y,z), 1\} \leq 1$, then $d'(x,y) = d(x,y) \leq 1$ and $d'(y,z) = d(y,z) \leq 1$. By the triangle inequality for metric d, it follows that

$$d'(x,z) \leq d(x,z) \leq d(x,y) + d(y,z) = d'(x,y) + d'(y,z).$$

Case 2: if $\min\{d(x,y), 1\} + \min\{d(y,z), 1\} > 1$, then

$$d'(x,z) \leq 1 < \min\{d(x,y), 1\} + \min\{d(y,z), 1\} = d'(x,y) + d'(y,z).$$

PROOF THAT d'' IS A METRIC: Let $x, y, z \in V$.

(D1) $d''(x,y)$ is the fraction of two nonnegative numbers by (D1) for metric d, hence nonnegative.

(D2) $d''(x,y) = 0$ if and only if $d(x,y) = 0$, which by (D2) for metric d is equivalent with $x = y$.

(D3) By (D3) for metric d: $d''(x,y) = d(x,y)/(1 + d(x,y)) = d(y,x)/(1 + d(y,x)) = d''(y,x)$.

(D4) To show:

$$\underbrace{\frac{d(x,z)}{1+d(x,z)}}_{=d''(x,z)} \leq \underbrace{\frac{d(x,y)}{1+d(x,y)}}_{=d''(x,y)} + \underbrace{\frac{d(y,z)}{1+d(y,z)}}_{=d''(y,z)}.$$

METHOD 1: The function f with $f(t) = t/(1+t)$ is increasing on $[0,\infty)$. Using the triangle inequality for metric d and its nonnegativity, we find:

$$\frac{d(x,z)}{1+d(x,z)} \leq \frac{d(x,y)+d(y,z)}{1+d(x,y)+d(y,z)} = \frac{d(x,y)}{1+d(x,y)+d(y,z)} + \frac{d(y,z)}{1+d(x,y)+d(y,z)} \leq \frac{d(x,y)}{1+d(x,y)} + \frac{d(y,z)}{1+d(y,z)}.$$

METHOD 2: Adding 1 to both sides of the triangle inequality for metric d, we know that

$$1 + d(x,z) \leq 1 + d(x,y) + d(y,z).$$

Hence,

$$\begin{aligned}\frac{d(x,z)}{1+d(x,z)} &= 1 - \frac{1}{1+d(x,z)} \leq 1 - \frac{1}{1+d(x,y)+d(y,z)} \\ &= \frac{d(x,y)+d(y,z)}{1+d(x,y)+d(y,z)} = \frac{d(x,y)}{1+d(x,y)+d(y,z)} + \frac{d(y,z)}{1+d(x,y)+d(y,z)} \\ &\leq \frac{d(x,y)}{1+d(x,y)} + \frac{d(y,z)}{1+d(y,z)}.\end{aligned}$$

8.8 Let $x, y, z \in X$. Rewriting the absolute value, we need to prove:

$$-d(x,z) \leq d(x,y) - d(y,z) \leq d(x,z).$$

The first inequality follows from

$$d(y,z) \overset{(I4)}{\leq} d(y,x) + d(x,z) \overset{(I3)}{=} d(x,y) + d(x,z),$$

and the second from

$$d(x,y) \overset{(I4)}{\leq} d(x,z) + d(z,y) \overset{(I3)}{=} d(x,z) + d(y,z).$$

8.9 (a) For all $x, y, z \in V$: $d(x+z, y+z) = \|(x+z) - (y+z)\| = \|x - y\| = d(x,y)$.

(b) For all $x, y \in V$ and all $\alpha \in \mathbb{R}$: $d(\alpha x, \alpha y) = \|\alpha x - \alpha y\| = \|\alpha(x-y)\| \overset{(N3)}{=} |\alpha|\|x-y\| = |\alpha| d(x,y)$.

(c) Using properties (D1) to (D4) of the metric d and the additional properties above, we show that $\|\cdot\|$ satisfies the properties of Definition 7.1:

(N1) For each $x \in V$: $\|x\| = d(x, \mathbf{0}) \overset{(D1)}{\geq} 0$.

(N2) $\|x\| = d(x, \mathbf{0}) = 0$ if and only if $x = \mathbf{0}$ by property (D2) of metric d.

(N3) For all $x \in V$ and $\alpha \in \mathbb{R}$: $\|\alpha x\| = d(\alpha x, \mathbf{0}) = d(\alpha x, \alpha \mathbf{0}) \overset{(b)}{=} |\alpha| d(x, \mathbf{0}) = |\alpha| \|x\|$.

(N4) For all $x, y \in V$ we have

$$\|x+y\| = d(x+y, \mathbf{0}) \overset{(D4)}{\leq} d(x+y, x) + d(x, \mathbf{0}) \overset{(a)}{=} d(y, \mathbf{0}) + d(x, \mathbf{0}) = \|x\| + \|y\|.$$

(d) For all $x, y \in V$: $\|x - y\| = d(x-y, \mathbf{0}) \overset{(a)}{=} d((x-y)+y, \mathbf{0}+y) = d(x,y)$.

8.10 (a) The intersection of your purchases with those of customer 1 is

{`Lila`, `Infinite jest`, `Tourmaline`},

with three elements and the union is

{`Lila`, `The blazing world`, `Infinite jest`, `Ways of going home`, `Tourmaline`, `Freshwater`},

with six elements, so the Jaccard distance to customer 1 is $1 - \frac{3}{6} = \frac{1}{2}$. Analogously, the Jaccard distance to customers 2 and 3 is $1 - \frac{2}{7} = \frac{5}{7}$ and $1 - \frac{3}{7} = \frac{4}{7}$, respectively. So customer 1 is nearest and this customer's purchase of `Freshwater` is likely to show up among your personal recommendations.

(b) Let A, B, and C be nonempty subsets of X. We prove the four properties of a metric:

(D1) $d(A,A) = 1 - \frac{|A\cap A|}{|A\cup A|} = 1 - \frac{|A|}{|A|} = 0$.

(D2) If $A = B$, then $A \cap B = A \cup B = A$, so $d(A,B) = 1 - \frac{|A\cap B|}{|A\cup B|} = 1 - \frac{|A|}{|A|} = 0$.

Conversely, if $d(A,B) = 1 - \frac{|A\cap B|}{|A\cup B|} = 0$, it follows that $|A \cap B| = |A \cup B|$. But $A \cap B$ is a subset of $A \cup B$; the only way they can have the same number of elements is if they are the same: $A \cap B = A \cup B$. This, in its turn, implies that $A = B$.

(D3) Since $A \cap B = B \cap A$ and $A \cup B = B \cup A$, it follows that $d(A,B) = 1 - \frac{|A\cap B|}{|A\cup B|} = 1 - \frac{|B\cap A|}{|B\cup A|} = d(B,A)$.

(D4) We follow the hint in the exercise and argue, one line at a time, why

$$d(A,B) + d(B,C) = \frac{|A\cup B| - |A\cap B|}{|A\cup B|} + \frac{|B\cup C| - |B\cap C|}{|B\cup C|} \tag{157}$$

$$\geq \frac{|T_1| + |T_2|}{|A\cup B\cup C|} + \frac{|T_2| + |T_3|}{|A\cup B\cup C|} \tag{158}$$

$$\geq \frac{|T_1| + |T_2| + |T_3|}{|A\cup B\cup C|} \tag{159}$$

$$= 1 - \frac{|V|}{|A\cup B\cup C|} \tag{160}$$

$$\geq 1 - \frac{|A\cap C|}{|A\cup C|} \tag{161}$$

$$= d(A,C). \tag{162}$$

☒ Equalities (157) and (162) holds by definition.

☒ By definition of the sets T_1, T_2, and T_3 (see the figure in the exercise),

$$|A\cup B| - |A\cap B| = |T_1| + |T_2| \qquad \text{and} \qquad |B\cup C| - |B\cap C| = |T_2| + |T_3|$$

and clearly

$$|A\cup B| \leq |A\cup B\cup C| \qquad \text{and} \qquad |B\cup C| \leq |A\cup B\cup C|.$$

So in going from (157) to (158), the numerators of the fractions have been kept the same, but the denominators have increased. And dividing by a larger number gives a smaller number.

☒ (158) is at least as large as (159) because

$$\frac{|T_1| + |T_2|}{|A\cup B\cup C|} + \frac{|T_2| + |T_3|}{|A\cup B\cup C|} = \frac{|T_1| + 2|T_2| + |T_3|}{|A\cup B\cup C|} \geq \frac{|T_1| + |T_2| + |T_3|}{|A\cup B\cup C|}.$$

☒ (159) equals (160) because $A \cup B \cup C = T_1 \cup T_2 \cup T_3 \cup V$ and all four sets T_1, T_2, T_3, and V are disjoint, so $|T_1| + |T_2| + |T_3| = |A \cup B \cup C| - |V|$.

☒ (160) is at least as large as (161), because $V \subseteq A \cap C$ and $A \cup C \subseteq A \cup B \cup C$, so we have (weakly) increased the numerator and decreased the denominator, leading to a larger fraction.

The proof of the triangle inequality is adapted from G. Gilbert (1972), "Distance between sets", *Nature* 239, p. 174. The titles of the books were not chosen at random: they are books that I like a lot.

9.1

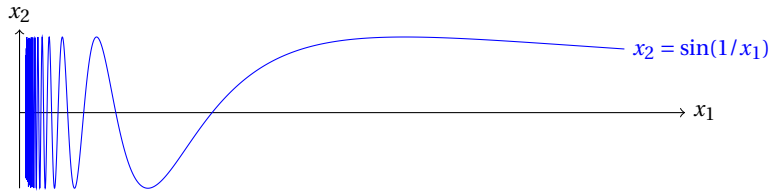
U	$\text{int}(U)$	$\text{cl}(U)$	$\text{bd}(U)$	$\text{acc}(U)$
$\{x : x_1 > 0\}$	$\{x : x_1 > 0\}$	$\{x : x_1 \geq 0\}$	$\{x : x_1 = 0\}$	$\{x : x_1 \geq 0\}$
$\{x : x_1^2 + x_2^2 = 4\}$	$\emptyset$	U	U	U
$\{x : x_1 \leq x_2\}$	$\{x : x_1 < x_2\}$	$\{x : x_1 \leq x_2\}$	$\{x : x_1 = x_2\}$	$\{x : x_1 \leq x_2\}$
$\{x : x_1 > 0, x_2 = \sin(1/x_1)\}$	$\emptyset$	$U \cup \{(0,x_2) : -1 \leq x_2 \leq 1\}$	$= \text{cl}(U)$	$= \text{cl}(U)$
$\{x : x_1 x_2 \in \mathbb{Q}\}$	$\emptyset$	$\mathbb{R}^2$	$\mathbb{R}^2$	$\mathbb{R}^2$

Find appropriate motivations yourself. This might help:

☒ For the set $\{x : x_1 > 0, x_2 = \sin(1/x_1)\}$, recall that the sine function is periodic with period 2π. Moreover, for each $k \in \mathbb{N}$, $\frac{1}{x_1} = k(2\pi)$ if and only if $x_1 = \frac{1}{k(2\pi)}$. So if you sketch the set, you will have to squeeze a sine curve on each interval with x_1 in

$$\left[\frac{1}{2(2\pi)}, \frac{1}{2\pi}\right] \text{ and } \left[\frac{1}{3(2\pi)}, \frac{1}{2(2\pi)}\right] \text{ and } \left[\frac{1}{4(2\pi)}, \frac{1}{3(2\pi)}\right] \text{ and...}$$

Intuitively, there are more and more sine curves, the closer x_1 gets to zero. In particular, for each $\varepsilon > 0$ there are infinitely many sine curves for $x_1 \in (0, \varepsilon)$. Here is a sketch of the set:



☒ For the set $\{x : x_1 x_2 \in \mathbb{Q}\}$ of vectors (x_1, x_2) where the product of the coordinates is rational, note:

1. If r is a real number, you can find a rational number arbitrarily nearby by truncating the decimal representation of r. For instance, $\pi = 3.1415926\cdots$, so truncating after 5 decimal places gives a rational number $\frac{314159}{10^5}$ within distance 10^{-5} of π.
2. If r is a rational number, you can find a irrational number arbitrarily nearby: $\sqrt{2}$ is irrational, so $\frac{\sqrt{2}}{n}$ is irrational for each $n \in \mathbb{N}$. Hence, $r + \frac{\sqrt{2}}{n}$ is irrational and you can make it arbitrarily close to r by choosing n sufficiently large.

To show, for instance, that the interior of $U = \{x : x_1 x_2 \in \mathbb{Q}\}$ is empty, let $x \in U$ and $\varepsilon > 0$. I construct a point in $B(x, \varepsilon)$ and U^c, showing that x is not an interior point of U.

CASE 1: $x_1 x_2 = q \in \mathbb{Q}$ for some $q \neq 0$. Then $x_1 \neq 0$. I construct a point $(x_1, x_2 + \delta)$ such that

1. its distance $|\delta|$ to x is less than ε,
2. $x_1(x_2 + \delta) = x_1 x_2 + \delta x_1 = q + \delta x_1 \notin \mathbb{Q}$.

The idea is to write δx_1 in the form $\frac{\sqrt{2}}{n}$, i.e., take $\delta = \frac{\sqrt{2}}{n x_1}$, and choose n so large that $|\delta| < \varepsilon$:

$$|\delta| = \left|\frac{\sqrt{2}}{n x_1}\right| = \frac{\sqrt{2}}{n|x_1|} < \varepsilon \quad \Leftrightarrow \quad n > \frac{\sqrt{2}}{\varepsilon |x_1|}.$$

Conclude: if $n > \frac{\sqrt{2}}{\varepsilon|x_1|}$, point $\left(x_1, x_2 + \frac{\sqrt{2}}{n x_1}\right)$ lies in the ε-ball around x but not in U.

CASE 2: $x_1 x_2 = 0$. Reasoning as above gives: if $x = (0,0)$, take $n \in \mathbb{N}$, $n > \frac{2\sqrt{2}}{\varepsilon^2}$. Point $\left(\frac{\sqrt[4]{2}}{\sqrt{n}}, \frac{\sqrt[4]{2}}{\sqrt{n}}\right) \in B(x, \varepsilon) \cap U^c$ shows that $x = (0,0)$ is not an interior point of U. If only one coordinate, say x_1, of x is zero, take $n \in \mathbb{N}$, $n > \frac{\sqrt{2}}{\varepsilon|x_2|}$. Point $\left(\frac{\sqrt{2}}{n x_2}, x_2\right) \in B(x, \varepsilon) \cap U^c$ shows that $x = (0, x_2)$ is not an interior point of U.

9.2 (a) All functions f_k are isolated points. For each $x \in [0,1]$:

$$f_1(x) \geq f_2(x) \geq f_3(x) \geq \cdots,$$

with strict inequality for $x \in (0,1)$. So for $k, \ell \in \mathbb{N}$:

$$d_\infty(f_k, f_\ell) \geq \begin{cases} d_\infty(f_k, f_{k+1}) & \text{if } \ell > k, \\ d_\infty(f_k, f_{k-1}) & \text{if } \ell < k. \end{cases}$$

Hence, if we take $0 < \varepsilon < \min\{d_\infty(f_k, f_{k+1}), d_\infty(f_k, f_{k-1})\}$, then $B(f_k, \varepsilon)$ cannot contain other f_ℓ.

There are no accumulation points either. Consider any $f \in C[0,1]$.

CASE 1: if $f(x) < 0$ for some $x \in (0,1)$, f can't be an accumulation point: the ball around f with radius $\varepsilon = -f(x) > 0$ contains none of the functions f_n, since

$$d_\infty(f_n, f) \geq f_n(x) - f(x) = x^n - f(x) > 0 - f(x) = \varepsilon.$$

CASE 2: if $f(x) > 0$ for some $x \in (0,1)$, f can't be an accumulation point: the ball around f with radius $\varepsilon = \frac{1}{2} f(x) > 0$ contains only finitely many of the functions f_n, since $f_n \in B(f, \varepsilon)$ implies that

$$f(x) - f_n(x) \leq d_\infty(f, f_n) < \varepsilon = \frac{1}{2} f(x).$$

Rewriting and using that $\ln x < 0$ if $x \in (0,1)$ gives:

$$\begin{aligned} f(x) - f_n(x) < \frac{1}{2} f(x) &\Leftrightarrow \frac{1}{2} f(x) < x^n \\ &\Leftrightarrow \ln\left(\frac{1}{2} f(x)\right) < n \ln x \\ &\Leftrightarrow n < \frac{\ln\left(\frac{1}{2} f(x)\right)}{\ln x}. \end{aligned}$$

Evidently, only finitely many $n \in \mathbb{N}$ satisfy this inequality, so f cannot be an accumulation point by Example 9.5.

CASE 3: if $f(x) = 0$ for all $x \in (0,1)$, f can't be an accumulation point: by continuity of f, also $f(1) = 0$. But $f_n(1) = 1^n = 1$ for all $n \in \mathbb{N}$, so the distance between f and f_n is at least one: the ball around f with radius $0 < \varepsilon < 1$ contains none of the functions f_n.

These three cases are exhaustive: no $f \in C[0,1]$ is an accumulation point.

(b) For distinct $k, \ell \in \mathbb{N}$,

$$d_\infty(f_k, f_\ell) \geq |f_k(1) - f_\ell(1)| = |k - \ell| \geq 1,$$

making distinct f_k and f_ℓ at least distance one apart. Hence all f_k are isolated: a ball around them with radius $0 < \varepsilon < 1$ contains only f_k. Similarly, there are no accumulation points: by the triangle inequality, a ball $B(f, 1/2)$ around any $f \in C[0,1]$ contains at most one function f_k. But then f can't be an accumulation point by Example 9.5.

(c) Reasoning as in (a), each f_k is an isolated point of $\{f_1, f_2, f_3, \ldots\}$.

Function $f \in C[0,1]$ with $f(x) = 0$ for all $x \in [0,1]$ is an accumulation point of $\{f_n : n \in \mathbb{N}\}$:

Let $\varepsilon > 0$. Choose $n \in \mathbb{N}$ with $n > 1/\varepsilon$. Then $f_n \in B(f, \varepsilon)$:

$$d_\infty(f_n, f) = \sup_{x \in [0,1]} |f_n(x) - f(x)| = \sup_{x \in [0,1]} \left| \frac{x}{n} - 0 \right| = \frac{1}{n} < \varepsilon.$$

Since $f \notin \{f_n : n \in \mathbb{N}\}$, it follows that

$$\big(B(f, \varepsilon) \setminus \{f\}\big) \cap \{f_n : n \in \mathbb{N}\} \neq \emptyset,$$

proving that f is an accumulation point.

