A short course in static and discrete-time dynamic optimization

Mark Voorneveld

August 25, 2023

Contents

Preface

“Be troubled, ye careless ones.”
— Isaiah 32:11, King James Bible (Cambridge edition)

“You lecturers are sucking my blood. [...] You clever people should keep quiet about your knowledge and come out with it only in cases of urgency”
— Peter Handke, *Absence*, Farrar, Strauss and Giroux, 1990.

These are lecture notes for a course in advanced mathematics for economic analysis. This should point out three things. Firstly, it is a course in mathematics, not in economics. Secondly, the topics have been decided upon in close collaboration with other teachers responsible for the M.Sc. program in economics and provide — at least in part — the mathematical tools that are required for understanding and solving advanced models in these courses. Thirdly, it is an advanced course and I will assume that you master certain material before entering the course.

To many students, this course will come as a bit of a shock. The main reason is that this course tries to bridge the gap between the bachelor level courses that are mainly computation-driven and the abstract mathematics and proof-driven reasoning of the academic economist. For many students this will be a first encounter, and it takes some getting used to. Therefore, it is important that you come prepared. Several requirements were spelled out on the course web and in correspondence sent to you during the spring.

All of this — together with my hard-earned reputation of being autistically literal-minded and inflexible in exam correction — is enough to scare anyone. But there is some light at the end of the tunnel and it does not come from an on-storming freight train: the course starts of with a ‘peak’ of abstraction in the first couple of weeks, but becomes more and more applied towards the end of the course. Sounds like a bad choice of planning, doesn’t it? But it is the only natural order in which to do things: section 1 is needed for section 2, which is needed for section 3, etc.

And, ironically, once we understand the hard stuff, it makes the other stuff look easy.

If it is any consolation, let me quote one of my old math professors, Giel Paardekooper: “You are never given unsolvable problems; this is not the Faculty of Theology.”

One of the twentieth-century’s greatest philosophers, Ludwig Wittgenstein, once said: “I don’t know why we are here, but I’m pretty sure it is not in order to enjoy ourselves.” If you could bear that in mind during the Fall term, I’d certainly appreciate it.

Some motivation for the contents

One of the course’s main goals is to say something intelligent about optimization problems.

A crucial result is the Extreme Value Theorem (Thm. 13.3), which assures the existence of solutions to optimization problems under certain regularity assumptions, notably ***continuity*** and ***compactness***. It is commonly applied to establish that problems in economics, including least squares estimation or other minimal distance problems, utility/profit maximization, intergenerational welfare optimization, etc., actually have solutions.

In the sections on dynamic optimization, the feasible set will consist of ***sequences*** of choices over time. Moreover, the optimal value function — assigning to each initial state the optimal value of the problem over time, like optimal intergenerational welfare or so — and the corresponding optimal decisions are often derived using a ***recursive formulation***, where the value function becomes a ***fixed point*** of some equation.

The existence of such a fixed point is then established using the ***Banach fixed point theorem*** and it is approximated using ***sequences*** of these functions, getting closer and closer to one another ***(Cauchy sequence)*** and eventually, their ***limit***.

This explains the tools we need and in particular all the work we do in the relatively abstract parts of the course:

1. on (vector) spaces of sequences, functions, and other vectors;
2. on normed and metric spaces to measure distances;
3. on topology to be able to talk about open/closed/... sets, which in its turn lies at the foundation of definitions of continuity, compactness, etc.
4. on different types of convergence. Sometimes, extra assumptions (like uniform convergence) are required to make sure that the limit has nice properties;
5. on completeness and the Banach contraction theorem. Among other applications, these will establish the existence of the optimal value function in dynamic optimization.

Courseweb and lecture notes

On the Canvas website for this course under the 'Files' tab, you will find four directories with material that you might find useful:

1. Lecture notes and study guide:
   (a) This file of lecture notes. We will not go through the entire collection of lecture notes. Some sections will be treated very briefly, others not at all.
   (b) A study guide that I will update after each lecture. I will specify for each lecture which parts of the notes I covered (obligatory reading), what sections of the recommended books you might find useful additional reading (not obligatory reading), and often give a brief informal discussion. This is an important aid: we will not cover all of the notes. In particular, some sections and some technical and long proofs will be skipped. I put them in mostly to make the notes reasonably self-contained. So with the help of the study guide, you can save yourself a lot of time: *you only need to read the parts of the notes specified for each lecture in the study guide.* You can skip or skim everything else. I expect everybody to read the study guide.
   (c) If necessary: a file with errata.
2. Prerequisites: a file that was mailed to you already during the spring. It lists some of the material that I expect you to know before the course starts.
3. This year's problem sets: the teaching assistant, Yifan Yang, will post the problem sets and their solutions here.
4. Old problem sets and exams: these files contain elaborate solutions. Please note that old problem sets and exams refer to theorem numbers etc. in old versions of the lecture notes, which may not be identical to the latest installment, because I update the notes every year.

Note that I put solutions to many of the exercises in an appendix to the lecture notes! These notes are always a work in progress. Some sections that I would like to include in future editions are still missing, other sections are bare skeletons and I would like to add many more worked exercises and examples. Your comments would be very much appreciated: if you have suggestions or find typos, grammatical errors, stylistic *faux pas*, and (hopefully not) mathematical errors, I would be happy to hear it and take it into account in future editions of these notes.

Rules for the home assignments

You will be given five sets of home assignments during the course. Together, they count for 30% of your final grade. The rules are:

- ☒ Groups of up to four students may submit one set of solutions (the larger the group, the happier my teaching assistant will be). Write each group member's name and student number clearly on the first page of your solutions.
- ☒ Group compositions are allowed to change from one home assignment to another.
- ☒ To obtain a good score, write clearly and logically, starting from definitions and correctly deducing and motivating your answers.

I will not interfere in the way you compose groups. In earlier years, one or two students volunteered to coordinate on randomizing groups: participants who were interested in this signed up and they were randomly divided into groups of four for each problem set. This is a wonderful initiative!

Rules for the exam

- ☒ The exam is open-book. You may bring and use paper versions of the following, but nothing else:

  1. all PDF files distributed on the Canvas site of course 5301, fall term 2023,
  2. notes you wrote yourself,
  3. your group's solutions to the problem sets.

- ☒ You do not need a calculator. That being said, use of a calculator is allowed. The list of permissible models can be found on the portal. At the time I am writing this, it is under Student Support — Examinations — Calculators. If you don't have a permissible model, you may be able to borrow one from MSc students in the second year. This list was decided upon by the Program Committee and is not open for discussion with course directors or the Examinations Office.
- ☒ None of this material will be provided for you. Borrowing material from other students at the exam is not allowed.

Mark Voorneveld

1 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 1.1 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 1.2 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 1.1 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 1.1 Definition 1.1 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. ◁

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 1.2 on page 5, 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 1.3 For arbitrary $n \in \mathbb{N}$, the set $\mathbb{R}^n$ consists of n-tuples $x = (x_1, \dots, x_n)$ of n real numbers. Two vectors $x = (x_1, \dots, x_n)$ and $y = (y_1, \dots, y_n)$ in $\mathbb{R}^n$ are defined to be equal if $x_i = y_i$ for all coordinates $i = 1, \dots, n$. $\mathbb{R}^n$ is a vector space under the operations of coordinatewise addition and scalar multiplication. Formally, for all $x = (x_1, \dots, x_n), y = (y_1, \dots, y_n) \in \mathbb{R}^n$, and all scalars $\alpha \in \mathbb{R}$:

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

Its zero element $\mathbf{0}$ is the vector $(0, \dots, 0)$ 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 1.4 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 $i \in \{1, \dots, m\}$ and column $j \in \{1, \dots, n\}$ is denoted by A_{ij} or $a_{ij} \in \mathbb{R}$. Entries $a_{i1}, a_{i2}, \dots, a_{in}$ form the i-th *row* of matrix A and entries $a_{1j}, a_{2j}, \dots, a_{mj}$ 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 $i = 1, \dots, m$ and $j = 1, \dots, n$.

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 1.5 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 1.6 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 $a \leq b$. Addition and scalar multiplication are defined as in Example 1.2: 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 1.7 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 1.8 (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 1.1

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

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

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 1.2 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 1.2

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 1.7.

Example 1.9 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 1.10 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 1.2, 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 1.3** 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 1.11

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

◁

Exercises section 1

1.1 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.

1.2 (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 $\mathbb{R}^{\mathbb{N}}$?

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

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

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

(a) Prove, by verifying the properties in Definition 1.1, that the product space $\times_{i \in I} V_i$ (see Example 1.8) is a vector space.

(b) Prove that Examples 1.1, 1.2, 1.3, 1.4, and 1.5 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.

1.6 Prove, by verifying the properties in Definition 1.1, that the set $P(\mathbb{R})$ of real polynomials is a vector space.

1.7 (a) Prove Theorem 1.2.

(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 $\alpha, \beta \in \mathbb{R}$.

2 Linear combinations, basis and dimension

2.1 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 2.1** 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 $w_1, \dots, w_n$ in W and scalars $\alpha_1, \dots, \alpha_n$ 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 $\operatorname{span}(W)$. By convention, $\operatorname{span}(\emptyset) = \{\mathbf{0}\}$.
>
> ☒ If $W' = \operatorname{span}(W)$, 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.