Arguing as in cases 1 and 2 of the answer above, it follows that f is the *only* accumulation point.

9.4 (a) METHOD 1: Using Theorem 9.4:

$$\begin{aligned} x \in \mathrm{bd}(U) &\overset{\text{Thm 9.4(e)}}{\Leftrightarrow} && \forall \varepsilon > 0: \quad B(x, \varepsilon) \cap U \neq \emptyset \text{ and } B(x, \varepsilon) \cap U^c \neq \emptyset, \\ &\overset{\text{Thm 9.4(a),(b)}}{\Leftrightarrow} && x \in \mathrm{cl}(U), x \notin \mathrm{int}(U) \\ &\overset{\text{def}}{\Leftrightarrow} && x \in \mathrm{cl}(U) \setminus \mathrm{int}(U). \end{aligned}$$

METHOD 2: Using Exercise 9.9:

$$\mathrm{bd}(U) \overset{\text{def}}{=} \mathrm{cl}(U) \cap \mathrm{cl}(U^c) \overset{\text{Exc 9.9(b)}}{=} \mathrm{cl}(U) \cap \mathrm{int}(U)^c \overset{\text{def}}{=} \mathrm{cl}(U) \setminus \mathrm{int}(U).$$

(b) We show that $\text{acc}(U)^c$ is open. Let $v \in \text{acc}(U)^c$. Then there is an $\varepsilon > 0$ with $(B(v,\varepsilon) \setminus \{v\}) \cap U = \emptyset$. But then v is an interior point of $\text{acc}(U)^c$, because all points in $B(v,\varepsilon)$ lie in $\text{acc}(U)^c$: for each $w \in B(v,\varepsilon)$ with $w \neq v$, choose $0 < \delta < \min\{\varepsilon - d(v,w), d(v,w)\}$. By example 9.2, $B(w,\delta) \subseteq B(v,\varepsilon)$. Since $\delta < d(v,w)$: $v \notin B(w,\delta)$. Hence $(B(w,\delta) \setminus \{w\}) \cap U \subseteq (B(v,\varepsilon) \setminus \{v\}) \cap U = \emptyset$, i.e., $w \in \text{acc}(U)^c$.

9.4 Call the metric space (X,d). The empty set has zero elements; it is closed by Theorem 9.3. Next, consider a subset $\{x_1,\ldots,x_n\}$ of $n \in \mathbb{N}$ elements. Why is it closed?
METHOD 1: By Example 9.3, a set of one element is closed. Since the union of finitely many closed sets is closed (Theorem 9.3), it follows that each finite subset of X is closed.
METHOD 2: A bit more directly: let's show that the complement $\{x_1,\ldots,x_n\}^c$ is open. If $x \in \{x_1,\ldots,x_n\}^c$, it follows that $d(x,x_i) > 0$ for all $i = 1,\ldots,n$. Take $\varepsilon = \min\{d(x,x_1),\ldots,d(x,x_n)\} > 0$. Then the ball around x with radius ε contains none of the points $x_1,\ldots,x_n$, making x an interior point of $\{x_1,\ldots,x_n\}^c$.

9.5 (a) By Theorem 9.1, the complement $V \setminus \emptyset = V$ of the empty set and $V \setminus V = \emptyset$ of V are open.

(b) By De Morgan's Laws, the complement of the intersection of arbitrarily many closed sets $\{U_i : i \in I\}$ is $\left(\cap_{i\in I} U_i\right)^c = \cup_{i\in I} U_i^c$, the union of open sets U_i^c, hence open by Theorem 9.1.

(c) By De Morgan's Laws, the complement of the union of finitely many closed sets $\{U_i : i \in I\}$ is $\left(\cup_{i\in I} U_i\right)^c = \cap_{i\in I} U_i^c$, the intersection of finitely many open sets U_i^c, hence open by Theorem 9.1.

9.6 (a) $\emptyset$ is a closed set and $\text{cl}(\emptyset)$ is the smallest closed set containing $\emptyset$, so $\text{cl}(\emptyset) = \emptyset$.

(b) $\text{cl}(U)$ is a closed set containing U, so $U \subseteq \text{cl}(U)$.

(c) Since $\text{cl}(U)$ is closed, it is the smallest closed set containing itself: $\text{cl}(\text{cl}(U)) = \text{cl}(U)$.

(d) By (b), $A \subseteq B \subseteq \text{cl}(B)$, so $\text{cl}(B)$ is a closed set containing A. Set $\text{cl}(A)$ is the smallest closed set containing A, so $\text{cl}(A) \subseteq \text{cl}(B)$.

(e) Since $A \subseteq A \cup B$, (d) gives $\text{cl}(A) \subseteq \text{cl}(A \cup B)$. Likewise, $\text{cl}(B) \subseteq \text{cl}(A \cup B)$. So $\text{cl}(A) \cup \text{cl}(B) \subseteq \text{cl}(A \cup B)$. For the converse inclusion, (b) gives $A \subseteq \text{cl}(A)$ and $B \subseteq \text{cl}(B)$, so $A \cup B \subseteq \text{cl}(A) \cup \text{cl}(B)$. As the union of two closed sets, $\text{cl}(A) \cup \text{cl}(B)$ is a closed set containing $A \cup B$. Set $\text{cl}(A \cup B)$ is the smallest closed set containing $A \cup B$, so $\text{cl}(A \cup B) \subseteq \text{cl}(A) \cup \text{cl}(B)$.

9.7 The proofs proceed along the same lines as those in exercise 9.6:

(a) $\emptyset$ is an open set and $\text{int}(\emptyset)$ is the largest open set contained in $\emptyset$, so $\text{int}(\emptyset) = \emptyset$.

(b) $\text{int}(U)$ is an open set contained in U, so $\text{int}(U) \subseteq U$.

(c) Since $\text{int}(U)$ is open, it is the largest open set containing itself: $\text{int}(\text{int}(U)) = \text{int}(U)$.

(d) By (b), $\text{int}(A) \subseteq A \subseteq B$, so $\text{int}(A)$ is an open set contained in B. Set $\text{int}(B)$ is the largest open set contained in B, so $\text{int}(A) \subseteq \text{int}(B)$.

(e) Since $U \cap V \subseteq U$, (d) gives $\text{int}(U \cap V) \subseteq \text{int}(U)$. Likewise, $\text{int}(U \cap V) \subseteq \text{int}(V)$. So $\text{int}(U \cap V) \subseteq \text{int}(U) \cap \text{int}(V)$. For the converse inclusion, (b) gives $\text{int}(U) \subseteq U$ and $\text{int}(V) \subseteq V$, so $\text{int}(U) \cap \text{int}(V) \subseteq U \cap V$. As the intersection of two open sets, $\text{int}(U) \cap \text{int}(V)$ is an open set contained in $U \cap V$. Set $\text{int}(U \cap V)$ is the largest open set contained in $U \cap V$, so $\text{int}(U) \cap \text{int}(V) \subseteq \text{int}(U \cap V)$.

9.8 (a) Take $U = (0,1), V = (1,2)$. Then $\text{cl}(U \cap V) = \text{cl}(\emptyset) = \emptyset$, but $\text{cl}(U) \cap \text{cl}(V) = [0,1] \cap [1,2] = \{1\}$.

(b) Take $U = (-\infty, 0)$. Then $\text{cl}(U^c) = \text{cl}[0,\infty) = [0,\infty)$, but $\text{cl}(U)^c = (-\infty,0]^c = (0,\infty)$.

(c) Take $U = (0,1), V = [1,2)$. Then $\text{int}(U \cup V) = \text{int}((0,2)) = (0,2)$, but $\text{int}(U) \cup \text{int}(V) = (0,1) \cup (1,2)$.

(d) Take $U = (-\infty, 0)$. Then $\text{int}(U^c) = (0,\infty)$, but $\text{int}(U)^c = [0,\infty)$.

9.9 (a) Note that a set W with $W \supseteq U$ is closed if and only if $W^c \subseteq U^c$ and W^c is open. Hence:

$$\begin{aligned}
\text{cl}(U)^c &= \left(\cap_{W\supseteq U, W \text{ is closed}} W\right)^c && \text{(by definition)} \\
&= \cup_{W\supseteq U, W \text{ is closed}} W^c && \text{(De Morgan's Law)} \\
&= \cup_{W^c\subseteq U^c, W^c \text{ is open}} W^c && \\
&= \cup_{V\subseteq U^c, V \text{ is open}} V && \text{(writing } V = W^c\text{)} \\
&= \text{int}(U^c) && \text{(by definition)}
\end{aligned}$$

(b) Note that a set W with $W \subseteq U$ is open if and only if $W^c \supseteq U^c$ and W^c is closed. Hence:

$$\begin{aligned}
\operatorname{int}(U)^c &= \left(\cup_{W\subseteq U, W \text{ is open}} W\right)^c && \text{(by definition)}\\
&= \cap_{W\subseteq U, W \text{ is open}} W^c && \text{(De Morgan's Law)}\\
&= \cup_{W^c\supseteq U^c, W^c \text{ is closed}} W^c && \\
&= \cup_{V\supseteq U^c, V \text{ is closed}} V && \text{(writing } V = W^c\text{)}\\
&= \operatorname{cl}(U^c) && \text{(by definition)}
\end{aligned}$$

9.10 It suffices to prove the chain of implications (a) ⇒ (b) ⇒ (c) ⇒ (d) ⇒ (a). In principle, I prove things straight from their definitions, but I will also provide some alternative proofs by invoking Theorem 9.4.

(a) ⇒ (b): Assume U is closed. Then $U \subseteq \operatorname{cl}(U) = \cap_{W\supseteq U, W \text{ is closed}} W \subseteq U$, so $U = \operatorname{cl}(U)$.

(b) ⇒ (c): Assume $U = \operatorname{cl}(U)$. Then $\operatorname{bd}(U) = \operatorname{cl}(U) \cap \operatorname{cl}(U^c) \subseteq \operatorname{cl}(U) = U$, so $\operatorname{bd}(U) \subseteq U$.
ALTERNATIVE PROOF: Assume $U = \operatorname{cl}(U)$. By Theorem 9.4, $U = \operatorname{cl}(U) = U \cup \operatorname{bd}(U)$. Equality $U = U \cup \operatorname{bd}(U)$ implies $\operatorname{bd}(U) \subseteq U$.

(c) ⇒ (d): Assume $\operatorname{bd}(U) \subseteq U$. Let $u \in \operatorname{acc}(U)$. To show: $u \in U$.

Suppose, to the contrary, that $u \notin U$. Then $u \in U^c \subseteq \operatorname{cl}(U^c)$. Also, $u \in \operatorname{cl}(U)$. Otherwise, if $u \notin \operatorname{cl}(U)$, it would lie in its complement $\operatorname{cl}(U)^c$, which is an open set: u is an interior point. This means that for some $\varepsilon > 0$,

$$B(u,\varepsilon) \subseteq \operatorname{cl}(U)^c \subseteq U^c,$$

where the final inclusion follows from the fact that $U \subseteq \operatorname{cl}(U)$. In particular,

$$B(u,\varepsilon) \setminus \{u\} \subseteq U^c.$$

But then u can't be an accumulation point of U.

We proved that $u \in \operatorname{cl}(U) \cap \operatorname{cl}(U^c) = \operatorname{bd}(U)$; and by assumption, $\operatorname{bd}(U) \subseteq U$, so $u \in U$. That contradicts our assumption that $u \notin U$!
ALTERNATIVE PROOF: Assume $\operatorname{bd}(U) \subseteq U$. By Theorem 9.4, $U = U \cup \operatorname{bd}(U) = \operatorname{cl}(U) = U \cup \operatorname{acc}(U)$. Equality $U = U \cup \operatorname{acc}(U)$ implies $\operatorname{acc}(U) \subseteq U$.

(d) ⇒ (a): Assume $\operatorname{acc}(U) \subseteq U$. To show: U is closed.

Suppose, to the contrary, that it isn't. Then its complement U^c is not open: some $u \in U^c$ is not an interior point. So, for each $\varepsilon > 0$:

$$B(u,\varepsilon) \cap U \neq \emptyset.$$

Since $u \notin U$, it follows that $u \in \operatorname{acc}(U) \subseteq U$. But that contradicts $u \notin U$!
ALTERNATIVE PROOF: Assume $\operatorname{acc}(U) \subseteq U$. By Theorem 9.4, $U = U \cup \operatorname{acc}(U) = \operatorname{cl}(U)$. Set $\operatorname{cl}(U)$ is closed, so U is closed.

9.11

- ☒ U is not open. Let $f \in U$ and $\varepsilon > 0$. The function $x \mapsto f(x) + \varepsilon/2$ is the sum of two continuous functions on $[0,1]$, hence continuous. It has distance $\varepsilon/2 < \varepsilon$ to f, but $0 \mapsto f(0) + \varepsilon/2 = 0 + \varepsilon/2 \neq 0$, so it does not belong to U. Conclude that no ε-ball around $f \in U$ lies entirely in U: U is not open.
- ☒ U is closed. We show that its complement in $C[0,1]$ is open. So let $f \in C[0,1]$ lie in U^c: $f(0) \neq 0$. Take $\varepsilon = |f(x)|/2$. For each $g \in B(f,\varepsilon)$, we have

  $$|f(0)| - |g(0)| \leq |f(0) - g(0)| < \varepsilon = |f(0)|/2,$$

  so $|g(0)| > |f(0)|/2 > 0$, showing that $g(0) \neq 0$: $g \in U^c$. Conclude that $B(f,\varepsilon) \subseteq U^c$: U^c is open.
- ☒ V is not open: if f is constant, then the function $x \mapsto f(x) + \frac{1}{2}\varepsilon x$ is nonconstant (but affine) and has distance $\frac{1}{2}\varepsilon$ to f, so there is no ε-ball around f that contains only constant functions.
- ☒ V is closed. To show this, we show that its complement in $C[0,1]$ is open. So let $f \in C[0,1]$ be a function that is not constant: $f(x) \neq f(y)$ for some $x, y \in [0,1]$. Then the ball around f with radius $\varepsilon = \frac{1}{2}|f(x) - f(y)|$ contains no constant functions.

☒ W is open. Let $f \in W$. We show that it is an interior point of W. Since f is continuous, it achieves a maximum in some point $y \in [0,1]$. Since $f \in W$: $f(y) < 1$. Take $\varepsilon = 1 - f(y)$. For each $g \in B(f,\varepsilon)$ and each $x \in [0,1]$, we have

$$g(x) < f(x) + \varepsilon = f(x) + 1 - f(y) = 1 + \underbrace{f(x) - f(y)}_{\leq 0} \leq 1,$$

so $g \in W$. Conclude that $B(f,\varepsilon) \subseteq W$.

☒ W is not closed, since its complement is not open: the constant function $f \in C[0,1]$ with $f(x) = 1$ lies in W^c, but is not an interior point of W^c, since a ε-ball around f contains the constant function $g \in C[0,1]$ with $g(x) = 1 - \varepsilon/2 < 1$, which lies in W. Hence, no ε-ball around f lies entirely in W^c: W^c is not open!

11.1 ☒ Assume $f : X \to \mathbb{R}^n$ is continuous and let $i \in \{1,\ldots,n\}$. To show: $f_i : X \to \mathbb{R}$ is continuous.
Let $x \in X$ and $\varepsilon > 0$. By continuity of f, there is a $\delta > 0$ such that all $y \in X$ with $d(x,y) < \delta$ have

$$|f_i(x) - f_i(y)| \leq \sqrt{\sum_{j=1}^{n} (f_j(x) - f_j(y))^2} < \varepsilon.$$

Hence, f_i is continuous at x.

☒ Assume that $f_i : X \to \mathbb{R}$ is continuous for each $i \in \{1,\ldots,n\}$. To show: $f : X \to \mathbb{R}^n$ is continuous.
Let $x \in X$ and $\varepsilon > 0$. For each i, f_i is continuous at x, so there is a $\delta_i > 0$ such that all $y \in X$ with $d(x,y) < \delta_i$ have $|f_i(x) - f_i(y)| < \varepsilon/n$. Hence, if $d(x,y) < \min\{\delta_1,\ldots,\delta_n\}$, it follows that

$$\|f(x) - f(y)\| \leq \sum_{i=1}^{n} |f_i(x) - f_i(y)| < \sum_{i=1}^{n} \varepsilon/n = \varepsilon.$$

Hence, f is continuous at x.

11.2 Function f is the composition $p \circ h$ of the functions $p : \mathbb{R}^2 \to \mathbb{R}$ and $h : X \to \mathbb{R}^2$ with

$$p(x_1,x_2) = x_1 x_2 \text{ for all } (x_1,x_2) \in \mathbb{R}^2 \qquad \text{and} \qquad h(x_1,x_2) = (x_1, 1/x_2) \text{ for all } (x_1,x_2) \in X.$$

Since the composition of continuous functions is again continuous (Thm. 11.3), it suffices to prove that p and h are continuous. We did so for p in Example 11.4. For h we can use the coordinate criterion (Example 11.6) and prove that each of its two coordinates, i.e., the functions $h_1 : X \to \mathbb{R}$ with $h_1(x_1,x_2) = x_1$ and $h_2 : X \to \mathbb{R}$ with $h_2(x_1,x_2) = 1/x_2$ are continuous. Function h_1 is linear, hence continuous by Example 11.3. And h_2 is the composition of the functions $(x_1,x_2) \mapsto x_2$ on X and $x_2 \mapsto 1/x_2$ on $\mathbb{R}\setminus\{0\}$, both continuous by linearity and Example 11.5 respectively.

11.3 Function $h : X \to \mathbb{R}^2$ with $h(x) = (f(x), g(x))$ for each $x \in X$ is continuous by the coordinate criterion (Example 11.6.

(a) Fix $c \in \mathbb{R}$. The function $m : \mathbb{R} \to \mathbb{R}$ with $m(x) = cx$ for each $x \in \mathbb{R}$ is linear, hence continuous (Example 11.3). The rescaled function cf is the composition $m \circ f$ of two continuous functions, hence continuous (Thm. 11.3).

(b) Function $s : \mathbb{R}^2 \to \mathbb{R}$ with $s(x_1,x_2) = x_1 + x_2$ for each $(x_1,x_2) \in \mathbb{R}^2$ is linear, hence continuous (Example 11.3). The sum function $f + g$ is the composition $s \circ h$ of two continuous functions, hence continuous (Thm. 11.3).

(c), (d) Same as (b), but taking the composition with the function $(x_1,x_2) \mapsto x_1 x_2$ from Example 11.4 and the function $(x_1,x_2) \mapsto x_1/x_2$ from Exercise 11.2.