By Theorem 1.2, the span of a subset W of a vector space V is a subspace of V.

Example 2.1 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 2.2 The set of polynomials over $\mathbb{R}$ is spanned by the 'monomials' $1, x, x^2, x^3, x^4, \dots$. 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 2.2 Let V be a vector space.

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

$$\alpha_1 v_1 + \cdots + \alpha_k v_k = \mathbf{0}. \tag{5}$$

They are ***linearly independent*** if, in contrast, the only scalars for which (5) is true are $\alpha_1 = \cdots = \alpha_k = 0$.

☒ A (possibly infinite) subset W of V is ***linearly dependent*** if it has a nonempty finite subset whose distinct elements are linearly dependent. Otherwise, W is ***linearly independent***.

So in the following examples we write down and solve equation (5) for the alphas. If all of them are zero, the vectors are linearly independent; otherwise they are linearly dependent.

Example 2.3 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 2.4 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 2.3.

2.2 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 2.3 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 2.5 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 2.6 (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:

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

The set $\{e_1,\dots,e_n\}$ 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 2.7 The set $\{1,x,x^2,x^3,\dots\}$ is a basis for the set of polynomials over $\mathbb{R}$, making it infinite-dimensional; similarly, the set $\{1,x,x^2,\dots,x^n\}$ is a basis for the set of polynomials with degree at most n, making that subspace $(n+1)$-dimensional. ◁

Example 2.8 Recalling that $\operatorname{span}(\emptyset) = \{\mathbf{0}\}$ 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 $\{e_1,\dots,e_n\}$ for $\mathbb{R}^n$ and the polynomials $\{1,x,x^2,x^3,\dots\}$ 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 2.1 (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 3 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 2.2

Let V be a vector space that has

☒ $m \in \mathbb{N}$ vectors $s_1, \dots, s_m$ that span V,

☒ $n \in \mathbb{N}$ vectors $v_1, \dots, v_n$ 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 $s_1, \dots, s_m$ span V, there are scalars $\alpha_1, \dots, \alpha_m$ such that

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

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

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

So $v_1, s_2, \dots, s_m$ 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 $\alpha_1, \dots, \alpha_m$ 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 $\alpha_2, \dots, \alpha_m$ 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 $v_1, v_2, s_3, \dots, s_m$ span the same set as $v_1, s_2, \dots, s_m$. So $v_1, v_2, s_3, \dots, s_m$ 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 $v_1, \dots, v_m$ spans V. This is impossible by linear independence of $v_1, \dots, v_m, v_{m+1}, \dots, v_n$, since v_n is not in the span of $v_1, \dots, v_m$. Hence, $n \leq m$.

For the final claim, V has a basis by Theorem 2.1. 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 2.9

☒ 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 2.3

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 2.1, 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 2.10 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 2.3 they are a basis of $\mathbb{R}^3$. ◁

Exercises section 2

2.1 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 $C[0,1]$.

2.2 Prove:

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

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

2.3 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 $w_1, \ldots, w_n$.

3 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 3.1 (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 3.1 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 3.2 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 $a \le b$) or $(-\infty, b] \subseteq (-\infty, a]$ (if $b \le a$). 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 3.1 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 3.1 (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 2.1 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 span(B): if $s \notin \text{span}(B)$, 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 $u_1, \dots, u_n$ be finitely many elements of U and suppose there are scalars $\alpha_1, \dots, \alpha_n$ such that $\alpha_1 u_1 + \dots + \alpha_n u_n = \mathbf{0}$. For each $i = 1, \dots, n$, $u_i \in U$ 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, $\{u_1, \dots, u_n\} \subseteq C_n$. The linear independence of C_n implies that $\alpha_1 = \dots = \alpha_n = 0$. Since each chain in $\mathscr{A}$ has an upper bound, $\mathscr{A}$ has a maximal element by Zorn's lemma!

Exercises section 3

3.1 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 $\mathbb{N} = \{1, 2, 3, \dots\}$;

(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.

4 Linear functions

Definition 4.1 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 4.1 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

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

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 4.2 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 4.3 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 4.1, 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. $\triangleleft$

I list a few properties related to linear functions that follow easily from the definitions (Exercise 4.4):

- ☒ 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.

Exercises section 4

4.1 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 $T(x) = Ax$.

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

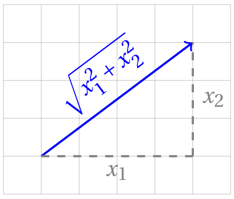
(e) $B_3 = \{(0,1,1,0,-1), (0,3,0,1,-2)\}$.

4.3 (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.

4.4 Prove the claims about linear functions after Example 4.3.

5 Normed vector spaces and inner product spaces

5.1 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 5.1 Define the length of vector $x = (x_1, \dots, x_n)$ 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.

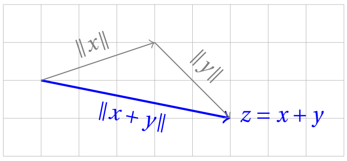
What is the length of 3 times the vector x (or -3 times, which has the same effect on the magnitude of the coordinates, but changes their sign)? Arguably, this is just three times the length of x. So:

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 5.1. ◁

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

Definition 5.1 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 5.1 becomes:

Example 5.2 $(\mathbb{R}^n, \|\cdot\|_2)$ 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 5.3 $(\mathbb{R}^n, \|\cdot\|_1)$ 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 5.4 $(\mathbb{R}^n, \|\cdot\|_\infty)$ 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 5.5 $(\mathbb{R}^n, \|\cdot\|_p)$ 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 5.3 and 5.2. The notation in Example 5.4 is explained by the fact that for each $x \in \mathbb{R}^n$:

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

Exercise 5.6 establishes the triangle inequality. ◁

Example 5.6 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 5.7 $(C[a,b], \|\cdot\|_\infty)$ 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 5.8 $(C[a,b], \|\cdot\|_1)$ is a normed vector space, where

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

◁

5.2 Inner product spaces

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

> **Definition 5.2** The ***inner product*** of two vectors $x, y \in \mathbb{R}^n$ is
>
> $$\langle x, y\rangle = x_1 y_1 + x_2 y_2 + \cdots + x_n y_n = \sum_{i=1}^n x_i y_i.$$

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 5.9 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 $p_i x_i$, this will cost you $\langle p, x\rangle = p_1 x_1 + \cdots + p_n x_n$. ◁

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

$$\langle x, x\rangle = x_1 x_1 + \cdots + x_n x_n = \sum_{i=1}^n x_i^2.$$

Comparing this with the definition of the Euclidean norm of a vector x in Example 5.2, 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 5.1.

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 ($x = \mathbf{0}$).
- ☒ $\langle x, y\rangle = \sum_{i=1}^n x_i y_i = \sum_{i=1}^n y_i x_i = \langle y, x\rangle$.
- ☒ $\langle x + y, z\rangle = \sum_{i=1}^n (x_i + y_i) z_i = \sum_{i=1}^n (x_i z_i + y_i z_i) = \sum_{i=1}^n x_i z_i + \sum_{i=1}^n y_i z_i = \langle x, z\rangle + \langle y, z\rangle$.
- ☒ $\langle \alpha x, y\rangle = \sum_{i=1}^n (\alpha x_i) y_i = \alpha \sum_{i=1}^n x_i y_i = \alpha \langle x, y\rangle$.

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 5.3 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.[1]

Example 5.10 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 5.1.

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***:

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

[1] 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 5.1 (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: $|\langle x, y\rangle| \le \|x\|\|y\|$.

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

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

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

$$\begin{aligned} 0 \le \|x-\alpha y\|^2 = \langle x-\alpha y, x-\alpha y\rangle &= \langle x-\alpha y, x\rangle - \alpha\underbrace{\langle x-\alpha y, y\rangle}_{=0} \\ = \langle x, x\rangle - \alpha\langle y, x\rangle = \langle x, x\rangle - \frac{\langle x, y\rangle^2}{\langle y, y\rangle} &= \|x\|^2 - \frac{\langle x, y\rangle^2}{\|y\|^2}. \end{aligned}$$

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

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

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

The triangle inequality follows by taking the square root. □

Exercises section 5

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

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

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

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

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

$$|\|x\| - \|y\|| \le \|x - y\|.$$

5.3 Verify that all our examples are normed vector spaces; for Example 5.5, consult Exercise 5.6.

5.4 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 $\|x-y\|$ and add or subtract the resulting equalities.

5.5 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.

5.6 (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 $x_1, \dots, x_n$ and $y_1, \dots, y_n$:

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

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***:

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

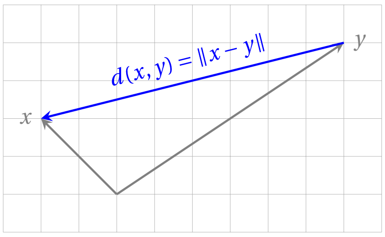
6 Metric spaces

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 6.1 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 6.1 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 6.2 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 6.3 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 6.2** 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 6.1, 6.3, and 6.4. 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 6.3** 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 6

6.1 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.

6.2 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.