11.4 Let $\varepsilon > 0$. Choose $\delta = \varepsilon/2$. Then for all $x, y \in (0,1)$ with $|x - y| < \delta$, we have

$$|x^2 - y^2| = |(x+y)(x-y)| = |x+y||x-y| \leq 2|x-y| < 2\delta = \varepsilon.$$

13.1 In exercises like this, it is often a good idea to divide both the numerator and denominator by the highest power of k to get some intuition about what the limit might be:

☒ In (a), we find

$$\frac{2k-3}{k+1} = \frac{2-3/k}{1+1/k}.$$

When k goes to infinity, the terms $-3/k$ and $1/k$ become really small, so our intuition is that the fraction converges to $\frac{2-0}{1+0} = 2$.

☒ In (b),

$$\frac{2k}{3k^2+4} = \frac{2/k}{3+4/k^2}.$$

When k goes to infinity, the terms $2/k$ and $4/k^2$ become really small, so our intuition is that the fraction converges to $\frac{0}{3+0} = 0$.

☒ In (c),

$$\frac{(-1)^k}{3k-1} = \frac{(-1)^k/k}{3-1/k}.$$

When k goes to infinity, the terms $(-1)^k/k$ and $1/k$ become really small, so our intuition is that the fraction converges to $\frac{0}{3-0} = 0$.

Now that we have figured out the candidate limits, it is time to get to the actual definition. Intuitively, we need to show that for each $\varepsilon > 0$, we can find large enough k such that the distance between x_k and its limit is less than ε: $d(x_k, x) = |x_k - x| < \varepsilon$. This is usually done by a chain of inequalities:

$$d(x_k, x) = |x_k - x| \leq \cdots < \varepsilon,$$

where the terms '$\cdots$' in the middle are constructed in such a way that they are easier to make small than the earlier terms. This is called ***majorizing*** (making larger) the expression. For fractions, this is usually done by increasing the numerator or decreasing the denominator: in a fraction of positive numbers, dividing by something smaller gives something larger. I will sketch this for the final problem; try it yourself for the other ones. We guessed that $\left(\frac{(-1)^k}{3k-1}\right)_{k\in\mathbb{N}}$ converges to zero, so I want to make $d(x_k, 0) < \varepsilon$. Here is a sketch:

$$d(x_k,0) = \left|\frac{(-1)^k}{3k-1} - 0\right| = \left|\frac{(-1)^k}{3k-1}\right| = \frac{\left|(-1)^k\right|}{|3k-1|} = \frac{1}{3k-1} \leq \frac{1}{3k-k} = \frac{1}{2k} < \varepsilon.$$

The final inequality is easy: $\frac{1}{2k} < \varepsilon$ whenever $k > \frac{1}{2\varepsilon}$. So *that* is going to be our candidate for the number N from which label onward the terms have a distance less than ε to the proposed limit zero.

(a) I prove that $\lim_{k\to\infty} \frac{2k-3}{k+1} = 2$. Let $\varepsilon > 0$. Choose $N > \frac{5}{\varepsilon}$. Then for each integer $k \geq N$:

$$d(x_k,2) = \left|\frac{2k-3}{k+1} - 2\right| = \left|\frac{2k-3}{k+1} - \frac{2(k+1)}{k+1}\right| = \left|\frac{-5}{k+1}\right| = \frac{5}{k+1} < \frac{5}{k} \leq \frac{5}{N} < \varepsilon.$$

(b) I prove that $\lim_{k\to\infty} \frac{2k}{3k^2+4} = 0$. Let $\varepsilon > 0$. Choose $N > \frac{2}{3\varepsilon}$. Then for each $k \geq N$:

$$d(x_k,0) = \left|\frac{2k}{3k^2+4} - 0\right| = \left|\frac{2/k}{3+4/k^2}\right| = \frac{2/k}{3+4/k^2} < \frac{2/k}{3} = \frac{2}{3k} \leq \frac{2}{3N} < \varepsilon.$$

(c) I prove that $\lim_{k\to\infty} \frac{(-1)^k}{3k-1} = 0$. Let $\varepsilon > 0$. Choose $N > \frac{1}{2\varepsilon}$. Then for each $k \geq N$:

$$d(x_k,0) = \left|\frac{(-1)^k}{3k-1} - 0\right| = \left|\frac{(-1)^k}{3k-1}\right| = \frac{\left|(-1)^k\right|}{|3k-1|} = \frac{1}{3k-1} \leq \frac{1}{3k-k} = \frac{1}{2k} \leq \frac{1}{2N} < \varepsilon.$$

13.2 Assume $\lim_{k\to\infty} x_k = x$ and consider a coordinate $i \in \{1,\ldots,n\}$. By definition, for $\varepsilon > 0$, there is an $N \in \mathbb{N}$ such that for all $k \geq N$:

$$|x_{ki} - x_i| = \sqrt{(x_{ki} - x_i)^2} \leq \sqrt{\sum_j (x_{kj} - x_j)^2} = \|x_k - x\| < \varepsilon,$$

showing that $\lim_{k\to\infty} x_{ki} = x_i$. Conversely, assume that $\lim_{k\to\infty} x_{ki} = x_i$ for all coordinates i. Then for each $\varepsilon > 0$ there is an $N_i \in \mathbb{N}$ such that for all $k \geq N_i$:

$$|x_{ki} - x_i| \leq \varepsilon/n.$$

Hence, for all $k \geq \max\{N_1, \ldots, N_n\}$:

$$\|x_k - x\| \leq \sum_{i=1}^{n} |x_{ki} - x_i| \leq \sum_{i=1}^{n} (\varepsilon/n) = \varepsilon,$$

showing that $\lim_{k\to\infty} x_k = x$.

13.3 This is an exotic case: recall that the discrete metric has $d(x,y) = 1$ if $x \neq y$ and $d(x,y) = 0$ if $x = y$. Consider any $x \in X$. What does it mean for a sequence $(x_k)_{k\in\mathbb{N}}$ to converge to x? For each $\varepsilon > 0$ there is an N such that $d(x_k, x) < \varepsilon$ for all $k \geq N$. But if we take $\varepsilon \in (0,1)$, the only point within distance ε from x is x itself! This means that a necessary (and evidently sufficient) condition for convergence to x is that the sequence $(x_k)_{k\in\mathbb{N}}$ is eventually constant and equal to x: there is an N such that $x_k = x$ for all $k \geq N$.

13.4 (a) No:

- ☒ If we approach $\mathbf{0}$ along the horizontal axis, via points of the form $(x_1, 0)$ with $x_1 \neq 0$, the function values are $f(x_1, 0) = \frac{x_1 \cdot 0}{x_1^2 + 0^2} = 0$. So if the limit exists, it must be 0.
- ☒ If we approach $\mathbf{0}$ along the vertical axis, via points of the form $(0, x_2)$ with $x_2 \neq 0$, the function values are $f(0, x_2) = \frac{0 \cdot x_2}{0^2 + x_2^2} = 0$. So if the limit exists, it must be 0.
- ☒ If we approach $\mathbf{0}$ diagonally, via points with equal coordinates (x_1, x_1) with $x_1 \neq 0$, the function values are $f(x_1, x_1) = \frac{x_1^2}{x_1^2 + x_1^2} = \frac{1}{2}$. So if the limit exists, it must be $\frac{1}{2}$.
- ☒ But if the limit exists, it has to be unique (as in Thm. 13.1). Since we have two different candidates, the limit can't exist.

This shows that approximating the point in which we need to compute the limit by changing only one coordinate at a time is not enough to draw meaningful conclusions about the limit!

(b) No:

- ☒ If we approach $\mathbf{0}$ along the horizontal axis, via points of the form $(x_1, 0)$ with $x_1 \neq 0$, the function values are $f(x_1, 0) = \frac{x_1^2 + 0^2}{|x_1 + 0| + |x_1 \cdot 0|} = \frac{x_1^2}{|x_1|} = |x_1|$, which goes to zero as x_1 goes to zero. So if the limit exists, it must be 0.
- ☒ If we approach $\mathbf{0}$ along the straight line where $x_2 = -x_1$ with $x_1 \neq 0$, the function values are $f(x_1, -x_1) = \frac{x_1^2 + (-x_1)^2}{|x_1 + (-x_1)| + |x_1 \cdot (-x_1)|} = \frac{2x_1^2}{|x_1^2|} = 2$. So if the limit exists, it must be 2.
- ☒ But if the limit exists, it has to be unique (as in Thm. 13.1). Since we have two different candidates, the limit can't exist.

(c) We prove that $\lim_{x\to\mathbf{0}} f(x) = 0$. Let $\varepsilon > 0$. Take $\delta = \varepsilon$. For each $x \in \mathbb{R}^2$ with $0 < d(x, \mathbf{0}) = \|x\| < \delta$, we have:

$$\begin{aligned} \big|f(x_1, x_2) - 0\big| &= \left| \frac{\sin(x_1 x_2)}{\sqrt{x_1^2 + x_2^2}} - 0 \right| = \left| \frac{\sin(x_1 x_2)}{\sqrt{x_1^2 + x_2^2}} \right| \\ &\leq \frac{|x_1 x_2|}{\sqrt{x_1^2 + x_2^2}} = \frac{|x_1||x_2|}{\sqrt{x_1^2 + x_2^2}} \\ &\leq \frac{\|x\|\,\|x\|}{\|x\|} = \|x\| < \delta = \varepsilon. \end{aligned}$$

By the way, to see why the hint is true, apply the mean value theorem to the sine function: for each $y \neq 0$, there is a z between 0 and y with $\sin(y) - \sin(0) = \sin'(z)(y - 0) = y\cos(z)$. Using that $|\cos(z)| \leq 1$ gives $|\sin(y)| = |\sin(y) - \sin(0)| = |y\cos(z)| \leq |y|$.

(d) Intuition: since fourth powers go to zero much faster than second powers, it seems a reasonable guess that the limit is zero. We prove this formally using Definition 8.1: let $\varepsilon > 0$. Choose $\delta = \sqrt{\varepsilon/2}$. Then for each $x \in \mathbb{R}^2 \setminus \{0\}$ with $0 < d(x, \mathbf{0}) = \|x\| < \delta$, we have

$$\begin{aligned}\left|f(x_1,x_2) - 0\right| &= \left|\frac{x_1^4 + x_2^4}{x_1^2 + x_2^2} - 0\right| = \frac{|x_1|^4 + |x_2|^4}{x_1^2 + x_2^2} \le \frac{\|x\|^4 + \|x\|^4}{\|x\|^2} \\ &= \frac{2\|x\|^4}{\|x\|^2} = 2\|x\|^2 < 2\delta^2 = 2(\sqrt{\varepsilon/2})^2 = \varepsilon.\end{aligned}$$

(e) Intuition: writing

$$f(x_1,x_2) = \frac{x_1^2 + 3x_1^2x_2 + x_2^2}{x_1^2 + x_2^2} = \frac{x_1^2 + x_2^2}{x_1^2 + x_2^2} + \frac{3x_1^2x_2}{x_1^2 + x_2^2} = 1 + \frac{3x_1^2x_2}{x_1^2 + x_2^2},$$

and realizing that the third-degree polynomial $3x_1^2x_2$ goes to zero faster than the second degree polynomial $x_1^2 + x_2^2$, it seems reasonable to guess that the limit must be 1. We prove this formally using Definition 8.1: let $\varepsilon > 0$. Take $\delta = \varepsilon/3$. Then for each $x \in \mathbb{R}^2 \setminus \{0\}$ with $0 < d(x, \mathbf{0}) = \|x\| < \delta$, we have

$$\left|f(x_1,x_2) - 0\right| = \left|\frac{x_1^2 + 3x_1^2x_2 + x_2^2}{x_1^2 + x_2^2} - 1\right| = \left|\frac{3x_1^2x_2}{x_1^2 + x_2^2}\right| \le \frac{3|x_1|^2\,|x_2|}{\|x\|^2} \le \frac{3\|x\|^2\,\|x\|}{\|x\|^2} = 3\|x\| < 3\delta = \varepsilon.$$

(f) We prove using Definition 8.1 that $\lim_{x\to\mathbf{0}} f(x_1,x_2) = 0$. Let $\varepsilon > 0$. Take $\delta = \varepsilon/4$. Then for each $x \in \mathbb{R}^2 \setminus \mathbf{0}$ with $0 < d(x, \mathbf{0}) = \|x\| < \delta$, we have

$$\left|f(x_1,x_2) - 0\right| = \left|\frac{4x_1x_2}{\sqrt{x_1^2 + x_2^2}} - 0\right| = \frac{4|x_1|\,|x_2|}{\sqrt{x_1^2 + x_2^2}} \le \frac{4\|x\|\,\|x\|}{\|x\|} = 4\|x\| < 4\delta = \varepsilon.$$

(g) ☒ If we approach $\mathbf{0}$ along the horizontal axis, via points of the form $(x_1, 0)$, the function values are $f(x_1,0) = \frac{x_1^2}{x_1^2+0^2} = 1$, making 1 the candidate limit.

☒ If we approach $\mathbf{0}$ along the vertical axis, via points of the form $(0, x_2)$, the function values are $f(0,x_2) = \frac{0^2}{0^2+x_2} = 0$, making 0 the candidate limit.

☒ But if the limit exists, it has to be unique (Theorem 8.1). Since we have two different candidates, the limit can't exist.

15.1 Let $f, g \in B(U, \mathbb{R})$. Then $f \le g + d_\infty(f,g)$, so

$$T(f) \overset{(a)}{\le} T(g + d_\infty(f,g)) \overset{(b)}{\le} T(g) + \beta d_\infty(f,g).$$

Reversing the roles of f and g gives $T(g) \le T(f) + \beta d_\infty(f,g)$. Combining the two inequalities gives $d_\infty(T(f), T(g)) \le \beta d_\infty(f,g)$.

15.2 The exercise is based on M. Edelstein (1962) "On fixed and periodic points under contractive mappings", J. London Math. Soc. 37, 74–79. The example in (d) is from D.G. Bennett and B. Fisher (1974) "On a fixed point theorem for compact metric spaces", Math. Magazine 47, 40–41.

(a) Suppose T has distinct fixed points x and y. Then $d(x,y) \overset{\text{fixed}}{=} d(T(x), T(y)) \overset{(43)}{<} d(x,y)$, a contradiction.

(b) f is the composition of continuous functions $x \mapsto (x, T(x))$ and $d : V \times V \to [0,\infty)$ and hence continuous. Since V is compact, f achieves a minimum at some $v \in V$. We show that $f(v) = d(v, T(v)) = 0$, so that $v = T(v)$, i.e., v is a fixed point of T. Suppose, to the contrary, that $f(v) > 0$. Then $v \neq T(v)$, so (43) gives $f(T(v)) = d(T(v), T^2(v)) < d(v, T(v)) = f(v)$, contradicting that f is minimal in v.

(c) Let $v_0 \in V$. For each $k \in \mathbb{N}$: $0 \le d(T^{k+1}(v_0), v) \overset{\text{fixed}}{=} d(T^{k+1}(v_0), T(v)) \overset{(43)}{\le} d(T^k(v_0), v)$, with equality only if $T^k(v_0) = v$, i.e., if the sequence of iterates has reached the fixed point. So the sequence of distances $d(T^k(v_0), v)$ is weakly decreasing, bounded from below by 0 and consequently converges to some $\ell \ge 0$. By sequential compactness, $T^k(v_0)$ has a convergent subsequence $(T^{k(n)}(v_0))_{n\in\mathbb{N}}$ with limit $w \in V$. By continuity of T: $T^{k(n)+1}(v_0) = T(T^{k(n)}(v_0)) \to T(w)$. By continuity of the metric, $d(T^{k(n)}(v_0), v) \to d(T(w), v)$ and $d(T^{k(n)+1}(v_0), v) \to d(T(w), v)$. Both limits must be ℓ, so $\ell = d(w, v) = d(T(w), v) = d(T(w), T(v))$. By (43), it follows that $w = v$, so $\ell = 0$.

(d) T is nonexpansive: for distinct x and y in $[0,1]$ we have

$$|T(x) - T(y)| = \left| \frac{x}{1+x} - \frac{y}{1+y} \right| = \frac{|x-y|}{(1+x)(1+y)} < \frac{|x-y|}{1 \cdot 1} = |x-y|.$$

We prove by induction on $k \in \mathbb{N}$ that $T^k(x) = x/(1+kx)$:

- ☒ For $k = 1$, this is true by definition of T.
- ☒ Now assume it is true for $k \in \mathbb{N}$ and let's prove that it is true for $k+1$:

$$\begin{aligned} T^{k+1}(x) &= T(T^k(x)) = T(x/(1+kx)) = \frac{x}{1+kx}\left(1 + \frac{x}{1+kx}\right)^{-1} \\ &= \frac{x}{1+kx}\left(\frac{(1+kx)+x}{1+kx}\right)^{-1} = \frac{x}{1+(k+1)x}. \end{aligned}$$

- ☒ By induction, the statement is true for all $k \in \mathbb{N}$.

To show that T^k is not a contraction, notice that $|T^k(x) - T^k(y)|/|x-y| = (1+kx)^{-1}(1+ky)^{-1}$ is arbitrarily close to 1 if x and y are sufficiently close to zero.

15.3 (a) Let $x \in \mathbb{R}$. Since $\sqrt{x^2+1} > \sqrt{x^2} = |x|$, we find that

$$|f'(x)| = \left| \frac{x}{\sqrt{x^2+1}} \right| = \frac{|x|}{\sqrt{x^2+1}} < 1.$$

(b) No. If x were a fixed point, we would have a contradiction:

$$f(x) = x \quad \Rightarrow \quad f(x)^2 = x^2 \quad \Rightarrow \quad x^2 + 1 = x^2 \quad \Rightarrow \quad 1 = 0.$$

(c) No: if it were, then it would have a fixed point on the complete metric space $\mathbb{R}$ with its usual distance.

17.1 According to Definition 17.1, a subset of $\mathbb{R}^n$ (with its usual distance) is compact if it is closed and bounded. I will give brief motivations whether the sets are compact.

(a) Not compact: the set is not closed, since it does not contain the boundary point 5.

(b) Compact: the set is finite (it has two elements) and every finite set is closed and bounded.

(c) Compact: the union can be written as $[0,5] \cup \{37\}$. So it is the union of two sets, $[0,5]$ and $\{37\}$, which are both closed and bounded and consequently closed and bounded itself.

(d) Not compact: the set is not bounded, since it contains the vector $(x_1, 1/x_1)$ for each $x_1 \neq 0$. The length $\|(x_1, 1/x_1)\| > x_1$ can be made arbitrarily large by letting x_1 go to infinity.

(e) Compact: if x lies in the set, its coordinates are bounded. For the first coordinate, for instance, we have

$$0 \le x_1 = x_1 + 0 \le x_1 + x_2 \le 3,$$

so the first coordinate is bounded from below by zero and from above by three. The same holds for the second coordinate. So the set is bounded. To see that it is closed, notice that it can be written as the intersection of four sets:

$$\{x \in \mathbb{R}^2 : 3x_1 - 2x_2 \le 6\}, \quad \{x \in \mathbb{R}^2 : x_1 + x_2 \le 3\}, \quad \{x \in \mathbb{R}^2 : x_1 \ge 0\}, \quad \{x \in \mathbb{R}^2 : x_2 \ge 0\}.$$

As the pre-image of a closed set under a continuous function, each of these four sets is closed (recall Theorem 11.2). The intersection of closed sets is a closed set.

(f) Not compact: the set is not closed, since it does not contain all of its boundary points. For instance, $(3,4)$ is a boundary point that does not belong to the set.

(g) Not compact: the set is not bounded, since it contains the vector $(1, x_2, 5+x_2)$ for each $x_2 \in \mathbb{R}^2$. The length $\|(1, x_2, 5+x_2)\| > x_2$ can be made arbitrarily large by letting x_2 go to infinity.

(h) Compact: the first restriction gives that $x_1^2 \leq 4$, so the first coordinate lies between -2 and 2. Likewise, the second coordinate lies between -1 and 1. The second restriction gives upper and lower bounds on x_3: since $x_3 = 1 - \frac{1}{2}x_1$ and x_1 lies between -2 and 2, it follows that x_3 lies between 0 and 2. Since each coordinate is bounded, the set is bounded. To see that it is closed, notice that it can be written as the intersection of two sets:

$$\{x \in \mathbb{R}^3 : x_1^2 + 4x_2^2 = 4\} \qquad \text{and} \qquad \{x \in \mathbb{R}^3 : x_1 + 2x_3 = 2\}.$$

As the pre-image of a closed set under a continuous function, each of these two sets is closed. The intersection of closed sets is closed.

(i) Compact: That the set is closed follows as in earlier cases. Why is it bounded? Let x be an element of the set. I will show that each of its coordinates is bounded. Conditions $0 \leq x_2 \leq x_1^3$ give $x_1 \geq 0$ and $x_2 \geq 0$, so both coordinates are bounded from below. Are they bounded from above? Since $x_2 \leq x_1^3$ and $x_2 \geq 2x_1^3 - 6x_1^2 + 12x_1 - 8$, we must have $x_1^3 \geq 2x_1^3 - 6x_1^2 + 12x_1 - 8$. Equivalently, $x_1^3 - 6x_1^2 + 12x_1 - 8 = (x_1-2)^3 \leq 0$. So $x_1 \leq 2$ and consequently $x_2 \leq x_1^3 \leq 8$: both coordinates are bounded from above. Conclude: the set is bounded.

REMARK: If you didn't realize that $x_1^3 - 6x_1^2 + 12x_1 - 8 = (x_1-2)^3$, there are other ways of obtaining upper bounds on x_1. For instance,

$$x_1^3 - 6x_1^2 + 12x_1 - 8 = x_1(x_1^2 - 6x_1 + 12) - 8 = x_1((x_1-3)^2 + 3) - 8 > 3(0+3) - 8 = 1$$

if $x_1 > 3$, so $x_1 \leq 3$.

17.2 Since $U \subseteq X$, (a) implies (b). To see that (b) implies (a), let $\varepsilon > 0$. For $\varepsilon/2 > 0$ there is a finite subset $x_1, \ldots, x_m$ of X such that the balls $B(x_1, \varepsilon/2), \ldots, B(x_m, \varepsilon/2)$ cover U. We may assume that $B(x_i, \varepsilon/2) \cap U \neq \emptyset$ for each x_i: otherwise that ball is not needed to cover U. So pick an element $u_i \in B(x_i, \varepsilon/2) \cap U$ for each x_i. By the triangle inequality, $B(x_i, \varepsilon/2) \subseteq B(u_i, \varepsilon)$, so the balls $B(u_1, \varepsilon), \ldots, B(u_m, \varepsilon)$ with centers in $U' = \{u_1, \ldots, u_m\} \subseteq U$ cover U.

17.3 Let C be a bounded subset of a metric space (X, d). Take $\varepsilon = 1$. Since C is totally bounded, there is a finite covering $B(c_1, 1), \ldots, B(c_k, 1)$ of C. Let $\delta = 1 + \max\{d(c_i, c_1) : i = 1, \ldots, k\}$. We show that $C \subseteq B(c_1, \delta)$. Let $c \in C$. Since the balls cover C, there is an i with $c \in B(c_i, 1)$. By the triangle inequality:

$$d(c, c_1) \leq d(c, c_i) + d(c_i, c_1) < 1 + d(c_i, c_1) \leq \delta.$$

17.4 Each totally bounded set is bounded (Exc. 17.3). Conversely, let U be bounded. I give two proofs that U is totally bounded.

METHOD 1: Each coordinate of a real vector can be approximated arbitrarily well by rounding it off to a large but finite number of decimal places. Since U is bounded, each coordinate is bounded and there are only finitely many points with that number of decimal places between the respective lower and upper bounds. So for each 'precision' $\varepsilon > 0$, we can find a finite set F of approximating vectors such that each element of U lies within distance ε of an element in F: U is totally bounded.

METHOD 2: Suppose U is not totally bounded: for some $\varepsilon > 0$ it is impossible to cover U by a finite number of balls with radius ε. Pick any u_1 in U. The ball $B(u_1, \varepsilon)$ doesn't cover U, so there is a u_2 in U that does not lie in this ball. Balls $B(u_1, \varepsilon)$ and $B(u_2, \varepsilon)$ do not cover U, so there is a u_3 in U that does not lie in these balls. Continue this way to find a sequence $(u_1, u_2, u_3, \ldots)$ in U where each u_k does not belong to balls $B(u_1, \varepsilon), \ldots, B(u_{k-1}, \varepsilon)$: its terms lie ε or more away from each other. But then it has no convergent subsequence, contradicting the Bolzano-Weierstrass theorem 13.2.

17.5 The closure of U is closed (Definition 9.3). If we can show that it is bounded as well, it is compact by Definition 17.1 (or Heine-Borel, Theorem 17.5). We establish boundedness in two ways:

METHOD 1: By definition (Def. 8.3), since U is bounded, there is an open ball $B(x,r)$ with

$$U \subseteq B(x,r) = \{y \in \mathbb{R}^n : d(x,y) < r\},$$

so in particular (switching from $<$ to $\leq$, which only makes the righthand set larger):

$$U \subseteq \{y \in \mathbb{R}^n : d(x,y) \leq r\}.$$

Since the set on the right is a closed set (Example 9.2) containing U and $\mathrm{cl}(U)$ is the *smallest* closed set containing U:

$$\mathrm{cl}(U) \subseteq \{y \in \mathbb{R}^n : d(x,y) \leq r\}.$$

This shows that $\mathrm{cl}(U)$ is bounded: it is contained in any open ball around x with radius larger than r.

METHOD 2: By definition (Def. 8.3), since U is bounded, there is an open ball $B(x,r)$ with

$$U \subseteq B(x,r) = \{y \in \mathbb{R}^n : d(x,y) < r\}.$$

We use the triangle inequality and Theorem 9.4(b), which says that

$$\mathrm{cl}(U) = \{x \in X \mid \text{ for each } \varepsilon > 0 : B(x,\varepsilon) \cap U \neq \emptyset\}$$

to show that

$$\mathrm{cl}(U) \subseteq B(x, r+1),$$

which makes $\mathrm{cl}(U)$ bounded. So let $y \in \mathrm{cl}(U)$. To show: $y \in B(x, r+1)$.

Taking $\varepsilon = 1$, we know that there is a $u \in B(y,1) \cap U$. In particular, $d(y,u) < 1$. And $U \subseteq B(x,r)$ implies $d(u,x) < r$. By the triangle inequality,

$$d(y,x) \leq d(y,u) + d(u,x) < 1 + r,$$

so $y \in B(x, r+1)$, as we had to prove!

17.6 (a) For each $x \in [0,1]$, the triangle inequality gives:

$$\begin{aligned} \big|f_b(x) - f_a(x)\big| &= |(b_1 x + b_2) - (a_1 x + a_2)| = |(b_1 - a_1)x + (b_2 - a_2)| \\ &\leq |(b_1 - a_1)x| + |b_2 - a_2| = |b_1 - a_1|\,|x| + |b_2 - a_2| \\ &\leq |b_1 - a_1| \cdot 1 + |b_2 - a_2| = |b_1 - a_1| + |b_2 - a_2|. \end{aligned}$$

(b) For each coordinate $i \in \{1,2\}$, $\|b-a\|_2 = \sqrt{(b_1-a_1)^2 + (b_2-a_2)^2} \geq \sqrt{(b_i-a_i)^2} = |b_i - a_i|$. Substituting $|b_i - a_i| \leq \|b-a\|_2$ into our previous answer we see that for all $x \in [0,1]$:

$$|f_b(x) - f_a(x)| \leq |b_1 - a_1| + |b_2 - a_2| \leq \|b-a\|_2 + \|b-a\|_2 = 2\|b-a\|_2.$$

Since this holds for all x in the domain of f_a and f_b, it follows that in particular

$$d_\infty(f_b, f_a) = \max_{x \in [0,1]} |f_b(x) - f_a(x)| \leq 2\|b-a\|_2.$$

(c) We prove that F is continuous at each point $a = (a_1, a_2)$ of its domain. Let $\varepsilon > 0$. Choose $\delta = \varepsilon/2$. Then for all $b \in \mathbb{R}^2$ with $d_2(b,a) = \|b-a\|_2 < \delta$ we have

$$d_\infty(F(b), F(a)) = d_\infty(f_b, f_a) \leq 2\|b-a\|_2 < 2\delta = \varepsilon.$$

(d) The intervals $[-1,4]$ and $[2,3]$ are compact subsets of $\mathbb{R}$ (Example 17.6). By Tychonoff's theorem, their Cartesian product $[-1,4] \times [2,3]$ is compact in $\mathbb{R}^2$. By Theorem 17.6(a), the continuous function F maps this compact set $[-1,4] \times [2,3]$ to a compact set $V = F([-1,4] \times [2,3])$. So V is compact.

17.7 Recall that a set is closed if its complement is open. The only open sets are $\emptyset$, which is the complement of X, and X, which is the complement of $\emptyset$. So only the empty set and X are closed.

Each subset of X is compact. The empty set is compact: since it has zero elements, we need zero open sets to cover it. To see that any nonempty subset Y is compact, suppose $\{O_i : i \in I\}$ is a covering. Since Y is nonempty, at least one of the open sets O_i is nonempty. The only nonempty open set in this exercise is X itself, so $O_i = X$ for some $i \in I$. Since $Y \subseteq X$, this single set $O_i = X$ is a finite subcovering.

REMARK: it is important to keep in mind that the indiscrete space in this exercise is highly exceptional: you rarely encounter cases where all sets are compact. The answer to the question what sets are compact depends on what sets are open and consequently on what sets are allowed in a covering. In the indiscrete space, few sets are open (only two); in a metric space, more sets are open. That will typically change the answer!

19.1 Assume C contains all nonnegative combinations of its elements. In particular, if $x, y \in C$, and $\lambda \geq 0$:

$$x + y = 1 \cdot x + 1 \cdot y \in C \qquad \text{and} \qquad \lambda x \in C,$$

showing that C is a convex cone. Conversely, assume C is a convex cone. By definition, it contains all nonnegative factors of one of its elements. For more than one term, we proceed by induction. Assume that C contains all nonnegative combinations of at most m terms. Then it also contains all nonnegative combinations of $m+1$ terms, since such a combination can be rewritten as

$$\lambda_1 v_1 + \cdots + \lambda_m v_m + \lambda_{m+1} v_{m+1} = (\lambda v_1 + \cdots + \lambda_m v_m) + \lambda_{m+1} v_{m+1}.$$

By assumption, $\lambda v_1 + \cdots + \lambda_m v_m \in C$ and $\lambda_{m+1} v_{m+1} \in C$. And since C is closed under addition, their sum is in C as well!

19.2 We solve the system of linear inequalities:

$$\begin{aligned} x_1 - x_2 &\leq 0 && (163)\\ x_1 - x_3 &\leq 0 && (164)\\ -x_1 + x_2 + 2x_3 &\leq 2 && (165)\\ -x_3 &\leq -1 && (166) \end{aligned}$$

We first eliminate x_1. In inequality (165), x_1 has a negative coefficient. It imposes a lower bound on x_1:

$$x_2 + 2x_3 - 2 \leq x_1.$$

Inequalities (163) and (164), where x_1 has a positive coefficient, impose upper bounds on x_1:

$$\begin{aligned} x_1 &\leq x_2\\ x_1 &\leq x_3 \end{aligned}$$

or, equivalently: $x_1 \leq \min\{x_2, x_3\}$. Inequality (166), where x_1 has coefficient zero (i.e., it does not occur in (166)) imposes no bounds on x_1.

So x_1 must satisfy

$$x_2 + 2x_3 - 2 \leq x_1 \leq \min\{x_2, x_3\}$$

We can find an x_1 between the lower and upper bounds if and only if the lower bound on x_1 does not exceed any of the upper bounds. Moreover, we still need (166). So there is a solution if and only if

$$\begin{aligned} x_2 + 2x_3 - 2 &\leq x_2\\ x_2 + 2x_3 - 2 &\leq x_3\\ -x_3 &\leq -1 \end{aligned}$$

has a solution. Simplify and rearrange terms so that all variables are on the lefthand side and all constants on the righthand side of the inequalities:

$$\begin{aligned} x_3 &\leq 1 \\ x_2 + x_3 &\leq 2 \\ -x_3 &\leq -1 \end{aligned}$$

Now we eliminate x_2, which is easy: the first and third inequality impose no restrictions on x_2. Only the second inequality imposes an upper bound on x_2:

$$x_2 \leq 2 - x_3.$$

In other words, for every feasible value for x_3, there is a solution by choosing x_2 sufficiently small. The restrictions on the feasible values of x_3 are given in the first and third inequality, which together give only one feasible candidate for x_3, namely $x_3 = 1$. Reading all of this backwards, we find that the set of solutions to the system of linear inequalities consists of all vectors $x = (x_1, x_2, x_3) \in \mathbb{R}^3$ with $x_3 = 1$, $x_2 \leq 2 - x_3 = 1$ and

$$x_2 + 2x_3 - 2 \leq x_1 \leq \min\{x_2, x_3\} \iff x_2 + 2 \cdot 1 - 2 = x_2 \leq x_1 \leq \min\{x_2, 1\}.$$

Since we know that $x_2 \leq 1$, it follows that $\min\{x_2, 1\} = x_2$, so

$$x_2 \leq x_1 \leq x_2 \qquad \text{or simply} \qquad x_1 = x_2.$$

19.3 If $P = \{x \in \mathbb{R}^n : Ax \leq \mathbf{0}\}$ for some matrix A, then P is obviously a polyhedron. It is also a convex cone: if $x, y \in P$ and $\lambda \geq 0$, then

- ☒ $Ax \leq \mathbf{0}$ and $Ay \leq \mathbf{0}$ imply $A(x+y) = Ax + Ay \leq \mathbf{0} + \mathbf{0} = \mathbf{0}$, so $x + y \in P$.
- ☒ $Ax \leq \mathbf{0}$ implies $A(\lambda x) = \lambda(Ax) \leq \lambda\mathbf{0} = \mathbf{0}$, so $\lambda x \in P$.

Conversely, assume that nonempty set P is a polyhedral cone. Since it is a polyhedron, there exist a matrix $A \in \mathbb{R}^{m \times n}$ and a vector $b \in \mathbb{R}^m$ with $P = \{x \in \mathbb{R}^n : Ax \leq b\}$. Let us prove that $P = \{x \in \mathbb{R}^n : Ax \leq \mathbf{0}\}$.

- ☒ Since P is nonempty and a convex cone, it follows that $\mathbf{0} \in P$. Consequently, $A\mathbf{0} = \mathbf{0} \leq b$. So if $Ax \leq \mathbf{0}$, then $Ax \leq b$, showing that $P \supseteq \{x \in \mathbb{R}^n : Ax \leq \mathbf{0}\}$.
- ☒ To establish the inclusion $P \subseteq \{x \in \mathbb{R}^n : Ax \leq \mathbf{0}\}$, we need to show that $Ax \leq \mathbf{0}$ for all $x \in P$. Suppose, to the contrary, that $a_i x > 0$ for some $x \in P$ and some row a_i of A. Since $\lambda x \in P$ for all $\lambda \geq 0$, it follows that $a_i(\lambda x) = \lambda(a_i x)$ is not bounded from above, contradicting that $a_i x \leq b_i$ for all $x \in P$.

19.4 A polytope is the convex hull of a finite set of vectors $\{v_1, \ldots, v_m\}$. Let $V \in \mathbb{R}^{n \times m}$ have these vectors as its columns. Then the polytope can be written as

$$\{x \in \mathbb{R}^n : \text{there is a } \lambda \in \mathbb{R}^m \text{ with } x = V\lambda, \lambda \geq \mathbf{0}, \sum_i \lambda_i = 1\}$$

which is the projection of the polyhedron

$$\{(x, \lambda) \in \mathbb{R}^{n+m} : x - V\lambda \leq 0, -x + V\lambda \leq 0, -\lambda \leq 0, \sum_i \lambda_i \leq 1, -\sum_i \lambda_1 \leq -1\}$$

by projecting away the coordinates of λ one at a time. By Fourier-Motzkin elimination, this is a polyhedron.

20.1 (a) Rewrite the inequalities in terms of upper and lower bounds on x_1:

$$\begin{aligned} x_1 &\leq 4 - x_2 - x_3 \\ x_1 &\leq 2x_2 + x_3 \\ -1 + x_2 + x_3 &\leq x_1 \\ 7 - 3x_2 - 4x_3 &\leq x_1 \end{aligned}$$

This means that x_1 must satisfy

$$\max\{-1 + x_2 + x_3, 7 - 3x_2 - 4x_3\} \leq x_1 \leq \min\{4 - x_2 - x_3, 2x_2 + x_3\}. \tag{167}$$

Thus, there is a solution if and only if each of the lower bounds on x_1 is less than or equal to each of the upper bounds on x_1:

$$\begin{aligned} -1+\ x_2+\ x_3 &\le 4-\ x_2-x_3 \\ -1+\ x_2+\ x_3 &\le \quad 2x_2+x_3 \\ 7-3x_2-4x_3 &\le 4-\ x_2-x_3 \\ 7-3x_2-4x_3 &\le \quad 2x_2+x_3 \end{aligned}$$

Rewrite the inequalities in terms of upper and lower bounds on x_2:

$$\begin{aligned} x_2 &\le \frac{5}{2}-x_3 \\ -1 &\le x_2 \\ \frac{3}{2}-\frac{3}{2}x_3 &\le x_2 \\ \frac{7}{5}-\ x_3 &\le x_2 \end{aligned}$$

This means that x_2 must satisfy

$$\max\{-1, \frac{3}{2}-\frac{3}{2}x_3, \frac{7}{5}-x_3\} \le x_2 \le \frac{5}{2}-x_3. \tag{168}$$

Thus, there is a solution if and only if each of the lower bounds on x_2 is less than or equal to the upper bound on x_2:

$$\begin{aligned} -1 &\le \frac{5}{2}-x_3 \\ \frac{3}{2}-\frac{3}{2}x_3 &\le \frac{5}{2}-x_3 \\ \frac{7}{5}-\ x_3 &\le \frac{5}{2}-x_3 \end{aligned}$$

Rewrite the inequalities in terms of upper and lower bounds on x_3:

$$\begin{aligned} x_3 &\le \frac{7}{2} \\ -2 &\le x_3 \\ \frac{7}{5} &\le \frac{5}{2} \end{aligned}$$

The final inequality is obviously true, no matter what x_3 is. It follows that x_3 must satisfy $-2 \le x_3 \le \frac{7}{2}$. Then the feasible x_2 follow from (168) and the feasible x_1 follow from (167).