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

6.4 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 $\{i \in \{1,\dots,n\} : x_i \neq y_i\}$. 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 6.2?

(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 $(\mathbb{R}^2, d_H)$.

6.5 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.

6.6 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)?

6.7 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.

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

6.9 We know how to turn a normed vector space into a metric space. This exercise is about the opposite direction. Recall from Example 6.1 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 $x,y,z \in V$: $d(x+z, y+z) = d(x,y)$.

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

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 $x,y \in V$.

6.10 (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:

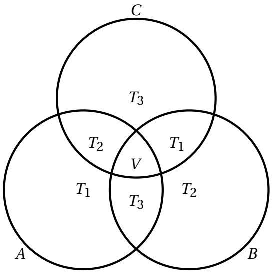
Customer 1:	{`Infinite jest`, `Tourmaline`, `Freshwater`, `Lila`},
Customer 2:	{`The blazing world`, `Ways of going home`, `Ghost wall`, `Elmet`},
Customer 3:	{`Lila`, `Infinite jest`, `The blazing world`, `Elmet`, `Freshwater`}.

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

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

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:

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

7 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 7.1 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 7.1 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 7.2 If (X,d) is a metric space, the open ball $B(x,r)$ from Definition 6.2 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

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

Example 7.3 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 7.1

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

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

(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 6.2, *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 7.4 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 7.2 A ***topology*** on a set X is a collection $\mathscr{O}$ of subsets of X that are called ***open*** sets and that satisfy:

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

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

By Theorem 7.1, metric spaces are topological spaces. But there are many others. Given an arbitrary set X, the ***trivial topology*** $\mathscr{O} = \{\emptyset, X\}$ calls only the empty set and X itself open; in the ***discrete topology*** $\mathscr{O} = P(X)$, the power set X, *all* subsets of X are called open.

Theorem 7.2 (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 7.1 and using De Morgan's Laws (Appendix A.2), we find:

Theorem 7.3

Let (X, d) be a metric space. Its collection $\mathscr{C}$ 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 7.1 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 8.2.

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

☒ The ***interior*** of U, denoted $\mathrm{int}(U)$, is the largest open set contained in U. It is the union of all open sets contained in U:

$$\mathrm{int}(U) = \cup_{W \subseteq U:\, W \text{ is open}} W. \tag{17}$$

☒ The ***closure*** of U, denoted $\mathrm{cl}(U)$, is the smallest closed set containing U. It is the intersection of all closed sets containing U:

$$\mathrm{cl}(U) = \cap_{W \supseteq U:\, W \text{ is closed}} W. \tag{18}$$

☒ The ***boundary*** of U, denoted $\mathrm{bd}(U)$, is the set

$$\mathrm{bd}(U) = \mathrm{cl}(U) \cap \mathrm{cl}(U^c). \tag{19}$$

As the intersection of two closed sets, it is a closed set. Elements of $\mathrm{bd}(U)$ 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 $\mathrm{acc}(U)$.

☒ 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 7.5 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 $\mathrm{cl}(U)$, the set of points around which each open ball contains at least one element of U, and the boundary $\mathrm{bd}(U)$ 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 7.4

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:

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

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

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

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

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

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 7.2) 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 7.4 implies a number of useful characterizations of closed sets; its proof is Exercise 7.10.

Theorem 7.5 (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: $\mathrm{acc}(U) \subseteq U$.

Since these properties are equivalent, any of them can be used as an alternative definition of a closed set. For instance, several textbooks on mathematics for economists define a set to be closed if it contains all its boundary points.

Example 7.6 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
$(0,1]$	$(0,1)$	$[0,1]$	$\{0,1\}$	$[0,1]$	$\emptyset$
$\mathbb{Q}$	$\emptyset$	$\mathbb{R}$	$\mathbb{R}$	$\mathbb{R}$	$\emptyset$
$\mathbb{N}$	$\emptyset$	$\mathbb{N}$	$\mathbb{N}$	$\emptyset$	$\mathbb{N}$
$\{0\} \cup (1,2]$	$(1,2)$	$\{0\} \cup [1,2]$	$\{0,1,2\}$	$[1,2]$	$\{0\}$

Exercises section 7

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

7.2 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 $x \in [0,1]$.

7.3 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.

7.4 Show that finite subsets of metric spaces are closed.

7.5 Prove Theorem 7.3.

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

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

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

7.9 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': $\mathrm{int}(U)^c = \mathrm{cl}(U^c)$.

7.10 Prove Theorem 7.5.

7.11 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?

8 Continuous functions

8.1 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 8.1 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 8.1 (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 8.1 Consider $f : \mathbb{R} \to \mathbb{R}$ with $f(x) = x^2$. Then

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

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 8.2 (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 8.2 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 8.2 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

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

The composition of continuous functions is continuous:

Theorem 8.3 (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. □

8.2 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 8.3 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 8.4 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 8.5 The function $f:\mathbb{R}\setminus\{0\} \to \mathbb{R}$ with $f(x) = 1/x$ is continuous: let $a \neq 0$ and $\varepsilon > 0$. Choose $\delta = \min\{\frac{\varepsilon|a|^2}{2}, \frac{|a|}{2}\}$. For each $x \neq 0$ 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

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

◁

Example 8.6 (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 $i = 1,\dots,n$:

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

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 8.3, 8.4, and 8.5) and recalling that the composition of continuous functions is continuous (Theorem 8.3), 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 8.7 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 8.4 (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 $x \in X$).

8.3 Uniform and Lipschitz continuity

There are two common, more restrictive notions of continuity:

Definition 8.3 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 8.8 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 8.9 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 8

8.1 Make the proof sketch in Example 8.6 precise.

8.2 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.

8.3 Prove Theorem 8.4.

8.4 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 8.8? Sure, but the function has a different domain now!)

9 The limit of a sequence and the limit of a function

9.1 The limit of a sequence

Next, we consider the limit of sequences. We defined sequences of real numbers in Example 1.5. Let us extend this to sequences in an arbitrary metric space.

Definition 9.1 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 9.1 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 9.2 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

for each $\varepsilon > 0$ there is a number N such that for each integer $k \geq N$ we have $d(x_k, x) < \varepsilon$.

Such a sequence is called ***convergent***.

Theorem 9.1 (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

for each integer $k \geq N$: $d(x_k, x) < \varepsilon$ and for each integer $k \geq N'$: $d(x_k, x') < \varepsilon$.

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:

$$d(x,x') = 2\varepsilon > d(x,x_k) + d(x_k,x') \geq d(x,x'). \qquad \square$$

Some relatively routine examples of convergent sequences of real numbers can be found in Exercise 9.1; its solution in the appendix is very elaborate and explains possible strategies for solving this type of problems. Here are a few more:

Example 9.2 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 9.3 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 9.4 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 9.3 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 9.2 (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 $i \in \{1,\dots,n\}$ 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 9.4 (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 9.3 (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 9.4

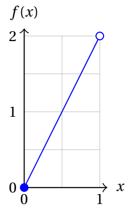
A subset U of a metric space X is closed if and only if for each convergent sequence in U, the limit belongs to U.

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! □

9.2 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 9.5 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

$$\text{each } x \in U \text{ with } 0 < d(x,a) < \delta \text{ has } d'(f(x),L) < \varepsilon. \tag{26}$$

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 9.1, if a limit exists, it must be unique by the Hausdorff property of metric space Y.

If you compare Definition 9.5 with the (ε,δ)-definition of continuity at point a (Thm. 8.1), 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 9.5

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 9.5 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 (26) implies that (22) holds as long as $x \neq a$. It holds trivially for $x = a$. □

Exercises section 9

9.1 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) $\left(\frac{(-1)^k}{3k-1}\right)_{k\in\mathbb{N}}$

9.2 (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 $i \in \{1,\dots,n\}$ the sequence $(x_{ki})_{k\in\mathbb{N}}$ of its i-th coordinates converges to $x_i \in \mathbb{R}$.

9.3 Consider a set X with the discrete metric. Which sequences are convergent?

9.4 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) $f(x_1,x_2) = \frac{x_1x_2}{x_1^2+x_2^2}$

(b) $f(x_1,x_2) = \frac{x_1^2+x_2^2}{|x_1+x_2|+|x_1x_2|}$

(c) $f(x_1,x_2) = \frac{\sin(x_1x_2)}{\sqrt{x_1^2+x_2^2}}$ HINT: $|\sin(y)| \le |y|$ 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) $f(x_1,x_2) = \frac{x_1^2+3x_1^2x_2+x_2^2}{x_1^2+x_2^2}$

(f) $f(x_1,x_2) = \frac{4x_1x_2}{\sqrt{x_1^2+x_2^2}}$

(g) $f(x_1,x_2) = \frac{x_1^2}{x_1^2+x_2^2}$

10 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

$$\text{for all } m,n \geq N: \qquad d(x_m,x_n) \leq d(x_m,x) + d(x,x_n) < \frac{1}{2}\varepsilon + \frac{1}{2}\varepsilon = \varepsilon. \tag{27}$$

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 10.1** 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 (27), every convergent sequence in a metric space is a Cauchy sequence. The converse is false:

Example 10.1 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 10.2 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 5.8. 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 10.2** 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 10.1 (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 $\max\{\|x_1\|_2, \dots, \|x_{N-1}\|_2, \|x_N\|_2 + 1\}$.
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 9.2.
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 10.2

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 9.4). This proves that each Cauchy sequence in U has a limit in U: U is complete.

Conversely, assume U is complete. We use Theorem 9.4 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 10.3 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

$$d_\infty(f,g) = \sup\{d(f(x), g(x)) : x \in X\}. \tag{28}$$

☒ 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 (28) 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 13.6 and 13.1(e);
4. Earlier, we wrote $B(\mathbb{N},\mathbb{R}) = B(\mathbb{N})$ and $C([a,b],\mathbb{R}) = C[a,b]$.

Theorem 10.3

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

$$f(x) = \lim_{k\to\infty} f_k(x) \in Y \tag{29}$$

exists. For the function $f : X \to Y$ defined in (29) 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

$$\text{for all } k, m \geq N: \qquad d_\infty(f_k, f_m) < 1. \tag{30}$$

By assumption, f_N is bounded:

$$\text{for some } y \in Y \text{ and some } \varepsilon > 0: \qquad f_N(X) \subseteq B(y, \varepsilon). \tag{31}$$

We will show that

$$f(X) \subseteq B(y, \varepsilon + 2). \tag{32}$$

Let $z \in f(X)$: there is an $x \in X$ with $f(x) = z$. By (29) for $\varepsilon = 1$, there is an $M \in \mathbb{N}$ such that for all $k \geq M$:

$$d(f(x), f_k(x)) < 1. \tag{33}$$

Let $k \geq \max\{N, M\}$. By the triangle inequality and (33), (30), (31):

$$d(f(x), y) \leq d(f(x), f_k(x)) + d(f_k(x), f_N(x)) + d(f_N(x), y) < 1 + 1 + \varepsilon,$$

proving (32).

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 $k, m \geq N$:

$$d(f_k(x), f(x)) \leq d(f_k(x), f_m(x)) + d(f_m(x), f(x)) < \frac{1}{3}\varepsilon + d(f_m(x), f(x)). \tag{34}$$

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 10.3 ($C[a,b]$ 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 10.2 is *not* complete, but we just argued that it *is* complete under the supremum metric. ◁

11 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 11.1 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 11.1 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 $x, y \in X$:

$$d(T(x), T(y)) \leq \alpha d(x,y). \tag{35}$$

Each contraction is Lipschitz continuous and consequently (uniformly) continuous.

Example 11.2 $f : [0,\infty) \to [0,\infty)$ 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 11.3 $f : [0,\varepsilon] \to [0,\varepsilon]$ 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 11.4 (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 $x \in \text{int}(U)$, then f is a contraction: for any two distinct $x, y \in U$, the Mean Value Theorem (see Theorem A.1 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 11.3). ◃

Definition 11.2 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 11.1 (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:

$$\text{for each } k \in \mathbb{N}: \qquad x_k = T(x_{k-1}) = T^k(x_0). \tag{36}$$

(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 $k \in \mathbb{N}$:

$$d(T^k(x_0), x) \le \frac{\alpha}{1-\alpha} d(T^{k-1}(x_0), T^k(x_0)) \quad \text{and} \quad d(T^k(x_0), x) \le \frac{\alpha^k}{1-\alpha} d(T(x_0), x_0). \tag{37}$$

Proof: (a) and (b): We show that the sequence of successive function values in (36) is a Cauchy sequence. For each $k \in \mathbb{N}$, definition (36) 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

$$d(x_{k+1}, x_k) \le \alpha^k d(x_1, x_0). \tag{38}$$

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 (38). □

The inequalities in (37) 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 11.5 (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 (37):

$$d(x, T^n(0)) \leq \frac{(1/2)^n}{1 - (\frac{1}{2})} \cdot \frac{1}{2} = \frac{1}{2^n}.$$

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 (35), to assume that

$$\text{for all } x, y \in X, \text{ if } x \neq y, \text{ then:} \qquad d(T(x), T(y)) < d(x, y), \tag{39}$$

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 11.2, however, we will see that nonexpansiveness does suffice if the metric space (X, d) is assumed to be compact.

Example 11.6 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 11.2

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 11.1, 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 11.3

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 $x = T(x) \in T(X_1) \subseteq X_2$. □

11.1 Application: ranking the relevance of webpages

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:

$$\begin{aligned} d_1(f(x), f(y)) &= \|Mx - My\|_1 = \|\alpha Bx - \alpha By\|_1 = \alpha \|B(x-y)\|_1 \\ &= \alpha \left\| \textstyle\sum_{i=1}^n (x_i - y_i) b_i \right\|_1 \leq \alpha \sum_{i=1}^n |x_i - y_i| \|b_i\|_1 \\ &= \alpha \|x - y\|_1 = \alpha d_1(x, y). \end{aligned}$$

Finally, to see that Δ_n is complete, note that $\mathbb{R}^n$ with the taxicab metric d_1 is complete.[2] And Δ_n is a closed subset, hence complete as well. So we can indeed apply the contraction theorem.

Exercises section 11

11.1 (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 $f \leq g$, then $T(f) \leq T(g)$. [Here, $f \leq g$ means $f(x) \leq g(x)$ 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 $c \in B(U, \mathbb{R})$:

$$T(f + c) \leq T(f) + \beta c;$$

then T is a contraction with modulus β in $(B(U, \mathbb{R}), d_\infty)$.

11.2 (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 $x_0 \in X$.

[2] 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: $\|x\|_2 \leq \|x\|_1 \leq \sqrt{n}\|x\|_2$ by Exc. 5.1.

(a) Show that T has at most one fixed point.

(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.

11.3 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?

12 Uniform convergence

12.1 Uniform convergence and the Cauchy criterion

In Examples 9.3 and 9.4, we studied convergence of sequences of functions $(f_k)_{k\in\mathbb{N}}$ in the metric space $(C[0,1], d_\infty)$. Such a sequence converges to function f if for each $\varepsilon > 0$ there is an $N \in \mathbb{N}$ such that

$$\text{for all } k \geq N: \qquad d_\infty(f_k, f) = \sup_{x\in[0,1]} |f_k(x) - f(x)| < \varepsilon.$$

Hence, $|f_k(x) - f(x)| < \varepsilon$ for all $x \in [0,1]$. The chosen N is consequently independent of x. One refers to this as “uniform convergence”. There is also a weaker type of convergence — pointwise convergence — where the choice of N is allowed to depend on x as well. Let us define both formally.

Definition 12.1 Let D be a subset of $\mathbb{R}$ and let $(f_k)_{k\in\mathbb{N}}$ be a sequence of functions $f_k : D \to \mathbb{R}$.

☒ The sequence $(f_k)_{k\in\mathbb{N}}$ ***converges pointwise*** to limit $f : D \to \mathbb{R}$ if for each $x \in D$: $\lim_{k\to\infty} f_k(x) = f(x)$. Stated differently, if for each $x \in D$ and each $\varepsilon > 0$,

$$\text{there is an } N \in \mathbb{N} \text{ such that if } k \geq N, \text{ then } |f_k(x) - f(x)| < \varepsilon.$$

☒ The sequence $(f_k)_{k\in\mathbb{N}}$ ***converges uniformly*** to limit $f : D \to \mathbb{R}$ if for each $\varepsilon > 0$

$$\text{there is an } N \in \mathbb{N} \text{ such that if } k \geq N, \text{ then for each } x \in D: \quad |f_k(x) - f(x)| < \varepsilon. \tag{40}$$

So in pointwise convergence, N is allowed to depend on both ε and x, whereas in uniform convergence it is allowed to depend on ε only: uniform convergence implies pointwise convergence, not the other way around.

This is tricky, so let's first go through the definitions once more, a bit more informally, and then do a bunch of examples. Let D be a subset of $\mathbb{R}$, let $(f_k)_{k\in\mathbb{N}}$ be a sequence of functions $f_k : D \to \mathbb{R}$, and $f : D \to \mathbb{R}$ a candidate limit.

1. What does it mean that sequence $(f_k)_{k\in\mathbb{N}}$ converges to f *pointwise*?

   For each x in the domain D of the function, the sequence of numbers

   $$f_1(x), f_2(x), f_3(x), \ldots$$

   converges to $f(x)$.

2. What does it mean that sequence $(f_k)_{k\in\mathbb{N}}$ converges to f *uniformly*?

   For each $\varepsilon > 0$ you must be able to do the following:

   ☒ Draw the graph of the function that lies ε above f and the graph of the function that lies ε below f.

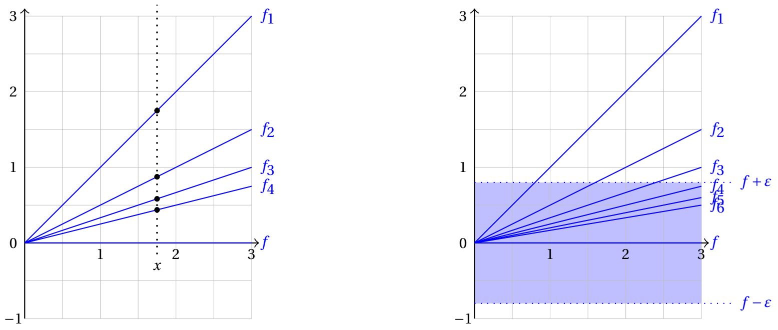
   ☒ Shade the area between them in some pretty color.