(b) Since

$$A = \begin{bmatrix} 1 & 1 & 1 \\ 1 & -2 & -1 \\ -1 & 1 & 1 \\ -1 & -3 & -4 \end{bmatrix} \text{ and } b = \begin{bmatrix} 4 \\ 0 \\ 1 \\ -7 \end{bmatrix},$$

we solve system $y^\top A = \mathbf{0}^\top$ or, equivalently, $A^\top y = \mathbf{0}$, by Gaussian elimination on the coefficient matrix $A^\top$. After several steps, we find reduced matrix

$$\begin{bmatrix} 1 & 0 & 0 & -\frac{5}{2} \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & -\frac{5}{2} \end{bmatrix}$$

It follows that

$$y_4 \in \mathbb{R} \text{ (free)}, y_3 = \frac{5}{2}y_4, y_2 = y_4, y_1 = \frac{5}{2}y_4.$$

The restriction $y \geq \mathbf{0}$ requires that $y_4 \geq 0$. But then substitution of the solution gives

$$y^\top b = 4y_1 + y_3 - 7y_4 = 4\left(\frac{5}{2}y_4\right) + \frac{5}{2}y_4 - 7y_4 = \frac{11}{2}y_4 \geq 0,$$

contradicting the requirement that $y^\top b < 0$.

20.2 (a) In set-theoretic notation we need to show:

$$\{x \in \mathbb{R}^n : Ax \geq b\} \neq \emptyset \Leftrightarrow \{y \in \mathbb{R}^m : y^\top A = \mathbf{0}^\top, y^\top b > 0, y \geq \mathbf{0}\} = \emptyset.$$

Using Theorem 20.3, we find:

$$\begin{aligned}\{x \in \mathbb{R}^n : Ax \geq b\} \neq \emptyset &\Leftrightarrow \{x \in \mathbb{R}^n : (-A)x \leq -b\} \neq \emptyset \\ &\Leftrightarrow \{y \in \mathbb{R}^m : y^\top(-A) = \mathbf{0}^\top, y^\top(-b) < 0, y \geq \mathbf{0}\} = \emptyset \\ &\Leftrightarrow \{y \in \mathbb{R}^m : y^\top A = \mathbf{0}^\top, y^\top b > 0, y \geq \mathbf{0}\} = \emptyset.\end{aligned}$$

(b) Using Theorem 20.3:

$$\begin{aligned}\{x \in \mathbb{R}^n : Ax = b\} \neq \emptyset &\Leftrightarrow \left\{x \in \mathbb{R}^n : \begin{bmatrix} A \\ -A \end{bmatrix} x \leq \begin{bmatrix} b \\ -b \end{bmatrix}\right\} \neq \emptyset \\ &\Leftrightarrow \left\{(y_1, y_2) \in \mathbb{R}^{m+m} : (y_1, y_2)^\top \begin{bmatrix} A \\ -A \end{bmatrix} = \mathbf{0}^\top, (y_1, y_2)^\top \begin{bmatrix} b \\ -b \end{bmatrix} < 0, (y_1, y_2) \geq \mathbf{0}\right\} = \emptyset \\ &\Leftrightarrow \left\{(y_1, y_2) \in \mathbb{R}^{m+m} : (y_1 - y_2)^\top A = \mathbf{0}^\top, (y_1 - y_2)^\top b < 0, (y_1, y_2) \geq \mathbf{0}\right\} = \emptyset \\ &\Leftrightarrow \left\{y \in \mathbb{R}^m : y^\top A = \mathbf{0}^\top, y^\top b < 0\right\} = \emptyset.\end{aligned}$$

(c) Using Theorem 20.3:

$$\begin{aligned}\{x \in \mathbb{R}^n : Ax \leq b, x \geq \mathbf{0}\} \neq \emptyset &\Leftrightarrow \left\{x \in \mathbb{R}^n : \begin{bmatrix} A \\ -I \end{bmatrix} x \leq \begin{bmatrix} b \\ \mathbf{0} \end{bmatrix}\right\} \neq \emptyset \\ &\Leftrightarrow \left\{(y_1, y_2) \in \mathbb{R}^{m+m} : (y_1, y_2)^\top \begin{bmatrix} A \\ -I \end{bmatrix} = \mathbf{0}^\top, (y_1, y_2)^\top \begin{bmatrix} b \\ \mathbf{0} \end{bmatrix} < 0, (y_1, y_2) \geq \mathbf{0}\right\} = \emptyset \\ &\Leftrightarrow \left\{(y_1, y_2) \in \mathbb{R}^{m+m} : y_1^\top A - y_2^\top = \mathbf{0}^\top, y_1^\top b < 0, (y_1, y_2) \geq \mathbf{0}\right\} = \emptyset \\ &\Leftrightarrow \left\{y \in \mathbb{R}^m : y^\top A \geq \mathbf{0}^\top, y^\top b < 0, y \geq \mathbf{0}\right\} = \emptyset.\end{aligned}$$

(d) Let $\mathbf{1} = (1, \ldots, 1) \in \mathbb{R}^n$ be the vector with all coordinates equal to 1. Using Theorem 20.1:

$$\begin{aligned}\{x \in \mathbb{R}^n : Ax = \mathbf{0}, \mathbf{0} \neq x \geq \mathbf{0}\} \neq \emptyset &\Leftrightarrow \{x \in \mathbb{R}^n : Ax = \mathbf{0}, \sum_i x_i = 1, x \geq \mathbf{0}\} \neq \emptyset \\ &\Leftrightarrow \left\{x \in \mathbb{R}^n : \begin{bmatrix} A \\ \mathbf{1}^\top \end{bmatrix} x = \begin{bmatrix} \mathbf{0} \\ 1 \end{bmatrix}, x \geq \mathbf{0}\right\} \neq \emptyset \\ &\Leftrightarrow \left\{(y, y_{m+1}) \in \mathbb{R}^{m+1} : (y, y_{m+1})^\top \begin{bmatrix} A \\ \mathbf{1}^\top \end{bmatrix} \geq \mathbf{0}^\top, (y, y_{m+1})^\top \begin{bmatrix} \mathbf{0} \\ 1 \end{bmatrix} < 0\right\} = \emptyset \\ &\Leftrightarrow \left\{(y, y_{m+1}) \in \mathbb{R}^{m+1} : y^\top A + y_{m+1}\mathbf{1}^\top \geq \mathbf{0}^\top, y_{m+1} < 0\right\} = \emptyset \\ &\Leftrightarrow \left\{y \in \mathbb{R}^m : y^\top A > \mathbf{0}^\top\right\} = \emptyset.\end{aligned}$$

(e) Rescaling vector x if necessary and using Theorem 20.3:

$$\begin{aligned}
\{x \in \mathbb{R}^n : Ax = \mathbf{0}, x > \mathbf{0}\} \neq \emptyset &\Leftrightarrow \{x \in \mathbb{R}^n : Ax = \mathbf{0}, x \geq \mathbf{1}\} \neq \emptyset \\
&\Leftrightarrow \left\{ x \in \mathbb{R}^n : \begin{bmatrix} A \\ -A \\ -I \end{bmatrix} x \leq \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ -\mathbf{1} \end{bmatrix} \right\} \neq \emptyset \\
&\Leftrightarrow \left\{ (y_1, y_2, y_3) \in \mathbb{R}^{3m} : (y_1, y_2, y_3)^\top \begin{bmatrix} A \\ -A \\ -I \end{bmatrix} = \mathbf{0}^\top, (y_1, y_2, y_3)^\top \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ -\mathbf{1} \end{bmatrix} < 0, (y_1, y_2, y_3) \geq \mathbf{0} \right\} = \emptyset \\
&\Leftrightarrow \left\{ (y_1, y_2, y_3) \in \mathbb{R}^{3m} : (y_1 - y_2)^\top A - y_3^\top = \mathbf{0}^\top, -y_3^\top \mathbf{1} < 0, (y_1, y_2, y_3) \geq \mathbf{0} \right\} = \emptyset \\
&\Leftrightarrow \left\{ (y_1, y_2, y_3) \in \mathbb{R}^{3m} : (y_1 - y_2)^\top A = y_3^\top, y_1, y_2 \geq \mathbf{0}, \mathbf{0} \neq y_3 \geq \mathbf{0} \right\} = \emptyset \\
&\Leftrightarrow \{y \in \mathbb{R}^m : \mathbf{0} \neq y^\top A \geq \mathbf{0}^\top\} = \emptyset.
\end{aligned}$$

(f) Using part (d):

$$\begin{aligned}
\{x \in \mathbb{R}^n : Ax \leq \mathbf{0}, \mathbf{0} \neq x \geq \mathbf{0}\} \neq \emptyset &\Leftrightarrow \left\{ (x, x') \in \mathbb{R}^{n+n} : Ax + Ix' = \begin{bmatrix} A & I \end{bmatrix} \begin{bmatrix} x \\ x' \end{bmatrix} = \mathbf{0}, \mathbf{0} \neq (x, x') \geq \mathbf{0} \right\} \neq \emptyset \\
&\Leftrightarrow \left\{ y \in \mathbb{R}^m : y^\top \begin{bmatrix} A & I \end{bmatrix} > \mathbf{0}^\top \right\} = \emptyset \\
&\Leftrightarrow \{y \in \mathbb{R}^m : \begin{bmatrix} y^\top A & y^\top I \end{bmatrix} > \mathbf{0}^\top\} = \emptyset \\
&\Leftrightarrow \{y \in \mathbb{R}^m : y^\top A > \mathbf{0}^\top, y > \mathbf{0}\} = \emptyset.
\end{aligned}$$

(g) Rescaling vector x if necessary and using Theorem 20.3:

$$\begin{aligned}
\{x \in \mathbb{R}^n : Ax \leq \mathbf{0}, x > \mathbf{0}\} \neq \emptyset &\Leftrightarrow \{x \in \mathbb{R}^n : Ax \leq \mathbf{0}, x \geq \mathbf{1}\} \neq \emptyset \\
&\Leftrightarrow \left\{ x \in \mathbb{R}^n : \begin{bmatrix} A \\ -I \end{bmatrix} x \leq \begin{bmatrix} \mathbf{0} \\ -\mathbf{1} \end{bmatrix} \right\} \neq \emptyset \\
&\Leftrightarrow \left\{ (y_1, y_2) \in \mathbb{R}^m \times \mathbb{R}^n : \begin{bmatrix} y_1^\top & y_2^\top \end{bmatrix} \begin{bmatrix} A \\ -I \end{bmatrix} = \mathbf{0}^\top, \begin{bmatrix} y_1^\top & y_2^\top \end{bmatrix} \begin{bmatrix} \mathbf{0} \\ -\mathbf{1} \end{bmatrix} < 0, (y_1, y_2) \geq \mathbf{0} \right\} = \emptyset \\
&\Leftrightarrow \left\{ (y_1, y_2) \in \mathbb{R}^m \times \mathbb{R}^n : y_1^\top A = y_2^\top, y_2^\top \mathbf{1} > 0, y_1, y_2 \geq \mathbf{0} \right\} = \emptyset \\
&\Leftrightarrow \left\{ y \in \mathbb{R}^m : \mathbf{0}^\top \neq y^\top A \geq \mathbf{0}, y \geq \mathbf{0} \right\} = \emptyset.
\end{aligned}$$

20.3 PROOF USING THEOREM 20.1: Rescaling vector x if necessary, we see that the first set is nonempty if and only if there is a solution x to the system $Ax \leq -\mathbf{1}, Bx \leq \mathbf{0}, Cx = \mathbf{0}$. Writing $x = x^+ - x^-$ with $x^+, x^- \geq \mathbf{0}$, and introducing slacks, this is equivalent with there being a *nonnegative* solution to

$$\begin{aligned}
A(x^+ - x^-) + Is_A &= -\mathbf{1} \\
B(x^+ - x^-) + Is_B &= \mathbf{0} \\
C(x^+ - x^-) &= \mathbf{0}
\end{aligned}$$

Or, in matrix notation, using O for a zero matrix, to the system

$$\begin{bmatrix} A & -A & I & O \\ B & -B & O & I \\ C & -C & O & O \end{bmatrix} \begin{bmatrix} x^+ \\ x^- \\ s_A \\ s_B \end{bmatrix} = \begin{bmatrix} -\mathbf{1} \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix}$$

By Theorem 20.1, this means that there is no solution $(y_1, y_2, y_3) \in \mathbb{R}^{m_1} \times \mathbb{R}^{m_2} \times \mathbb{R}^{m_3}$ to the system

$$\begin{bmatrix} y_1^\top & y_2^\top & y_3^\top \end{bmatrix} \begin{bmatrix} A & -A & I & O \\ B & -B & O & I \\ C & -C & O & O \end{bmatrix} \geq \mathbf{0}^\top$$

$$\begin{bmatrix} y_1^\top & y_2^\top & y_3^\top \end{bmatrix} \begin{bmatrix} -\mathbf{1} \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix} < 0.$$

Rewriting, this means that there is no solution $(y_1, y_2, y_3) \in \mathbb{R}^{m_1} \times \mathbb{R}^{m_2} \times \mathbb{R}^{m_3}$ to $y_1^\top A + y_2^\top B + y_3^\top C = \mathbf{0}^\top, \mathbf{0} \neq y_1 \geq \mathbf{0}, y_2 \geq \mathbf{0}$.

PROOF USING THEOREM 20.3: Rescaling vector x if necessary, we see that the first set is nonempty if and only if there is a solution x to the system $Ax \leq -\mathbf{1}, Bx \leq \mathbf{0}, Cx = \mathbf{0}$. In matrix notation, this system becomes

$$\begin{bmatrix} A \\ B \\ C \\ -C \end{bmatrix} x \leq \begin{bmatrix} -\mathbf{1} \\ \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix}$$

By Theorem 20.3, this means that there is no solution $(y_1, y_2, y_3, y_4) \in \mathbb{R}^{m_1} \times \mathbb{R}^{m_2} \times \mathbb{R}^{m_3} \times \mathbb{R}^{m_3}$ to the system

$$\begin{bmatrix} y_1^\top & y_2^\top & y_3^\top & y_4^\top \end{bmatrix} \begin{bmatrix} A \\ B \\ C \\ -C \end{bmatrix} = \mathbf{0}^\top$$

$$\begin{bmatrix} y_1^\top & y_2^\top & y_3^\top & y_4^\top \end{bmatrix} \begin{bmatrix} -\mathbf{1} \\ \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix} < 0$$

$$y_1, y_2, y_3, y_4 \geq \mathbf{0}$$

Writing out these (in)equalities and replacing the difference of the nonnegative vectors $y_3 - y_4$ by an unconstrained vector y_3, this means that there is no solution (y_1, y_2, y_3) to $y_1^\top A + y_2^\top B + y_3^\top C = \mathbf{0}^\top, \mathbf{0} \neq y_1 \geq \mathbf{0}, y_2 \geq \mathbf{0}$.

20.4 With some sign changes and in matrix notation, we must show that there exist $(x, y) \in \mathbb{R}^n \times \mathbb{R}^m$ with

$$\begin{bmatrix} -A & -I \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} < \mathbf{0}$$

$$\begin{bmatrix} -A & O \\ O & -I \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \leq \mathbf{0}$$

$$\begin{bmatrix} O & A^\top \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \mathbf{0}$$

Suppose no such x and y exist. Then Exercise 20.3 implies that there are $z_1, z_2, z_3 \in \mathbb{R}^m, z_4 \in \mathbb{R}^n$ with

$$z_1^\top \begin{bmatrix} -A & -I \end{bmatrix} + \begin{bmatrix} z_2^\top & z_3^\top \end{bmatrix} \begin{bmatrix} -A & O \\ O & -I \end{bmatrix} + z_4^\top \begin{bmatrix} O & A^\top \end{bmatrix} = \mathbf{0}^\top$$

and $\mathbf{0} \neq z_1 \geq \mathbf{0}, z_2, z_3 \geq \mathbf{0}$. Writing out the matrix product, this means that the z_i satisfy

$$(z_1 + z_2)^\top A = \mathbf{0}^\top, Az_4 = z_1 + z_3, \mathbf{0} \neq z_1 \geq \mathbf{0}, z_2, z_3 \geq \mathbf{0}.$$

But this gives a contradiction: on one hand $\underbrace{(z_1 + z_2)^\top A}_{=\mathbf{0}^\top} z_4 = 0$, whereas on the other

$$(z_1 + z_2)^\top Az_4 = (z_1 + z_2)^\top (z_1 + z_3) = \underbrace{z_1^\top z_1}_{>0} + \underbrace{z_1^\top z_3}_{\geq 0} + \underbrace{z_2^\top z_1}_{\geq 0} + \underbrace{z_2^\top z_3}_{\geq 0} > 0.$$

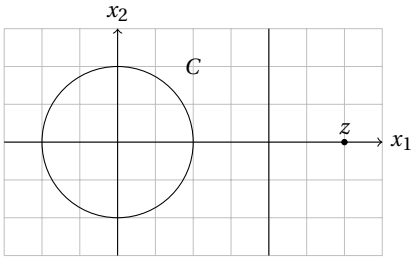
21.1 By the parallelogram law:

$$\begin{aligned}\|x-z\|^2 &= \|\tfrac{1}{2}(x_1+x_2)-z\|^2 = \|\tfrac{1}{2}(x_1-z)+\tfrac{1}{2}(x_2-z)\|^2 \\ &= 2\|\tfrac{1}{2}(x_1-z)\|^2+2\|\tfrac{1}{2}(x_2-z)\|^2-\|\tfrac{1}{2}(x_1-z)-\tfrac{1}{2}(x_2-z)\|^2 = \tfrac{1}{2}\|x_1-z\|^2+\tfrac{1}{2}\|x_2-z\|^2-\tfrac{1}{4}\|x_1-x_2\|^2.\end{aligned}$$

If $x_1 \neq x_2$, this is smaller than the squared distance of x_1 or x_2 to z.