   ☒ You should be able to find some integer N such that the graphs of the functions

   $$f_N, f_{N+1}, f_{N+2}, \ldots$$

   all lie in this shaded area.

Consider, as a simple example, the sequence $(f_k)_{k\in\mathbb{N}}$ of functions $f_k : [0,3] \to \mathbb{R}$ with $f_k(x) = \frac{x}{k}$ and the function $f : [0,3] \to \mathbb{R}$ with $f(x) = 0$. The left figure illustrates why this sequence converges to f pointwise; the right figure illustrates why it converges to f uniformly.



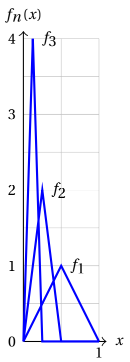
This is meant to convey the intuition behind pointwise and uniform convergence only. Here are some more detailed examples:

Example 12.1 Let $D = [0,1]$ and $f_k(x) = x^k$. Take $f(x) = 0$ if $0 \le x < 1$ and $f(1) = 1$. The sequence $(f_k)_{k\in\mathbb{N}}$ converges to f pointwise, but not uniformly, since $\lim_{x\to 1} f_k(x) = 1$. So $f_k(x)$ is not close to $f(x) = 0$ if we choose $x \in (0,1)$ sufficiently large. ◁

Example 12.2 Let $D = [-1,1]$ and $f_k(x) = x^k$. The sequence $(f_k)_{k\in\mathbb{N}}$ does not converge pointwise: $f_k(-1) = (-1)^k$ is 1 if k is even and -1 if k is odd, so the sequence of function values $(f_k(-1))_{k\in\mathbb{N}} = (-1,1,-1,1,\ldots)$ does not converge. ◁

Example 12.3 For each $n \in \mathbb{N}$, define the following 'tent function' $f_n : [0,1] \to \mathbb{R}$:

$$f_n(x) = \begin{cases} 2^{2n-1}x & \text{if } 0 \le x < \frac{1}{2^n}; \\ -2^{2n-1}\left(x - \frac{2}{2^n}\right) & \text{if } \frac{1}{2^n} \le x < \frac{2}{2^n}; \\ 0 & \text{if } \frac{2}{2^n} \le x \le 1. \end{cases}$$



So f_n starts in $f_n(0) = 0$ and its function value increases linearly to 2^{n-1} at $x = \frac{1}{2^n}$, after which it decreases linearly to zero at $x = \frac{2}{2^n}$ and remains zero thereafter. The figure to the right contains the graphs of f_1, f_2, and f_3. Notice that the height of the triangles doubles with n, whereas their bases decrease by a factor $1/2$. Since the base of the triangle shrinks to negligible size it is easy to show that the sequence $(f_n)_{n\in\mathbb{N}}$ converges pointwise to the zero function. The convergence is not uniform: since $f_n\left(\frac{1}{2^n}\right) = 2^{n-1} \ge 1$, it follows that for $\varepsilon \in (0,1)$, there is no $m \in \mathbb{N}$ such that $f_n(x) = |f_n(x) - 0| < \varepsilon$ for all $x \in [0,1]$. ◁

In Example 12.1, we saw that the pointwise limit of a sequence of continuous functions need not be continuous. Under uniform convergence, however, the limit function is continuous as well:

Theorem 12.1

Let $D \subseteq \mathbb{R}$ and let $(f_k)_{k\in\mathbb{N}}$ be a sequence of continuous functions $f_k : D \to \mathbb{R}$. If $(f_k)_{k\in\mathbb{N}}$ converges uniformly to $f : D \to \mathbb{R}$, then f is continuous.

Proof: Let $d \in D$ and $\varepsilon > 0$. We need to show that there is a $\delta > 0$ such that for all $x \in D$ with $|x - d| < \delta$ we have that $|f(x) - f(d)| < \varepsilon$.

By uniform convergence of $(f_k)_{k\in\mathbb{N}}$, there is an $N \in \mathbb{N}$ such that

$$\text{if } k \geq N \text{ and } x \in D, \text{ then } |f_k(x) - f(x)| < \frac{\varepsilon}{3}. \tag{41}$$

This holds in particular for the function f_N. By continuity of f_N, there is a $\delta > 0$ such that

$$\text{if } x \in D \text{ has } |x - d| < \delta, \text{ then } |f_N(x) - f_N(d)| < \frac{\varepsilon}{3}. \tag{42}$$

It follows from (41), (42), and the triangle inequality that if $x \in D$ has $|x - d| < \delta$, then

$$\begin{aligned} |f(x) - f(d)| &= |f(x) \underbrace{- f_N(x) + f_N(x)}_{=0} \underbrace{- f_N(d) + f_N(d)}_{=0} - f(d)| \\ &\leq |f(x) - f_N(x)| + |f_N(x) - f_N(d)| + |f_N(d) - f(d)| \\ &< \frac{\varepsilon}{3} + \frac{\varepsilon}{3} + \frac{\varepsilon}{3} = \varepsilon. \end{aligned}$$

□

Hence, if a sequence of continuous functions has a *discontinuous* pointwise limit, then the convergence cannot be uniform. Notice, however, from Example 12.3 that even if a sequence of continuous functions converges pointwise to a continuous limit, the convergence still need not be uniform. One way of recognizing uniformly convergent sequences is:

Theorem 12.2 (Cauchy criterion for uniform convergence)

Let $D \subseteq \mathbb{R}$ and let $(f_k)_{k\in\mathbb{N}}$ be a sequence of functions $f_k : D \to \mathbb{R}$. The following two claims are equivalent:

(a) The sequence $(f_k)_{k\in\mathbb{N}}$ converges uniformly;

(b) For each $\varepsilon > 0$ there is an $N \in \mathbb{N}$ such that for all $k, \ell \geq N$ and all $x \in D$:

$$|f_k(x) - f_\ell(x)| < \varepsilon. \tag{43}$$

Its proof is a tricky exercise (Exercise 12.2) using Cauchy sequences.

12.2 Uniform convergence and integration

Theorem 12.3

Let $(f_k)_{k\in\mathbb{N}}$ be a sequence of continuous functions $f_k : [a,b] \to \mathbb{R}$ that converges uniformly to $f : [a,b] \to \mathbb{R}$. Then 'the limit of the integrals is the integral of the limit':

$$\lim_{k\to\infty} \int_a^b f_k(x)\,\mathrm{d}x = \int_a^b \lim_{k\to\infty} f_k(x)\,\mathrm{d}x.$$

Proof: By Theorem 12.1, the uniform limit f is continuous. Let $\varepsilon > 0$. By uniform convergence, there is an $N \in \mathbb{N}$ such that for all $k \geq N$ and all $x \in [a,b]$:

$$|f_k(x) - f(x)| < \frac{\varepsilon}{b-a}.$$

Hence, for all $k \geq m$:

$$\left|\int_a^b \big(f_k(x) - f(x)\big)\,\mathrm{d}x\right| \leq \int_a^b \big|f_k(x) - f(x)\big|\,\mathrm{d}x \leq \int_a^b \frac{\varepsilon}{b-a}\,\mathrm{d}x = \varepsilon. \qquad \square$$

The theorem does not necessarily hold if convergence is only pointwise:

Example 12.4 Recall the 'tent' functions from Example 12.3. For each n, the integral $\int_0^1 f_n(x)\,\mathrm{d}x$ is simply the area of the triangle with base $\frac{2}{2^n}$ height 2^{n-1} and , i.e.:

$$\int_0^1 f_n(x)\,\mathrm{d}x = \frac{1}{2}\cdot\frac{2}{2^n}\cdot 2^{n-1} = \frac{1}{2}.$$

The sequence $(f_n)_{n\in\mathbb{N}}$ converges pointwise to the zero function. But $\int_0^1 0\,\mathrm{d}x = 0 \neq \frac{1}{2}$: the limit of the integrals is different from the integral of the limit! ◁

12.3 Uniform convergence and differentiation

Is it allowed to exchange the order of limits and differentiation? In other words, is the derivative of the limit the limit of the derivatives? Not even uniform convergence suffices for this:

Example 12.5 The sequence of functions $(f_k)_{k\in\mathbb{N}}$ on $[0,1]$ with $f_k(x) = x^k/k$ converges uniformly to the zero function $\mathbf{0}$: for each $\varepsilon > 0$, take $N \in \mathbb{N}$ with $N > \frac{1}{\varepsilon}$. It follows that for each $k \geq N$ and each $x \in [0,1]$:

$$\big|f_k(x) - 0\big| = \left|\frac{x^k}{k} - 0\right| = \frac{x^k}{k} \leq \frac{1}{k} \leq \frac{1}{N} < \varepsilon.$$

The derivative of f_k is $f_k'(x) = x^{k-1}$, so $f_k'(1) = 1$. It follows that the sequence of derivatives does not converge (not even pointwise) to the derivative of its limit. ◁

Only under heavy additional assumptions, the order of limits and differentiation can be reversed:

Theorem 12.4

Consider a sequence $(f_k)_{k\in\mathbb{N}}$ of functions $f_k : [a,b] \to \mathbb{R}$ with the following three properties:

- ☒ the functions f_k are differentiable on (a,b);
- ☒ there is a point $c \in (a,b)$ such that the sequence of numbers $(f_k(c))_{k\in\mathbb{N}}$ converges;
- ☒ the sequence of derivatives $(f_k')_{k\in\mathbb{N}}$ converges uniformly to a limit g.