21.2 (a) Both $x=(1,0)$ and $y=(-1,0)$ belong to C, but convex combination $\frac{1}{2}x+\frac{1}{2}y=(0,0)$ does not.

(b) $z=(3,0)$ can be separated from C by the hyperplane $\{x\in\mathbb{R}^2 : x_1=2\}$, i.e., the hyperplane $H=\{x\in\mathbb{R}^2 : c^\top x=\delta\}$ with normal $c=(1,0)$ and $\delta=2$.



(c) $z=(0,0)$ cannot be separated from C by a hyperplane. Suppose, to the contrary, that we can find a normal $c\neq\mathbf{0}$ with $c^\top x\leq c^\top z=0$ for all $x\in C$. This leads to a contradiction: considering the four elements $(1,0),(-1,0),(0,1)$, and $(0,-1)$ of C, normal c must satisfy

$$\begin{array}{lcrl} c^\top(1,0) & = & c_1 & \leq 0, \\ c^\top(-1,0) & = & -c_1 & \leq 0, \\ c^\top(0,1) & = & c_2 & \leq 0, \\ c^\top(0,-1) & = & -c_2 & \leq 0, \end{array}$$

so $c=(0,0)=\mathbf{0}$. But by definition of a hyperplane, its normal c is not allowed to be the zero vector.

(d) $C_1\cap C_2=\{x\in\mathbb{R}^2 : x_1=0, x_2=0\}=\{(0,0)\}$: their intersection is the origin of $\mathbb{R}^2$.
C_1 and C_2 cannot be separated by a hyperplane. Suppose, to the contrary, that we can find a normal $c\neq\mathbf{0}$ with $c^\top x\leq c^\top y$ for all $x\in C_1$ and $y\in C_2$.

$$\left\{\begin{array}{ll} x=(0,0)\in C_1, & y=(1,0)\in C_2 \\ x=(0,0)\in C_1, & y=(-1,0)\in C_2 \\ x=(0,1)\in C_1, & y=(0,0)\in C_2 \\ x=(0,-1)\in C_1, & y=(0,0)\in C_2 \end{array}\right. \text{ gives } \left\{\begin{array}{rcrcl} c^\top x= & 0 & \leq & c_1 & =c^\top y, \\ c^\top x= & 0 & \leq & -c_1 & =c^\top y, \\ c^\top x= & c_2 & \leq & 0 & =c^\top y, \\ c^\top x= & -c_2 & \leq & 0 & =c^\top y, \end{array}\right.$$

so $c=(0,0)=\mathbf{0}$. But by definition of a hyperplane, its normal c is not allowed to be the zero vector.

(e) Let $A=\left[\begin{smallmatrix}2 & -1\\ 2 & -2\end{smallmatrix}\right]$ be the matrix with v_1 and v_2 as its columns. Then z belongs to $\text{cone}\{v_1,v_2\}$ if and only if there is a nonnegative solution $x\geq\mathbf{0}$ to $Ax=z$. To show that no such solution exists, by Farkas' Lemma (Theorem 20.1), is equivalent to showing that there is a vector y with $y^\top A\geq\mathbf{0}^\top$ and $y^\top z<0$. This is equivalent to solving the system

$$\left\{\begin{array}{l} 2y_1+2y_2\geq 0 \\ -y_1-2y_2\geq 0 \\ -y_1+y_2<0 \end{array}\right.$$

Vector $y=(3,-2)$ is one solution to this system. Taking this as the normal, $H=\{x\in\mathbb{R}^2 : y^\top x=0\}=\{x\in\mathbb{R}^2 : 3x_1-2x_2=0\}$ is a hyperplane that *weakly* separates z from $\text{cone}\{v_1,v_2\}$: since $y^\top v_1\geq 0$ and $y^\top v_2\geq 0$, it follows that

$$y^\top z=3\cdot(-1)-2\cdot 1=-5<0\leq y^\top v \qquad \text{for all } v\in\text{cone}\{v_1,v_2\}.$$

For $-5<\delta<0$, the previous line shows that hyperplane $H=\{x\in\mathbb{R}^2 : y^\top x=\delta\}=\{x\in\mathbb{R}^2 : 3x_1-2x_2=\delta\}$ *strictly* separates z from $\text{cone}\{v_1,v_2\}$.

22.2 (a) Function $f:\mathbb{R}\to\mathbb{R}$ with $f(x)=0$ if $x\le 0$ and $f(x)=1$ otherwise is weakly increasing, hence quasiconcave. It is not continuous at $x=0$.

(b) Arguing as above, the functions f and g with $f(x)=\max\{0,x\}$ and $g(x)=\max\{0,-x\}$ are quasiconcave. Their sum, the function h with $h(x)=\max\{x,-x\}=|x|$ is not quasiconcave: for each $r>0$ the set

$$\{x\in\mathbb{R}: f(x)\ge r\}=(-\infty,-r]\cup[r,\infty)$$

is not convex.

(c) The function $f:[0,\infty)\to\mathbb{R}$ with $f(x)=x^2$ is increasing on $[0,\infty)$, hence quasiconcave. Its second derivative is $f''(x)=2\ge 0$, so it is convex. It is not concave since

$$f(2)=f\left(\frac{1}{2}\cdot 0+\frac{1}{2}\cdot 4\right)=4<8=\frac{1}{2}f(0)+\frac{1}{2}f(4).$$

22.3 (a) No: the linear function $f:\mathbb{R}\to\mathbb{R}$ with $f(x)=x$ is concave and the function $g:\mathbb{R}\to\mathbb{R}$ with $g(x)=x^3$ is strictly increasing, but their composition $g\circ f$ is equal to g, which is not concave (its second derivative has $g''(x)=6x$ and is not everywhere less than or equal to zero).

(b) Yes: let $x,y\in C$ and $\lambda\in[0,1]$. To show:

$$(g\circ f)\big(\lambda x+(1-\lambda)y\big)\ge\min\big\{(g\circ f)(x),(g\circ f)(y)\big\}.$$

Without loss of generality, $f(x)\le f(y)$. Since g is strictly increasing, $(g\circ f)(x)=g(f(x))\le g(f(y))=(g\circ f)(y)$, so $\min\big\{(g\circ f)(x),(g\circ f)(y)\big\}=g(f(x))$.

Since f is quasiconcave,

$$f\big(\lambda x+(1-\lambda)y\big)\ge\min\{f(x),f(y)\}=f(x).$$

Since g is strictly increasing,

$$(g\circ f)\big(\lambda x+(1-\lambda)y\big)=g\big(f\big(\lambda x+(1-\lambda)y\big)\big)\ge g\big(f(x)\big)=\min\big\{(g\circ f)(x),(g\circ f)(y)\big\},$$

as we set out to prove.

22.4 Function $f:\mathbb{R}\to\mathbb{R}$ with $f(x)=\max\{0,x\}$ is weakly increasing, hence quasiconcave. Each point $x<0$ is a local maximum because all points in its neighborhood $(-\infty,0)$ have the same function value 0, so $f(x)\ge f(y)$ for all $y\in(-\infty,0)$. Such an x is not a global maximum as each $z>0$ has a higher function value: $f(x)=0<z=f(z)$.

22.5 (a) Suppose, to the contrary, that x is not a global maximum: there is a $z\in C$ with $f(z)>f(x)$.

Since $\lambda x+(1-\lambda)z$ tends to x as $\lambda\in(0,1)$ tends to 1, we can choose a $\lambda\in(0,1)$ sufficiently close to 1 so that $\lambda x+(1-\lambda)z$ lies in the neighborhood U of x. This implies that $f(x)>f(\lambda x+(1-\lambda)z)$.

But quasiconcavity implies the opposite: $f(\lambda x+(1-\lambda)z)\ge\min\{f(x),f(z)\}=f(x)$, a contradiction.

(b) No. The set of global maxima is convex by Theorem 22.14. So if z were another global maximum, each point $\lambda x+(1-\lambda)z$ on the linepiece between x and z would be a global maximum as well. But as we argued above, for λ close to 1 such points lie in the neighborhood U of x and therefore have a lower function value than x, a contradiction.

22.6 For each $z\in C$, define the auxiliary function $h_z:C\to\mathbb{R}$ by $h_z(x)=f(z)+a_z^\top(x-z)$. This function is affine, hence convex. By assumption, f lies above these tangents: $f\ge h_z$ for each z. Moreover, in the point x, the functions f and h_x achieve the same value: $f(x)=h_x(x)$. So $f=\sup_{z\in C}h_z$ is the pointwise supremum of the convex functions h_z, hence convex itself by Theorem thm: constructing convex functions.

22.7 ($\Rightarrow$): as in Theorem 22.1, but with a strict inequality.
($\Leftarrow$): Assume that the set $\{(x,t)\in C\times\mathbb{R}: t>f(x)\}$ is convex. To see that f is convex, let $x,y\in C$ and $\lambda\in[0,1]$. I show:

$$\text{for each } \varepsilon>0:\qquad f(\lambda x+(1-\lambda)y)<\lambda f(x)+(1-\lambda)f(y)+\varepsilon. \tag{169}$$

This implies that f is convex, i.e., that

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y).$$

Indeed, if we'd have $f(\lambda x + (1-\lambda)y) > \lambda f(x) + (1-\lambda) f(y)$, then $\varepsilon = f(\lambda x + (1-\lambda)y) - \big(\lambda f(x) + (1-\lambda) f(y)\big) > 0$ and substituting this into (169) gives a contradiction.

To see that (169) holds, note that for each $\varepsilon > 0$ points $(x, f(x)+\varepsilon)$ and $(y, f(y)+\varepsilon)$ lie in $\{(x,t) \in C \times \mathbb{R} : t > f(x)\}$. By convexity of this set, so does

$$\lambda\big(x, f(x)+\varepsilon\big) + (1-\lambda)\big(y, f(y)+\varepsilon\big) = \big(\lambda x + (1-\lambda)y, \lambda f(x) + (1-\lambda) f(y) + \varepsilon\big),$$

which is equivalent with (169).

22.8 First, assume that f is concave. Let $a \in \mathbb{R}^n$. Function g is the sum of the concave function f and the linear, hence concave function $x \mapsto a^\top x$, so g is concave. In particular, g is quasiconcave.

Next, assume that for each vector $a \in \mathbb{R}^n$, the function $g : C \to \mathbb{R}$ with $g(x) = f(x) + a^\top x$ is quasiconcave. To see that f is concave, choose distinct x and y in C and $\lambda \in (0,1)$. Tilt f by choosing an $a \in \mathbb{R}^n$ so that $g(x) = g(y)$, i.e.,

$$f(x) + a^\top x = f(y) + a^\top y. \tag{170}$$

By quasiconcavity of $x \mapsto f(x) + a^\top x$,

$$\begin{aligned} f(\lambda x + (1-\lambda)y) + a^\top\left[\lambda x + (1-\lambda)y\right] &\geq \min\left\{f(x) + a^\top x, f(y) + a^\top y\right\} \\ &= \lambda\left[f(x) + a^\top x\right] + (1-\lambda)\left[f(y) + a^\top y\right], \qquad \text{(using (170))} \end{aligned}$$

whence

$$f(\lambda x + (1-\lambda)y) \geq \lambda f(x) + (1-\lambda) f(y),$$

making f concave.

It remains to pick a such that (170) holds. Since $x \neq y$, there is a coordinate i with $x_i \neq y_i$, so we can take

$$a = \frac{f(y) - f(x)}{x_i - y_i} e_i.$$

24.1 (a) ☒ The inequalities in standard form are $h_1(x) = x_1^3 - x_2 \leq 0$ and $h_2(x) = x_2 - 3x_1 - 2 \leq 0$ with gradients $\nabla h_1(x) = (3x_1^2, -1)$ and $\nabla h_2(x) = (-3, 1)$.

☒ In feasible points where only one constraint h_j is binding, the corresponding gradient $\nabla h_j(x)$ is distinct from the zero vector, so the set $\{\nabla h_j(x)\}$ is linearly independent: no such point belongs to X_{LD}.

☒ If both h_1 and h_2 are binding, then $h_1(x) = x_1^3 - x_2 = 0$ and $h_2(x) = x_2 - 3x_1 - 2 = 0$, so $x_1^3 = 3x_1 + 2$. Rewriting, $x_1^3 - 3x_1 - 2 = (x_1+1)^2(x_1 - 2) = 0$ has two solutions, $x_1 = -1$ and $x_1 = 2$.

CASE 1: $x_1 = -1$ gives $x_2 = x_1^3 = -1$. At $x = (-1,-1)$ the set of gradients of the binding constraints is $\{\nabla h_1(-1,-1), \nabla h_2(-1,-1)\} = \{(3,-1),(-3,1)\}$, which is linearly dependent, since $(3,-1)+(-3,1) = (0,0)$. So $(-1,-1) \in X_{LD}$.

CASE 2: $x_1 = 2$ gives $x_2 = x_1^3 = 8$. At $x = (2,8)$ the set of gradients of the binding constraints is $\{\nabla h_1(2,8), \nabla h_2(2,8)\} = \{(12,-1),(-3,1)\}$, which is linearly independent, since solving

$$\alpha(12,-1) + \beta(-3,1) = (0,0)$$

gives $\alpha = \beta = 0$.

☒ Conclude: only the point $(-1,-1)$, where both gradients are binding, belongs to X_{LD}.

(b) ☒ The inequalities in standard form are $h_1(x) = -x_2 \leq 0$, $h_2(x) = x_2 - x_1^3 \leq 0$, $h_3(x) = 2x_1^3 - 6x_1^2 + 12x_1 - 8 - x_2 \leq 0$ with gradients $\nabla h_1(x) = (0,-1)$, $\nabla h_2(x) = (-3x_1^2, 1)$, $\nabla h_3(x) = (6x_1^2 - 12x_1 + 12, -1)$.

☒ In feasible points where only one constraint h_j is binding, the corresponding gradient $\nabla h_j(x)$ is distinct from the zero vector, so the set $\{\nabla h_j(x)\}$ is linearly independent: no such point belongs to X_{LD}.

☒ Next, consider feasible points where two constraints are binding:

CASE 1: h_1 and h_2 are binding: $h_1(x) = -x_2 = 0, h_2(x) = x_2 - x_1^3 = 0$ gives $x = (0,0)$. At $x = (0,0)$, the set of gradients of the binding constraints is $\{\nabla h_1(0,0), \nabla h_2(0,0)\} = \{(0,-1),(0,1)\}$, which is linearly dependent since $(0,-1)+(0,1) = (0,0)$. So $(0,0) \in X_{LD}$.

CASE 2: h_1 and h_3 are binding: $h_1(x) = -x_2 = 0, h_3(x) = 2x_1^3 - 6x_1^2 + 12x_1 - 8 - x_2 = 0$ gives $x_2 = 0$ and $2x_1^3 - 6x_1^2 + 12x_1 - 8 = 2(x_1-1)(x_1^2 - 2x_1 + 4) = 2(x_1-1)((x_1-1)^2 + 3) = 0$, so $x_1 = 1$. At $x = (1,0)$, the set of gradients of the binding constraints is $\{\nabla h_1(1,0), \nabla h_3(1,0)\} = \{(0,-1),(6,-1)\}$, which is linearly independent, since solving

$$\alpha(0,-1) + \beta(6,-1) = (0,0)$$

gives $\alpha = \beta = 0$.

CASE 3: h_2 and h_3 are binding: $h_2(x) = x_2 - x_1^3 = 0, h_3(x) = 2x_1^3 - 6x_1^2 + 12x_1 - 8 - x_2 = 0$ gives $x_2 = x_1^3 = 2x_1^3 - 6x_1^2 + 12x_1 - 8$. So $x_1^3 - 6x_1^2 + 12x_1 - 8 = (x_1-2)^3 = 0$ gives $x_1 = 2$ and $x_2 = x_1^3 = 8$. At $x = (2,8)$, the set of gradients of the binding constraints is $\{\nabla h_2(2,8), \nabla h_3(2,8)\} = \{(-12,1),(12,-1)\}$, which is linearly dependent since $(-12,1)+(12,-1) = (0,0)$. So $(2,8) \in X_{LD}$.

☒ Next, consider feasible points where all three constraints are binding. No such points exist: the first two constraints require $x_1 = x_2 = 0$, violating the third constraint.

☒ Conclude: $X_{LD} = \{(0,0),(2,8)\}$.

24.2 Argue yourself that in all three cases the Extreme Value Theorem (Thm. 17.3) assures that a solution exists.

(a) The problem in standard form is

$$\begin{array}{ll} \text{maximize} & f(x) = (x_1+x_2)^2 + 2x_1 + x_2^2 \\ \text{with} & h_1(x) = -x_1 \le 0 \\ & h_2(x) = -x_2 \le 0 \\ & h_3(x) = x_1 + 3x_2 - 4 \le 0 \\ & h_4(x) = 2x_1 + x_2 - 3 \le 0 \end{array}$$

We can apply case 3 of Theorem 24.4: the constraints are affine functions. So a maximum must satisfy the KKT conditions. Assigning multipliers $\mu_1, \ldots, \mu_4$ to the constraints, the Lagrangian is

$$\mathscr{L}(x,\mu) = (x_1+x_2)^2 + 2x_1 + x_2^2 + \mu_1 x_1 + \mu_2 x_2 - \mu_3(x_1 + 3x_2 - 4) - \mu_4(2x_1 + x_2 - 3).$$

The KKT conditions require that in a solution, its partial derivatives w.r.t. x_1 and x_2 are zero:

$$\begin{aligned} 2(x_1+x_2) + 2 + \mu_1 - \mu_3 - 2\mu_4 &= 0 \\ 2(x_1+x_2) + 2x_2 + \mu_2 - 3\mu_3 - \mu_4 &= 0 \end{aligned}$$

and that the feasibility and complementary slackness conditions hold. So $\mu_3 + 2\mu_4 = 2(x_1+x_2) + 2 + \mu_1 \ge 2$: at least one of μ_3, μ_4 is positive. By complementary slackness, at least one of the constraints h_3 and h_4 is binding. So consider two cases.

First, maximum candidates where the third constraint is binding: $x_1 = 4 - 3x_2$. Plugging this into the constraints gives that x_2 must satisfy $1 \le x_2 \le 4/3$ and maximize $f(4-3x_2, x_2) = 5x_2^2 - 22x_2 + 24$. Comparing boundary points 1 and 4/3 and potential interior solutions, we see that the maximum is achieved at $x_2 = 1$ and that the KKT conditions are satisfied by $x = (1,1)$ with multipliers $\mu = (0,0,6/5,12/5)$.