Then also the sequence $(f_k)_{k\in\mathbb{N}}$ converges uniformly, namely to a function f with $f' = g$.

We will skip the proof; do verify which of the three properties is violated in Example 12.5.

Exercises section 12

12.1 Consider the following sequences $(f_k)_{k\in\mathbb{N}}$ of functions $f_k : D \to \mathbb{R}$. Do they converge pointwise? Do they converge uniformly?

(a) $f_k(x) = \frac{\sin kx}{\sqrt{k}}$ on $D = \mathbb{R}$

(b) $f_k(x) = e^{-kx^2}$ on $D = \mathbb{R}_+$

(c) $f_k(x) = \frac{kx}{1+k\sqrt{k}}$ on $D = \mathbb{R}_+$

(d) $f_k(x) = x\cos\frac{x}{x+k}$ on $D = [0,1]$

(e) $f_k(x) = \frac{kx^2+1}{kx+1}$ on $D = [0,1]$

(f) $f_k(x) = \frac{kx^2+1}{kx+1}$ on $D = [1,2]$

12.2 Prove Theorem 12.2.

13 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 13.2 and will derive a number of pleasant properties that follow from compactness in this section — whose highlight for optimization purposes is Theorem 13.3 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 13.1 A subset of $\mathbb{R}^n$ with its usual distance is ***compact*** if it is closed and bounded.

This concern is well-motivated: Definition 13.1 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 13.2 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 13.1 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 $x_1, \ldots, x_n$ 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 $\min\{\frac{1}{2}x_1, \ldots, \frac{1}{2}x_n\} > 0$. ◁

Example 13.2 Let X be an infinite set and d the discrete metric (see Example 6.2). 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 13.3 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 13.4 In a metric space (X,d), each finite subset is compact.

Let $F = \{x_1, \dots, x_n\}$ 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 $O(x_1), \dots, O(x_n)$ are a finite subcovering of F. ⊲

Example 13.5 In a metric space (X,d), if sequence $(x_k)_{k\in\mathbb{N}}$ has limit x, then the set $C = \{x, x_1, x_2, x_3, \dots\}$ 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 $x_N, x_{N+1}, x_{N+2}, \dots$ from the N-th term onward. For each of the remaining terms $x_k \in \{x_1, \dots, x_{N-1}\} \subseteq C$ there is a set $O(x_k)$ in the covering with $x_k \in O(x_k)$. Conclude: $O(x), O(x_1), \dots, O(x_{N-1})$ is a finite subcovering of C. ⊲

Example 13.6 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 13.3 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. 13.2). Every totally bounded set is bounded (Exc. 13.3). But a bounded set need not be totally bounded: see Examples 13.2 and 13.3.

Theorem 13.1

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 13.1(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 13.4). But the union of all these compact sets is $\mathbb{R}$, which is not compact, since it is not bounded (Theorem 13.1(e)).

Theorem 13.2 (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 13.3 (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 $c \in C$: $f(m) \le f(c) \le f(M)$.

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 $C' = \{y_1,\dots,y_k\} \subseteq C$ such that $\{L(y) : y \in C'\}$ covers C. Since C' is finite, it contains a y^* with highest function value: $f(y^*) = \max\{f(y_1),\dots,f(y_k)\}$. 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 8.2, every continuous function is upper semicontinuous. But upper semicontinuity is a much weaker requirement than continuity.

Example 13.7 Function $f : \mathbb{R} \to \mathbb{R}$ with $f(x) = 1$ if $x \ge 0$ and $f(x) = 0$ otherwise is upper semicontinuous:

$$f^{-1}((-\infty,a)) = \begin{cases} \mathbb{R} & \text{if } a > 1, \\ (-\infty, 0) & \text{if } 0 < a \le 1, \\ \emptyset & \text{if } a \le 0. \end{cases}$$

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 13.4 (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 $\{B(x_1,\varepsilon(x_1)),\dots,B(x_k,\varepsilon(x_k))\}$. I claim that $\varepsilon = \min\{\varepsilon(x_1),\dots,\varepsilon(x_k)\}$ 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:

$$B(x,\varepsilon) \subseteq B(x,\varepsilon(x_i)) \subseteq B(x_i,2\varepsilon(x_i)) \subseteq O(x_i).$$

STEP 3: U is totally bounded, so for ε from step 2 there are finitely many $x_1,\dots,x_n$ 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 13.1, 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 13.1 and 13.2) of compactness of sets in $\mathbb{R}^n$.

Theorem 13.5 (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 13.1. 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 9.2. C is closed, so its limit lies in C by Theorem 9.4. By Theorem 13.4, C is compact. □

Theorem 13.6

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

$$d(y, x_i) < \frac{1}{2}\delta(x_i) < \delta(x_i),$$
$$d(x_i, z) \le d(x_i, y) + d(y, z) < \frac{1}{2}\delta(x_i) + \delta \le \delta(x_i).$$

Hence, using the triangle inequality and (47):

$$d'(f(y), f(z)) \le d'(f(y), f(x_i)) + d'(f(x_i), f(z)) \le \frac{1}{2}\varepsilon + \frac{1}{2}\varepsilon = \varepsilon.$$

□

Finally, let us state — without proof — the following useful theorem:

Theorem 13.7 (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 13.8 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 13

13.1 Use Definition 13.1 to check whether the following sets are compact:

(a) $\{x \in \mathbb{R} : 0 \leq x < 5\}$,

(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) $\{x \in \mathbb{R}^2 : 3x_1 - 2x_2 \leq 6, x_1 + x_2 \leq 3, x_1 \geq 0, x_2 \geq 0\}$,

(f) $\{x \in \mathbb{R}^2 : 0 \leq x_1 \leq 3, 1 \leq x_2 < 4\}$,

(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) $\{x \in \mathbb{R}^2 : 0 \leq x_2 \leq x_1^3, x_2 \geq 2x_1^3 - 6x_1^2 + 12x_1 - 8\}$.

13.2 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 $U \subseteq \cup_{u \in U'} B(u,\varepsilon)$.

13.3 Show that every totally bounded set in a metric space is bounded.

13.4 Show that in $\mathbb{R}^n$ with its usual distance, a set U is bounded if and only if it is totally bounded.

13.5 Show that if U is a bounded subset of $\mathbb{R}^n$, then its closure $\operatorname{cl}(U)$ is compact.

13.6 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 13.6(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 13.6(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 $x \in [0,1]$:

$$|f_b(x) - f_a(x)| \leq |b_1 - a_1| + |b_2 - a_2|.$$

(b) Use this to show that $d_\infty(f_b, f_a) \leq 2\|b - a\|_2$.

(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 13.6(a) to show that V is compact.

13.7 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?

14 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.

14.1 Convex sets

Definition 14.1 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 14.2 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

$$x = \lambda_1 v_1 + \cdots + \lambda_m v_m. \tag{48}$$

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 14.1

(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 14.3 The ***convex hull*** 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: $\text{conv}(S) = \cap_{C \subseteq X : C \text{ is convex}, S \subseteq C}\, C$.

By Theorem 14.1, 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 $S = \{v_1, \dots, v_m\}$ of vectors, in which case

$$\operatorname{conv}(S) = \operatorname{conv}(\{v_1, \dots, v_m\}) = \{\lambda_1 v_1 + \cdots + \lambda_m v_m : \lambda_1, \dots, \lambda_m \geq 0, \sum_i \lambda_i = 1\}.$$

Here are a few other examples of convex sets:

Example 14.1 (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

$$\{x \in \mathbb{R}^n : c^\top x = \delta\} \qquad \text{for some vector } c \in \mathbb{R}^n, c \neq \mathbf{0}, \text{ and a number } \delta \in \mathbb{R}. \tag{49}$$

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 (49): 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. $\lhd$

Example 14.2 (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 $\lambda x + (1-\lambda)y \in B(v, \varepsilon)$. $\lhd$

Theorem 14.2

If C is a convex subset of a normed vector space X, then also its closure $\operatorname{cl}(C)$ and its interior $\operatorname{int}(C)$ are convex.

Proof: **(Closure)** If $\operatorname{cl}(C)$ is empty, it is convex. So assume it is nonempty and let $x, y \in \operatorname{cl}(C)$ and $\lambda \in [0,1]$. Since $x, y \in \operatorname{cl}(C)$, 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,

$$\lim_{k\to\infty} \left(\lambda x_k + (1-\lambda)y_k\right) = \lambda \lim_{k\to\infty} x_k + (1-\lambda) \lim_{k\to\infty} y_k = \lambda x + (1-\lambda)y,$$

so $\lambda x + (1-\lambda)y \in \operatorname{cl}(C)$.

(Interior) If $\operatorname{int}(C)$ is empty, it is convex. So assume it is nonempty and let $x, y \in \operatorname{int}(C)$ and $\lambda \in [0,1]$. Since $x, y \in \operatorname{int}(C)$, 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. □

14.2 Polyhedra and Fourier-Motzkin elimination

Definition 14.4 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 14.3 Let us solve the system of linear inequalities

$$\begin{aligned} -2x_1 - x_2 &\leq -2 && (50)\\ 3x_1 + x_2 &\leq 9 && (51)\\ -x_1 + 2x_2 &\leq 4 && (52)\\ - x_2 &\leq 0 && (53) \end{aligned}$$

by eliminating x_1. In inequalities (50) and (52), x_1 has a negative coefficient. They impose lower bounds on x_1:

$$\begin{aligned} 1 - \tfrac{1}{2}x_2 &\leq x_1\\ -4 + 2x_2 &\leq x_1 \end{aligned}$$

Inequality (51), where x_1 has a positive coefficient, imposes an upper bound on x_1:

$$x_1 \leq 3 - \tfrac{1}{3}x_2.$$

Inequality (53), 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

$$\begin{aligned} 1 - \tfrac{1}{2}x_2 &\leq 3 - \tfrac{1}{3}x_2\\ -4 + 2x_2 &\leq 3 - \tfrac{1}{3}x_2\\ - x_2 &\leq 0 \end{aligned}$$

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 14.3 (Fourier-Motzkin elimination)

Consider the projection $\pi : \mathbb{R}^n \to \mathbb{R}^{n-1}$ that omits the first coordinate:

$$\pi(x_1,\dots,x_n) = (x_2,\dots,x_n).$$

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

$$-\frac{a_{i2}}{a_{i1}}x_2 - \cdots - \frac{a_{in}}{a_{i1}}x_n + \frac{b_i}{a_{i1}} \le x_1. \tag{54}$$

Inequalities in G give upper bounds on x_1: inequality $j \in G$ can be rewritten as

$$x_1 \le -\frac{a_{j2}}{a_{j1}}x_2 - \cdots - \frac{a_{jn}}{a_{j1}}x_n + \frac{b_j}{a_{j1}}. \tag{55}$$

Inequalities in E don't involve x_1 and impose no restrictions on x_1. Now $(x_2,\dots,x_n) \in \pi(P)$ if and only if there is an x_1 with $(x_1,x_2,\dots,x_n) \in P$. This happens if and only if $(x_2,\dots,x_n)$ satisfies the inequalities in E and we can 'squeeze in' a number x_1 between the lower bounds in (54) and the upper bounds in (55). Equivalently, the inequalities in E must hold, as well as the inequalities

$$-\frac{a_{i2}}{a_{i1}}x_2 - \cdots - \frac{a_{in}}{a_{i1}}x_n + \frac{b_i}{a_{i1}} \le -\frac{a_{j2}}{a_{j1}}x_2 - \cdots - \frac{a_{jn}}{a_{j1}}x_n + \frac{b_j}{a_{j1}}, \tag{56}$$

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! □

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.

14.3 Convex cones

Definition 14.5 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 14.4

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 14.1; induction is your friend. □

Now mimic the discussion after Theorem 14.1: 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 14.5 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 14.6

⊠ The ***convex cone*** 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 14.4, every finitely generated cone is of the form

$$\text{cone}(\{v_1, \ldots, v_m\}) = \{\lambda_1 v_1 + \cdots + \lambda_m v_m : \lambda_1, \ldots, \lambda_m \in \mathbb{R}, \lambda_1, \ldots, \lambda_m \geq 0\} \tag{57}$$

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 (57) can be rewritten as

$$\text{cone}(\{v_1, \ldots, v_m\}) = \{x \in \mathbb{R}^n : \text{there is a } \lambda \in \mathbb{R}^m \text{ with } x = V\lambda \text{ and } \lambda \geq \mathbf{0}\}. \tag{58}$$

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

$$v_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \quad \text{and} \quad v_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

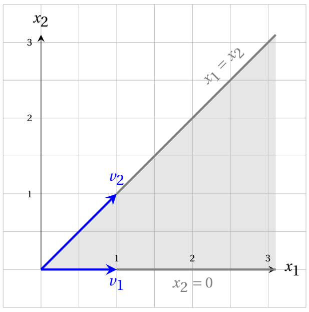
is the shaded area in the figure below; it is the set

$$\{x \in \mathbb{R}^2 : x_2 \geq 0, x_1 - x_2 \geq 0\} \tag{59}$$

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

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \lambda_1 \begin{bmatrix} 1 \\ 0 \end{bmatrix} + \lambda_2 \begin{bmatrix} 1 \\ 1 \end{bmatrix} \tag{60}$$

$$\lambda_1, \lambda_2 \geq 0 \tag{61}$$



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 (60) gives $\lambda_2 = x_1 - \lambda_1$ and the second that $\lambda_2 = x_2$. Together with (61) 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 (59).

I kept the equalities to simplify calculations. In the proof of Theorem 14.5, 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 14.5

A finitely generated cone is a polyhedron (and hence closed).

Proof: Let $C \subseteq \mathbb{R}^n$ be a finitely generated cone. By (58), we can write

$$\begin{aligned} C &= \{x \in \mathbb{R}^n : \text{there is a } \lambda \in \mathbb{R}^m \text{ with } x = V\lambda \text{ and } \lambda \geq \mathbf{0}\} \\ &= \{x \in \mathbb{R}^n : \text{there is a } \lambda \in \mathbb{R}^m \text{ with } x - V\lambda \leq \mathbf{0}, -x + V\lambda \leq \mathbf{0}, -\lambda \leq \mathbf{0}\}. \end{aligned}$$

This is the projection of the polyhedron

$$P = \{(x, \lambda) \in \mathbb{R}^{n+m} : x - V\lambda \leq \mathbf{0}, -x + V\lambda \leq \mathbf{0}, -\lambda \leq \mathbf{0}\} \tag{62}$$

by projecting away the m coordinates of λ one at a time. By Fourier-Motzkin elimination (Theorem 14.3), 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 (56) and the fact that the righthand sides in (62) 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 $\{x : Ax \le \mathbf{0}\}$ for some matrix A; see Exercise 14.3.

Exercises section 14

14.1 Prove Theorem 14.4.

14.2 Solve the following system of linear inequalities using Fourier-Motzkin elimination:

$$x_1 - x_2 \le 0, \qquad x_1 - x_3 \le 0, \qquad -x_1 + x_2 + 2x_3 \le 2, \qquad -x_3 \le -1.$$

14.3 Show that a nonempty set $P \subseteq \mathbb{R}^n$ is a polyhedral cone if and only if $P = \{x \in \mathbb{R}^n : Ax \le \mathbf{0}\}$ for some $m \times n$ matrix A.

14.4 Show that a polytope in $\mathbb{R}^n$ is a polyhedron and consequently closed.

15 Farkas' Lemma and some variants

15.1 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 15.1 (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 14.5, 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 I will not prove here (it is a special case of Exercise 15.2(d)) but that is useful in our treatment of optimization problems later is:

Theorem 15.2 (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

$$\begin{gathered} v_1^\top d > 0, \\ \vdots \\ v_k^\top d > 0. \end{gathered}$$

(b) There are nonnegative numbers $\mu_1, \ldots, \mu_k$, not all equal to zero, with

$$\mu_1 v_1 + \cdots + \mu_k v_k = \mathbf{0}.$$

15.2 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 ($a \le b$ and $b \le a$).
2. Real vectors can be written as the difference of two nonnegative vectors: if $x \in \mathbb{R}^n$, then $x = x^+ - x^-$ with $x^+, x^- \ge \mathbf{0}$ defined as follows:
$$\text{for each coordinate } i = 1, \dots, n: \quad x_i^+ = \max\{x_i, 0\} \qquad \text{and} \qquad x_i^- = \max\{-x_i, 0\}.$$
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,
$$x_1 + 2x_2 \le 3 \Leftrightarrow x_1 + 2x_2 + s = 3 \text{ for some } s \ge 0.$$

Let's practise on one variant of Farkas' Lemma:

Theorem 15.3

Exactly one of the following two problems has a solution:

(i) $Ax \le b$;

(ii) $y^\top A = \mathbf{0}^\top, y \ge \mathbf{0}, y^\top b < 0$.

Proof: $Ax \le b$ has a solution if and only if $A(x^+ - x^-) + w = b$ has a nonnegative solution (x^+, x^-, w). Thus, with $B = [A, -A, I]$:

$$Ax \le b \text{ has a solution} \qquad \Longleftrightarrow \qquad B \begin{bmatrix} x^+ \\ x^- \\ w \end{bmatrix} = b \text{ has a nonnegative solution.}$$

Using Farkas' Lemma with B instead of A, the second statement is true if and only if there is no y such that $y^\top B \ge \mathbf{0}^\top$ and $y^\top b < 0$. In other words, there is no y such that $y^\top A \ge \mathbf{0}^\top, y^\top(-A) \ge \mathbf{0}^\top, y^\top I \ge \mathbf{0}^\top$, and $y^\top b < 0$. Rewriting once more gives that there is no y such that $y^\top A = \mathbf{0}^\top, y \ge \mathbf{0}$, and $y^\top b < 0$. □

Exercises section 15

15.1 (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 15.3, 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.