Second, maximum candidates where the fourth constraint is binding: $x_2 = 3 - 2x_1$. Plugging this into the constraints gives that x_1 must satisfy $1 \le x_1 \le 3/2$ and maximize $f(x_1, 3-2x_1) = 5x_1^2 - 16x_1 + 18$, which leads to the same unique candidate $x = (1,1)$ as above.

Therefore, the goal function is maximized in $x = (1,1)$.

COMMENT: I solved the problem by searching for points that satisfy the KKT conditions *and* were maxima on part of the domain. A more traditional way would be to find all solutions to the KKT conditions and only then figure out which are maxima. In this example, that's a bit more time consuming; pretty much by distinguishing the same cases as above, it turns out that only $x = (1,1)$ satisfies the KKT conditions.

(b) A solution to the minimization problem must solve the following *maximization* problem in standard form:

$$\begin{array}{ll} \text{maximize} & f(x) = -4x_1 + 3x_2 \\ \text{with} & h_1(x) = x_1 + x_2 - 4 \leq 0 \\ & h_2(x) = -x_2 - 7 \leq 0 \\ & h_3(x) = (x_1 - 3)^2 - x_2 - 1 \leq 0 \end{array}$$

We can apply case 2 of Theorem 24.4: the constraints are convex functions and hold with strict inequality in $x_0 = (3,0)$. So a maximum must satisfy the KKT conditions. Assigning multipliers μ_1, μ_2, μ_3 to the constraints, the Lagrangian is

$$\mathscr{L}(x,\mu) = -4x_1 + 3x_2 - \mu_1(x_1 + x_2 - 4) - \mu_2(-x_2 - 7) - \mu_3((x_1 - 3)^2 - x_2 - 1).$$

The KKT conditions require that in a solution, its partial derivatives w.r.t. x_1 and x_2 are zero:

$$\begin{aligned} -4 - \mu_1 - 2\mu_3(x_1 - 3) &= 0 \\ 3 - \mu_1 + \mu_2 + \mu_3 &= 0 \end{aligned}$$

and that the feasibility and complementary slackness conditions hold. So $\mu_1 = 3 + \mu_2 + \mu_3 > 0$. By complementary slackness, the first constraint is binding. So is the third. If not, complementary slackness gives $\mu_3 = 0$, so that $\mu_1 = -4$, contradicting $\mu_1 \geq 0$. Solving $h_1(x) = 0$ and $h_3(x) = 0$ gives two candidates: $x = (1,3)$ and $x = (4,0)$. Some linear algebra shows that only $x = (1,3)$ solves the KKT conditions (for $\mu_1 = 16/3, \mu_2 = 0, \mu_3 = 7/3$): this is the desired optimum.

(c) The problem in standard form is:

$$\begin{array}{ll} \text{maximize} & f(x) = x_2 - 2x_1^3 + 2x_1^2 - x_1 \\ \text{with} & h_1(x) = -x_2 \leq 0 \\ & h_2(x) = x_2 - x_1^3 \leq 0 \\ & h_3(x) = 2x_1^3 - 6x_1^2 + 12x_1 - 8 - x_2 \leq 0 \end{array}$$

Let's solve the KKT conditions. Assigning multipliers μ_1, μ_2, μ_3 to the constraints, the Lagrangian is

$$\mathscr{L}(x,\mu) = x_2 - 2x_1^3 + 2x_1^2 - x_1 + \mu_1 x_2 - \mu_2(x_2 - x_1^3) - \mu_3(2x_1^3 - 6x_1^2 + 12x_1 - 8 - x_2).$$

The KKT conditions require that its partial derivatives w.r.t. x_1 and x_2 are zero:

$$-6x_1^2 + 4x_1 - 1 + 3\mu_2 x_1^2 - \mu_3(6x_1^2 - 12x_1 + 12) = 0 \tag{171}$$

$$1 + \mu_1 - \mu_2 + \mu_3 = 0 \tag{172}$$

and that the feasibility and complementary slackness conditions hold. So $\mu_2 = 1 + \mu_1 + \mu_3 > 0$. By complementary slackness, the second constraint is binding: $x_2 = x_1^3$. The first constraint then gives $x_1 \geq 0$ and the third gives

$$2x_1^3 - 6x_1^2 + 12x_1 - 8 - x_1^3 = x_1^3 - 6x_1^2 + 12x_1 - 8 = (x_1 - 2)^3 \leq 0,$$

so $x_1 \leq 2$. So solutions to the KKT conditions must be of the form (x_1, x_1^3) with $0 \leq x_1 \leq 2$.

If $x_1 = 0$, the KKT conditions cannot be satisfied: (171) gives $\mu_3 = -1/12$, contradicting $\mu_3 \geq 0$.

If $x_1 = 2$, the KKT conditions cannot be satisfied: (171) becomes $\mu_2 - \mu_3 = 17/12$, whereas (172) becomes $\mu_2 - \mu_3 = 1$ and we cannot have both!

If $0 < x_1 < 2$, complementary slackness gives $\mu_1 = \mu_3 = 0$ and with (172) that $\mu_2 = 1$. Substituting these into (171), we must have

$$0 = -6x_1^2 + 4x_1 - 1 + 3x_1^2 = -3x_1^2 + 4x_1 - 1 = (x_1 - 1)(-3x_1 + 1),$$

so $x_1 = 1$ or $x_1 = 1/3$.

So the KKT conditions hold in two points, $x = (1,1)$ and $x = (1/3, 1/27)$, both with $(\mu_1, \mu_2, \mu_3) = (0,1,0)$.

From Exercise 24.1, we have two more candidates, $x = (0,0)$ and $x = (2,8)$, where the gradients of the binding constraints are linearly dependent. Comparing all their function values, we find that $f(1,1) = 0, f(1/3, 1/27) = -4/27, f(0,0) = 0, f(2,8) = -2$. So the maxima are achieved at $(0,0)$ and $(1,1)$.

24.3 (a) A solution to the minimization problem must solve the following maximization problem in standard form:

$$\begin{array}{ll} \text{maximize} & f(x) = -x_1^2 + x_2^2 - 4x_3^2 \\ \text{with} & h_1(x) = -1 - x_2 \leq 0 \\ & h_2(x) = 1 - x_1 - x_3 \leq 0 \\ & h_3(x) = -10 - x_3 \leq 0 \end{array}$$

Assigning multipliers μ_1, μ_2, μ_3 to the constraints, its Lagrangian is

$$\mathscr{L}(x, \mu) = -x_1^2 + x_2^2 - 4x_3^2 - \mu_1(-1 - x_2) - \mu_2(1 - x_1 - x_3) - \mu_3(-10 - x_3).$$

The KKT conditions are:

☒ Partial derivatives of the Lagrangian w.r.t. x_1, x_2, and x_3 are zero:

$$\begin{aligned} -2x_1 + \mu_2 &= 0 \\ 2x_2 + \mu_1 &= 0 \\ -8x_3 + \mu_2 + \mu_3 &= 0 \end{aligned}$$

Equivalently:

$$\mu_1 = -2x_2, \qquad \mu_2 = 2x_1, \qquad \mu_3 = 8x_3 - 2x_1. \tag{173}$$

☒ Feasibility: $h_i(x) \leq 0$ for $i = 1, 2, 3$.

☒ Complementary slackness: $\mu_i \geq 0$ and $\mu_i h_i(x) = 0$ for $i = 1, 2, 3$.

Point $x = (4/5, 0, 1/5)$ is feasible and (173) requires $(\mu_1, \mu_2, \mu_3) = (0, 8/5, 0)$. It follows by substitution that complementary slackness is satisfied. So the KKT conditions hold in this point.

Point $x = (4/5, -1, 1/5)$ is feasible and (173) requires $(\mu_1, \mu_2, \mu_3) = (2, 8/5, 0)$. Again, substitution shows that complementary slackness is satisfied. So the KKT conditions hold in this point.

Point $x = (2, -1, -1)$ is feasible and (173) requires $(\mu_1, \mu_2, \mu_3) = (2, 4, -12)$, contradicting that μ_3 must be nonnegative. So the KKT conditions do not hold in this point.

(b) No: there is no minimum. For each $x_2 \geq -1$, the point $x = (1, x_2, 0)$ is feasible and $x_1^2 - x_2^2 + 4x_3^2 = 1 - x_2^2$ can be made arbitrarily small by letting x_2 tend to infinity.

24.4 (a) A solution to the minimization problem must solve the following maximization problem in standard form:

$$\begin{array}{ll} \text{maximize} & f(x) = -(x_1 - x_2 + x_3)^2 \\ \text{with} & g_1(x) = x_1 + 2x_2 - x_3 - 5 = 0 \\ & g_2(x) = x_1 - x_2 - x_3 - 1 = 0 \end{array}$$

The gradient $\nabla f(x) = -2(x_1 - x_2 + x_3)(1, -1, 1)$ at $(3/2, 2, 1/2)$ equals $(0, 0, 0)$ and those of the constraints are $\nabla g_1(x) = (1, 2, -1)$ and $\nabla g_2(x) = (1, -1, -1)$. If $\mu_0 = 1, \lambda_1 = \lambda_2 = 0$, we get $\mu_0 \nabla f(x) - \lambda_1 \nabla g_1(x) - \lambda_2 \nabla g_2(x) = \mathbf{0}$: the Fritz John conditions are satisfied.

(b) Yes: the point is feasible and has function value $(x_1 - x_2 + x_3)^2 = 0$. Since a square of a real number is always nonnegative, 0 is indeed the minimal value of the goal function.

As an aside, this is the *only* solution to the problem: Gaussian elimination shows that the feasible points are all points of the form $x = (x_1, 2, x_1 - 1)$ with $x_1 \in \mathbb{R}$. So we can rewrite the minimization problem as: minimize $(x_1 - x_2 + x_3)^2 = (x_1 - 2 + x_1 - 1)^2 = (2x_1 - 3)^2$, with a unique solution at $x_1 = 3/2$.

24.5 (a) For each $x_1 \in \mathbb{R}$, the point $x = (x_1, 1, 5 + x_1)$ is feasible. Its function value $x_1^2 + 1^2 + (5 + x_1)^2$ can be made arbitrarily large by letting x_1 go to infinity.

(b) Argue as in the first step of the proof of Theorem 21.1.

(c) A solution to the minimization problem must solve the following maximization problem in standard form:

$$\begin{array}{ll} \text{maximize} & f(x) = -x_1^2 - x_2^2 - x_3^2 \\ \text{with} & g(x) = x_3 - x_1 x_2 - 5 = 0 \end{array}$$

We apply Theorem 24.7. We don't need to check the conditions in (b) and (c) of that theorem, since there are no inequality constraints. And the gradient $\nabla g(x) = (-x_2, -x_1, 1)$ of g does not equal the zero vector, so there are no feasible points where the gradient(s) of the binding constraint(s) are linearly dependent. Hence, it suffices to solve the conditions in (a): in a solution x, there must be a λ different from zero such that $\nabla f(x) - \lambda \nabla g(x) = \mathbf{0}$. Equivalently:

$$-2x_1 + \lambda x_2 = 0 \tag{174}$$
$$-2x_2 + \lambda x_1 = 0 \tag{175}$$
$$-2x_3 - \lambda \quad = 0 \tag{176}$$

Consider two cases:

CASE 1: $x_1 = 0$. Then (175) gives $x_2 = 0$, feasibility gives $x_3 = 5$, (176) gives $\lambda = -10$. So $x = (0,0,5)$ with function value $f(x) = -25$ is one solution candidate.

CASE 2: $x_1 \neq 0$. Then (175) gives $x_2 \neq 0$. Rewriting (174) and (175) gives $\frac{x_1}{x_2} = \frac{\lambda}{2}$ and $\frac{x_1}{x_2} = \frac{2}{\lambda}$, so $\frac{\lambda}{2} = \frac{2}{\lambda}$, so $\lambda \in \{-2, 2\}$.

If $\lambda = -2$, (176) gives $x_3 = 1$, (175) gives $x_2 = \frac{\lambda}{2} x_1 = -x_1$. By feasibility, $5 = x_3 - x_1 x_2 = 1 + x_1^2$, so $x_1 = -2$ or $x_1 = 2$. This gives two solution candidates $x = (-2, 2, 1)$ and $x = (2, -2, 1)$ with function value $-x_1^2 - x_2^2 - x_3^2 = -9$.

If $\lambda = 2$, (176) gives $x_3 = -1$, (175) gives $x_2 = \frac{\lambda}{2} x_1 = x_1$. By feasibility, $5 = x_3 - x_1 x_2 = -1 - x_1^2$, so $x_1^2 = -6$, which has no solution. This case gives no solution candidates.

Comparing the three candidates and translating everything back to the original *minimization* problem, we conclude that there are two minima, namely at $(-2, 2, 1)$ and $(2, -2, 1)$, both with function value $f(x) = 9$.

COMMENT: There are other ways to solve the FJ conditions, for instance by adding (174) and (175) to obtain $(2 + \lambda)(x_1 + x_2) = 0$ and distinguishing the two cases $2 + \lambda = 0$ and $x_1 + x_2 = 0$.

(d) For all (x_1, x_2) in $\mathbb{R}^2$ there is precisely one x_3 such that (x_1, x_2, x_3) is feasible: $x_3 = 5 + x_1 x_2$. Substituting this into the goal function, we can rewrite it to: minimize $x_1^2 + x_2^2 + (5 + x_1 x_2)^2$ over $\mathbb{R}^2$. The first-order conditions require its partial derivatives to be zero:

$$2x_1 + 2x_2(5 + x_1 x_2) = 0$$
$$2x_2 + 2x_1(5 + x_1 x_2) = 0$$

Multiply the first equation with x_1, the second with x_2, and subtract to obtain $2(x_1^2 - x_2^2) = 2(x_1 + x_2)(x_1 - x_2) = 0$. So $x_2 = -x_1$ or $x_2 = x_1$.

If $x_2 = -x_1$, we must have $2x_1 - 2x_1(5 - x_1^2) = 2x_1(x_1^2 - 4) = 0$, so $x_1 \in \{-2, 0, 2\}$. This gives three candidates, $x = (-2, 2, 1)$, $x = (0, 0, 5)$, and $x = (2, -2, 1)$ with function values 9, 25, and 9, respectively.

If $x_2 = x_1$, we must have $2x_1 + 2x_1(5 + x_1^2) = 2x_1(6 + x_1^2) = 0$, so $x_1 = 0$, again giving candidate $x = (0, 0, 5)$ with function value 25.

Comparing these candidates, conclude that there are two minima at $x = (-2, 2, 1)$ and $x = (2, -2, 1)$.

24.6 If x_1 is the length and x_2 the width of the rectangle, then its area is $x_1 x_2$ and its perimeter is $2x_1 + 2x_2$. So the problem is to maximize $x_1 x_2$ with $x_1 \geq 0, x_2 \geq 0, 2x_1 + 2x_2 = 1$.

The feasible set is nonempty and compact (verify) and the goal function is polynomial, hence continuous. So a solution exists by the Extreme Value Theorem.

Before doing any computations, note that in a maximum the inequality constraints cannot be binding: if length or width is zero, so is the area, which is clearly not maximal. By complementary slackness, the corresponding multipliers will be zero, so they drop out of the gradient expression in the Fritz John conditions. Let's apply Theorem 24.8. Assigning multiplier λ to the equality constraint, this allows us to simplify this expression to

$$\nabla f(x) - \lambda \nabla g(x) = \mathbf{0},$$

where $f(x) = x_1 x_2$ and $g(x) = 2x_1 + 2x_2 - 1$. Since f has gradient (x_2, x_1) and g has gradient $(2,2)$, this gives

$$\begin{aligned} x_2 - 2\lambda &= 0, \\ x_1 - 2\lambda &= 0, \end{aligned}$$

so x_1 and x_2 are equal. With a perimeter of one, we find that the rectangle has length and width equal to $1/4$.

The following answers are more concise than previous ones. Try out intermediate steps yourself.

24.7 (a) Maxima and minima exist by the Extreme Value Theorem. All feasible points are regular: the constraint $g(x) = x_1^2 + 4x_2^2 - 72 = 0$ has gradient $\nabla g(x) = (2x_1, 8x_2)$, which equals $\mathbf{0}$ if and only if $x = \mathbf{0}$, which is not feasible. So we may solve the FJ conditions with $\mu_0 = 1$. Since there are only equality constraints, the FJ conditions are the same for maxima and minima. We must solve $\nabla f(x) - \lambda \nabla g(x) = \mathbf{0}$ and $g(x) = 0$. The condition on the gradients is linear in x_1 and x_2:

$$\begin{aligned} 2(1-\lambda)x_1 + \quad & 6x_2 = 0 \\ 6x_1 + 8(1-\lambda)x_2 \quad & = 0 \end{aligned}$$

Gaussian elimination on the coefficient matrix gives

$$\begin{bmatrix} 6 & 8(1-\lambda) \\ 2(1-\lambda) & 6 \end{bmatrix} \sim \begin{bmatrix} 1 & \frac{4}{3}(1-\lambda) \\ 0 & 6 - \frac{8}{3}(1-\lambda)^2 \end{bmatrix}$$

If $6 - \frac{8}{3}(1-\lambda)^2 \neq 0$, the only solution is $x = \mathbf{0}$, which is not feasible. So $6 - \frac{8}{3}(1-\lambda)^2 = 0$, i.e., $\lambda \in \{-\frac{1}{2}, \frac{5}{2}\}$.

CASE 1: $\lambda = -\frac{1}{2}$ gives $x_1 + 2x_2 = 0$. Together with $x_1^2 + 4x_2^2 = 72$, this gives two solutions to the FJ conditions: $x = (6,-3)$ and $x = (-6,3)$, both with function value -36.

CASE 2: $\lambda = \frac{5}{2}$ gives $x_1 - 2x_2 = 0$. Together with $x_1^2 + 4x_2^2 = 72$, this gives two solutions to the FJ conditions: $x = (6,3)$ and $x = (-6,-3)$, both with function value 180.

Conclude: the set of points satisfying the FJ conditions is $\{(6,-3),(-6,3),(6,3),(-6,-3)\}$; the first two are minima, the last two are maxima.

(b) There is neither a maximum nor a minimum. If $x \in \mathbb{R}^3$ is feasible, then x_1 is arbitrary, $x_2 = x_1 + 1$, and $x_3 = -x_1 + 2x_2 - 3 = x_1 - 1$. Its function value

$$f(x) = f(x_1, x_1+1, x_1-1) = x_1^3 + (x_1+1)(x_1-1) = x_1^2(x_1+1) - 1$$

can be made arbitrarily large by letting x_1 go to infinity and arbitrarily small by letting x_1 go to minus infinity.

24.8 (a) Maxima and minima exist by the Extreme Value Theorem. Write the constraints as $g_1(x) = x_1^2 + 4x_2^2 - 4 = 0$ and $g_2(x) = x_1 + 2x_3 - 2 = 0$. The gradients $\nabla g_1(x) = (2x_1, 8x_2, 0)$ and $\nabla g_2(x) = (1,0,2)$ are linearly independent in each feasible point. So a feasible x satisfies the FJ conditions if there are λ_1 and λ_2 such that $\nabla f(x) - \lambda_1 \nabla g_1(x) - \lambda_2 \nabla g_2(x) = \mathbf{0}$:

$$\begin{aligned} 2x_1 - 2\lambda_1 x_1 - \quad & \lambda_2 = 0 \\ 4x_2 - 8\lambda_1 x_2 \quad & = 0 \\ 4x_3 \quad & - 2\lambda_2 = 0 \end{aligned}$$

CASE 1: $x_2 = 0$ gives two feasible points satisfying the FJ conditions: $x = (-2,0,2)$ with $(\lambda_1, \lambda_2) = (2,4)$ and $x = (2,0,0)$ with $(\lambda_1, \lambda_2) = (1,0)$.

CASE 2: $x_2 \neq 0$ gives $\lambda_1 = 1/2$ and, with some algebra, $\lambda_2 = 1, x_1 = 1, x_3 = 1/2$, and $x_2 \in \{-\sqrt{3}/2, \sqrt{3}/2\}$.

So the set of feasible points satisfying the FJ conditions is $\{(-2,0,2),(2,0,0),(1,-\sqrt{3}/2,1/2),(1,\sqrt{3}/2,1/2)\}$. Comparing their function values 12, 4, 3, and 3, we see that $(-2,0,2)$ is the maximum and $(1,-\sqrt{3}/2,1/2)$ and $(1,\sqrt{3}/2,1/2)$ are the minima.

(b) There is neither a maximum, nor a minimum. For each $\alpha \neq 0$, the point $(x_1, x_2, x_3, x_4) = (1, \alpha, -2/\alpha, 0)$ is feasible. Its function value is $f(1, \alpha, -2/\alpha, 0) = 1 + \alpha + 4/\alpha^2$. This value can be made arbitrarily large by letting α go to infinity and arbitrarily small by letting α go to minus infinity.

24.9 (a) A maximum exists by the Extreme Value Theorem. Since the constraints are affine, it must satisfy the KKT conditions. In the usual notation, the condition $\nabla f(x) - \sum_{i=1}^{4} \mu_i \nabla h_i(x) = \mathbf{0}$ becomes

$$\begin{aligned} 3\mu_1 + \mu_2 - \mu_3 &= 2(1 - x_1) \\ -2\mu_1 + \mu_2 - \mu_4 &= 1 \end{aligned}$$

By nonnegativity of the μ_i: $\mu_2 = 1 + 2\mu_1 + \mu_4 > 0$. By complementary slackness, the second constraint is binding: $x_1 + x_2 = 3$. By elementary linear algebra, at most two of the constraints can be binding: if two constraints bind, they determine a unique feasible point where the others are not binding (sketch feasible set!). So we find all x satisfying the KKT conditions by looking at 4 cases:

- ☒ No other constraints are binding: then $\mu = (0, 1, 0, 0)$ and $x = (1/2, 5/2)$.
- ☒ Constraints one and two binding: $x = (12/5, 3/5)$ gives $\mu = (-19/25, -13/25, 0, 0)$, contradicting $\mu_1, \mu_2 \geq 0$.
- ☒ Constraints two and three binding: $x = (0, 3)$ gives $\mu = (0, 1, -1, 0)$, contradicting $\mu_3 \geq 0$.
- ☒ Constraints two and four binding: $x = (3, 0)$ gives $\mu = (0, -4, 0, -5)$, contradicting $\mu_2, \mu_4 \geq 0$.

So only $x = (1/2, 5/2)$ satisfies the KKT conditions and must be the maximum.

(b) A maximum exists by the Extreme Value Theorem. Since the constraints are convex functions and satisfied with strict inequality in, for instance, $x = \mathbf{0}$, a maximum must satisfy the KKT conditions. In the usual notation, the condition $\nabla f(x) - \mu_1 \nabla h_1(x) - \mu_2 \nabla h_2(x) = \mathbf{0}$ becomes

$$\begin{aligned} \mu_1 + 2\mu_2 x_1 &= 3x_1^2 \\ 2\mu_2 x_2 &= 1 \\ 2\mu_1 x_3 &= 0 \end{aligned}$$

Together with the nonnegativity conditions on μ_i, we see that $\mu_2 > 0$ and $x_2 > 0$. By complementary slackness $x_1^2 + x_2^2 = \frac{2}{3}$. We find all points satisfying the KKT conditions by looking at two cases:

- ☒ If $\mu_1 > 0$, then $x_3 = 0$ and $x_1 + x_3^2 = 1$, so $x_1 = 1$, contradicting feasibility.
- ☒ If $\mu_1 = 0$, then $2\mu_2 x_1 = 3x_1^2$, so $x_1 = 0$ or $2\mu_2 = 3x_1$.

  If $x_1 = 0$, we find that all $x = \left(0, \sqrt{\frac{2}{3}}, x_3\right)$ with $x_3^2 \leq 1$ satisfy the KKT conditions with $\mu = \left(0, \frac{\sqrt{3}}{2\sqrt{2}}\right)$.

  If $2\mu_2 = 3x_1$, then $1 = 2\mu_2 x_2 = 3x_1 x_2$ gives $x_2 = \frac{1}{3x_1}$. So $x_1^2 + \frac{1}{9x_2^2} = \frac{2}{3}$. Solving this for x_1^2 (!) gives $x_1^2 = \frac{1}{3}$.

  Since x_1 must be nonnegative ($3x_1 = 2\mu_2 \geq 0$): $x_1 = \frac{1}{\sqrt{3}}$. All $x = \left(\frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}, x_3\right)$ with $x_3^2 \leq 1 - x_1 = 1 - \frac{1}{\sqrt{3}}$ solve the KKT conditions with $\mu = (0, \frac{1}{2}\sqrt{3})$.

So we find infinitely many solutions to the KKT conditions:

$$\left\{\left(0, \sqrt{\frac{2}{3}}, x_3\right) : x_3^2 \leq 1\right\} \cup \left\{\left(\frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}, x_3\right) : x_3^2 \leq 1 - \frac{1}{\sqrt{3}}\right\}.$$

Elements in the former set have function value $\sqrt{\frac{2}{3}} \approx 0.82$, those in the latter $\frac{4}{3\sqrt{3}} \approx 0.77$. So *all* the former are maxima.

(c) Write constraints $h_1(x) = x_1^2 + x_2^2 + x_3^2 - 5 \leq 0$ and $h_2(x) = x_1^2 + x_3^2 - 1 \leq 0$. The goal function is concave, the constraints convex, so the KKT conditions are necessary and sufficient for a maximum. The gradient requirement $\nabla f(x) - \mu_1 \nabla h_1(x) - \mu_2 \nabla h_2(x) = \mathbf{0}$ can be written as

$$\begin{aligned} 1 - 2\mu_1 x_1 - 2\mu_2 x_1 &= 0 \\ 2 - 2\mu_1 x_2 &= 0 \\ -2\mu_1 x_3 - 2\mu_2 x_3 &= 0 \end{aligned}$$

which together with the nonnegativity constraints on μ_1 and μ_2 gives that $\mu_1 > 0, x_2 > 0, x_1 > 0, x_3 = 0$.

CASE 1: $\mu_2 = 0$ gives $x_1/x_2 = 1/2$ and with $h_1(x) = 0$ that $x = (1,2,0)$ and $(\mu_1, \mu_2) = (1/2, 0)$.

CASE 2: $\mu_2 \neq 0$ implies that both constraints are binding. Again we find $x = (1,2,0)$ and $(\mu_1, \mu_2) = (1/2, 0)$, contradicting the assumption that $\mu_2 \neq 0$.

Conclude: only $x = (1,2,0)$ solves the KKT solutions and must be the maximum.

27.1 (a)
- ☒ Since the inner product is a linear function of its first argument we have, for each $w \in W$, that $\langle \mathbf{0}, w\rangle = 0$. So $\mathbf{0} \in W^\perp$.
- ☒ Let $v_1, v_2 \in W^\perp$. By linearity of the inner product in its first argument we have, for each $w \in W$, that $\langle v_1 + v_2, w\rangle = \langle v_1, w\rangle + \langle v_2, w\rangle = 0 + 0 = 0$, so $v_1 + v_2 \in W^\perp$.
- ☒ Likewise, if $v \in W^\perp$, then for each scalar α, $\alpha v \in W^\perp$.
- ☒ By Theorem 3.2, $W^\perp$ is a subspace of V.

(b) Consider a sequence of vectors $(v_k)_{k\in\mathbb{N}}$ in $W^\perp$ with limit $v \in V$. To show: $v \in W^\perp$. Well, recall that the inner product is a continuous function in its first argument, so for each $w \in W$:

$$0 = \lim_{k\to\infty} \langle v_k, w\rangle = \langle \lim_{k\to\infty} v_k, w\rangle = \langle v, w\rangle.$$

Therefore, $v \in W^\perp$.

(c) Since W and $W^\perp$ are both subspaces, they both contain the zero vector. Their intersection contains nothing else: if v lies in both W and $W^\perp$, it must be orthogonal to itself: $\langle v, v\rangle = 0$. By (I2), $v = \mathbf{0}$!

29.1 Eigenvalues are roots of the characteristic polynomial, so we show that A and $A^\top$ have the same characteristic polynomial:

$$\det(A - \lambda I) = \det(A^\top - \lambda I).$$

Using that the identity matrix is symmetric ($I = I^\top$), this follows from property (DET6):

$$\det(A^\top - \lambda I) = \det(A^\top - \lambda I^\top) = \det((A - \lambda I)^\top) = \det(A - \lambda I).$$

But A and $A^\top$ need not have the same eigenvectors. For instance, triangular matrix

$$A = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix},$$

has its only eigenvalue $\lambda = 0$ on the diagonal. Solving $(A - \lambda I)x = \mathbf{0}$ we see that each eigenvector is a multiple of $(0,1)$. But if we repeat this for $A^\top$, we see that each eigenvector is a multiple of $(1,0)$ instead.

29.2
- ☒ B is symmetric: $B^\top = \left(\frac{1}{2}A + \frac{1}{2}A^\top\right)^\top = \frac{1}{2}A^\top + \frac{1}{2}\left(A^\top\right)^\top = \frac{1}{2}A^\top + \frac{1}{2}A = \frac{1}{2}A + \frac{1}{2}A^\top = B$.
- ☒ For each $x \in \mathbb{R}^n$, $x^\top Ax$ is a 1×1 matrix and consequently its own transpose:

$$x^\top Ax = \left(x^\top Ax\right)^\top = x^\top A^\top x.$$

Hence

$$x^\top Bx = x^\top\left(\frac{1}{2}A + \frac{1}{2}A^\top\right)x = \frac{1}{2}x^\top Ax + \frac{1}{2}x^\top A^\top x = \frac{1}{2}x^\top Ax + \frac{1}{2}x^\top Ax = x^\top Ax.$$

B.1 (a) $v + w = (4 - i) + (-3 + 2i) = 1 + i$.

(b) $2w - iv = 2(-3 + 2i) - i(4 - i) = -6 + 4i - 4i + i^2 = -7$.

(c) $vw = (4 - i)(-3 + 2i) = -12 + 8i + 3i - 2i^2 = -10 + 11i$.

(d) $\overline{v} = \overline{4 - i} = 4 + i$.

(e) $\frac{v}{w} = \frac{4-i}{-3+2i} = \frac{4-i}{-3+2i} \cdot \frac{-3-2i}{-3-2i} = \frac{-12-8i+3i+2i^2}{13} = -\frac{14}{13} - \frac{5}{13}i$.

(f) $\dfrac{wi}{(1-i)v} = \dfrac{(-3+2i)i}{(1-i)(4-i)} = \dfrac{-2-3i}{4-i-4i+i^2} = \dfrac{-2-3i}{3-5i} = \dfrac{3+5i}{3+5i}\cdot\dfrac{-2-3i}{3-5i} = \dfrac{-6-9i-10i-15i^2}{9+25} = \dfrac{9-19i}{34}$
$= \dfrac{9}{34} - \dfrac{19}{34}i.$

B.2 Throughout the exercise we argue as follows: if

$$a+bi = r\left(\cos\varphi + i\sin\varphi\right),$$

then

$$r = \sqrt{a^2+b^2}, \qquad \cos\varphi = a/r, \qquad \sin\varphi = b/r.$$

(a) $r = \sqrt{48} = 4\sqrt{3}$, $\cos\varphi = \frac{1}{2}$, $\sin\varphi = \frac{-6}{4\sqrt{3}} = -\frac{1}{2}\sqrt{3}$, so $\varphi = \frac{5}{3}\pi$.

(b) $r = 1$, $\cos\varphi = -1$, $\sin\varphi = 0$, so $\varphi = \pi$.

(c) $r = 2$, $\cos\varphi = \frac{1}{2}$, $\sin\varphi = \frac{1}{2}\sqrt{3}$, so $\varphi = \frac{1}{3}\pi$.

(d) $r = 3\sqrt{2}$, $\cos\varphi = \sin\varphi = -\frac{1}{2}\sqrt{2}$, so $\varphi = \frac{5}{4}\pi$.

B.3 Write $z = a+bi$. Then $z^2 = a^2 - b^2 + (2ab)i$. If z^2 is real, then $2ab = 0$, so $a = 0$ or $b = 0$. Suppose $b \neq 0$. Then $a = 0$, so $z^2 = -b^2 < 0$, contradicting that $z^2 \geq 0$. So $b = 0$: z is a real number.

B.4 $\frac{a_1+b_1 i}{a_2+b_2 i} = \frac{a_1+b_1 i}{a_2+b_2 i}\cdot\frac{a_2-b_2 i}{a_2-b_2 i} = \frac{(a_1a_2+b_1b_2)+(b_1a_2-a_1b_2)i}{a_2^2+b_2^2} = \frac{a_1a_2+b_1b_2}{a_2^2+b_2^2} + \frac{b_1a_2-a_1b_2}{a_2^2+b_2^2}\,i.$

B.5 Write $z_1 = a+bi$ and $z_2 = c+di$. Then $|z_1| = \sqrt{a^2+b^2}$ is the square root of a nonnegative real number, hence nonnegative itself. It equals zero if and only if $a^2+b^2 = 0$. Since the square of a real number is nonnegative, that means that both a^2 and b^2, and consequently a and b themselves, must be zero. So $|z_1| = 0$ if and only if $z_1 = 0+0i = 0$. Next,

$$|z_1z_2| = |(a+bi)(c+di)| = |(ac-bd)+(ad+bc)i| = \sqrt{(ac-bd)^2+(ad+bc)^2}$$
$$= \sqrt{a^2c^2-2abcd+b^2d^2+a^2d^2+2abcd+b^2c^2} = \sqrt{(a^2+b^2)(c^2+d^2)} = \sqrt{a^2+b^2}\sqrt{c^2+d^2} = |z_1||z_2|.$$

Finally, vectors $v = (a,b)$ and $w = (c,d)$ in $\mathbb{R}^2$ satisfy the triangle inequality:

$$\|v+w\| \leq \|v\| + \|w\| \qquad \text{or equivalently} \qquad \sqrt{(a+c)^2+(b+d)^2} \leq \sqrt{a^2+b^2} + \sqrt{c^2+d^2}.$$

The inequality's left side equals $|z_1+z_2|$, its right side is $|z_1|+|z_2|$, so indeed $|z_1+z_2| \leq |z_1|+|z_2|$.

B.6 Using identity $e^ae^b = e^{a+b}$, we have

$$e^{i\varphi_1}e^{i\varphi_2} = e^{i(\varphi_1+\varphi_2)} = \cos(\varphi_1+\varphi_2) + i\sin(\varphi_1+\varphi_2).$$

And using polar coordinates,

$$\left(\cos\varphi_1 + i\sin\varphi_1\right)\left(\cos\varphi_2 + i\sin\varphi_2\right) = \left(\cos\varphi_1\cos\varphi_2 - \sin\varphi_1\sin\varphi_2\right) + i\left(\sin\varphi_1\cos\varphi_2 + \cos\varphi_1\sin\varphi_2\right).$$

The complex numbers in these two expressions are the same, so equating their real and imaginary parts gives (152) and (153).

Claim (154) is true by definition if $n = 1$. Now let $n \in \mathbb{N}$ and assume that

$$(a+bi)^n = r^n\left(\cos n\varphi + i\sin n\varphi\right). \tag{177}$$

Then

$$\begin{aligned}
(a+bi)^{n+1} &= (a+bi)^n(a+bi) \\
&= r^n\left(\cos n\varphi + i\sin n\varphi\right) r\left(\cos\varphi + i\sin\varphi\right) && \text{by (177)} \\
&= r^{n+1}\left((\cos n\varphi\cos\varphi - \sin n\varphi\sin\varphi) + i(\cos n\varphi\sin\varphi + \sin n\varphi\cos\varphi)\right) && \text{by (144)} \\
&= r^{n+1}\left(\cos(n+1)\varphi + i\sin(n+1)\varphi\right) && \text{by (152) and (153)}
\end{aligned}$$

So the claim is true for $n+1$. By induction, the claim is true for all n.

B.7 $\left|p(0)+\alpha^k u^k q(\alpha u)\right|^2 - \left|p(0)\right|^2$

$= \left(p(0)+\alpha^k u^k q(\alpha u)\right)\overline{\left(p(0)+\alpha^k u^k q(\alpha u)\right)} - p(0)\overline{p(0)}$

$= \alpha^k\left[p(0)\overline{u^k q(\alpha u)} + \overline{p(0)}u^k q(\alpha u)\right] + \alpha^{2k}\left[u^k q(\alpha u)\overline{u^k q(\alpha u)}\right]$ ($\overline{\alpha}=\alpha$ since α is real)

$= 2\alpha^k \operatorname{Re}\left[\overline{p(0)}u^k q(\alpha u)\right] + \alpha^{2k}\left|u^k q(\alpha u)\right|$ since $\operatorname{Re}(z_1 z_2) = \dfrac{z_1\overline{z_2}+\overline{z_1}z_2}{2}$

$= 2\alpha^k \operatorname{Re}\left[\overline{p(0)}u^k q(\alpha u)\right] + \alpha^{2k}\left|q(\alpha u)\right|$ since $|z_1 z_2| = |z_1||z_2|$ and $|u|=1$