15.2 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)	$Ax \le \mathbf{0}, x > \mathbf{0}$	$\mathbf{0}^\top \ne y^\top A \ge \mathbf{0}^\top, y \ge \mathbf{0}$

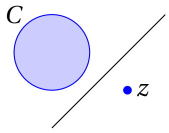
15.3 Let $A \in \mathbb{R}^{m_1 \times n}, B \in \mathbb{R}^{m_2 \times n}, C \in \mathbb{R}^{m_3 \times n}$. Show that exactly one of the following sets is nonempty:

$$\{x \in \mathbb{R}^n : Ax < \mathbf{0}, Bx \leq \mathbf{0}, Cx = \mathbf{0}\} \quad \text{or} \quad \{(y_1, y_2, y_3) \in \mathbb{R}^{m_1} \times \mathbb{R}^{m_2} \times \mathbb{R}^{m_3} : y_1^\top A + y_2^\top B + y_3^\top C = \mathbf{0}^\top, \mathbf{0} \neq y_1 \geq \mathbf{0}, y_2 \geq \mathbf{0}\}.$$

15.4 Let $A \in \mathbb{R}^{m \times n}$. Show that there exist $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^m$ with $Ax \leq \mathbf{0}, x \geq \mathbf{0}, y^\top A \geq \mathbf{0}^\top$, and $Ax + y > \mathbf{0}$.

16 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 16.1

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

$$\text{for all } y \in C: \qquad (z-x)^\top (y-x) \le 0. \tag{63}$$

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 $\{x \in \mathbb{R}^n : \|x - z\| \le \|y - z\|\} \cap C$. 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 16.1.
PROPERTY (63). 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:

$$\text{for all } \lambda \in (0,1): \qquad \|\lambda y + (1-\lambda)x - z\|^2 = \|\lambda(y-x) + (x-z)\|^2 \ge \|x-z\|^2.$$

Rewrite in terms of inner products:

$$\text{for all } \lambda \in (0,1): \qquad \lambda^2 (y-x)^\top (y-x) + 2\lambda (y-x)^\top (x-z) + (x-z)^\top (x-z) \ge (x-z)^\top (x-z).$$

Simplifying this expression and dividing by $\lambda \in (0,1)$ gives:

$$\text{for all } \lambda \in (0,1): \qquad 2(z-x)^\top (y-x) \le \lambda \|y-x\|^2.$$

Letting λ go down to zero, the right side becomes arbitrarily small, so $(z-x)^\top (y-x) \le 0$, proving (63).□

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 16.2 (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 16.1, there is an $x \in C$ with minimal distance to z. Define $c = z - x \neq \mathbf{0}$. By (63), $c^\top y \le c^\top x$ 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 16.3

Set $C \subseteq \mathbb{R}^n$ is closed and convex if and only if there is a collection $\mathscr{H}$ of halfspaces such that $C = \cap_{H \in \mathscr{H}} H$.

Proof: **⇒:** Let $\mathscr{H}$ be the collection of halfspaces H with $C \subseteq H$. Clearly, $C \subseteq \cap_{H \in \mathscr{H}} H$. To show that $C \supseteq \cap_{H \in \mathscr{H}} H$, let $x \in \cap_{H \in \mathscr{H}} H$ and suppose that $x \notin C$. By Theorem 16.2, there is a halfspace H' with $C \subseteq H'$ and $x \notin H'$. So $H' \in \mathscr{H}$ and $x \notin \cap_{H \in \mathscr{H}} H$, a contradiction.

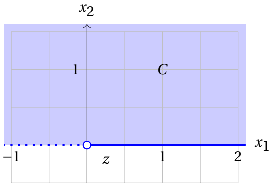
⇐: Let C be the intersection of halfspaces. Each halfspace is closed and convex. As the intersection of closed sets, C is closed; as the intersection of convex sets, C is convex. □

In Theorem 16.2, we need to assume that C is closed:

Example 16.1 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 16.4 (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 15.2, 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 13.2) 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 16.5 (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 16.4, 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 16

16.1 Prove that the vector $x \in C$ minimizing the distance to z in Theorem 16.1 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.

16.2 Separating hyperplane theorems typically say when two *convex* sets (in Theorems 16.2 and 16.4, 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 16.2 and 16.4 and $C_1 \cap C_2 = \emptyset$ in Theorem 16.5). 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 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.)

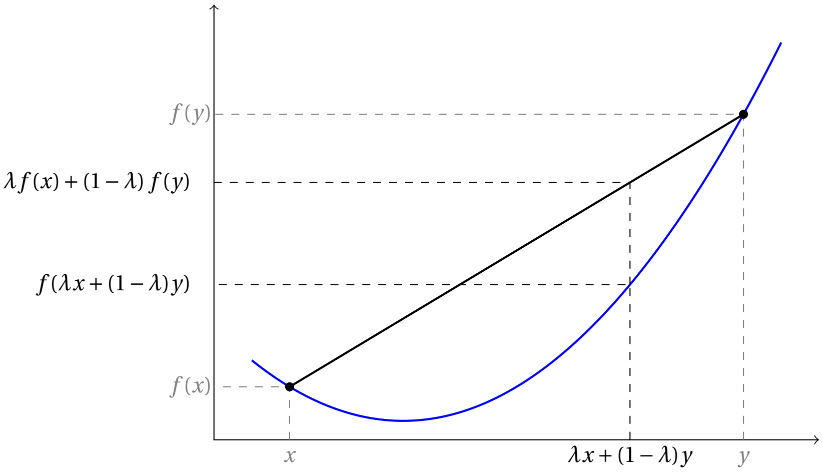
17 Convex functions and variants

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 17.1;
- ☒ that there are various ways to transform convex functions to new ones (Theorem 17.6);
- ☒ that variants of convex functions facilitate and add structure to the solution of optimization problems (subsection 17.4).

17.1 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.



$f(y)$
$\lambda f(x) + (1-\lambda) f(y)$
$f(\lambda x + (1-\lambda) y)$
$f(x)$
x
$\lambda x + (1-\lambda) y$
y

Definition 17.1 Let $C \subseteq \mathbb{R}^n$ be a convex set. A function $f : C \to \mathbb{R}$ is ***convex*** if

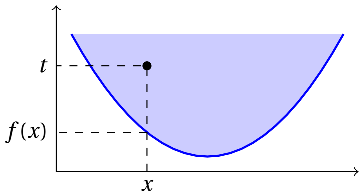
$$\text{for all } x, y \in C \text{ and all } \lambda \in [0,1]: \qquad f(\lambda x + (1-\lambda) y) \le \lambda f(x) + (1-\lambda) f(y). \tag{64}$$

Example 17.1 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

$$\begin{aligned} f\big(\lambda x + (1-\lambda) y\big) \le \lambda f(x) + (1-\lambda) f(y) \quad &\iff \quad \big(\lambda x + (1-\lambda) y\big)^2 \le \lambda x^2 + (1-\lambda) y^2 \\ &\iff \quad \lambda^2 x^2 + 2\lambda(1-\lambda) xy + (1-\lambda)^2 y^2 \le \lambda x^2 + (1-\lambda) y^2 \\ &\iff \quad \lambda(1-\lambda)(x^2 - 2xy + y^2) \ge 0 \\ &\iff \quad \lambda(1-\lambda)(x-y)^2 \ge 0. \end{aligned}$$

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 17.2 The ***epigraph*** of a function $f : C \to \mathbb{R}$ with domain $C \subseteq \mathbb{R}^n$ is the set

$$\text{epi}(f) = \{(x, t) \in C \times \mathbb{R} : t \geq f(x)\} \subseteq \mathbb{R}^{n+1}.$$

Theorem 17.1 (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 epi(f) is a convex set.

Proof: **(⇒)** Assume f is convex. Let $(x, t), (y, t') \in \text{epi}(f)$, and $\lambda \in [0, 1]$. To show:

$$\lambda(x, t) + (1-\lambda)(y, t') = (\lambda x + (1-\lambda)y, \lambda t + (1-\lambda)t') \in \text{epi}(f).$$

Since $(x, t), (y, t') \in \text{epi}(f)$ 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 $(\lambda x + (1-\lambda)y, \lambda t + (1-\lambda)t') \in \text{epi}(f)$.

(⇐) Assume epi(f) is a convex set. Then for all $x, y \in C$ and $\lambda \in [0, 1]$: $(x, f(x)), (y, f(y)) \in \text{epi}(f)$, so $(\lambda x + (1-\lambda)y, \lambda f(x) + (1-\lambda) f(y)) \in \text{epi}(f)$. The latter means $f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y)$. □

Example 17.2 The absolute-value function $f : \mathbb{R} \to \mathbb{R}$ with $f(x) = |x| = \max\{-x, x\}$ is convex: its epigraph

$$\text{epi}(f) = \left\{(x, t) \in \mathbb{R}^2 : t \geq f(x) = \max\{-x, x\}\right\} = \left\{(x, t) \in \mathbb{R}^2 : t \geq -x, t \geq x\right\}$$

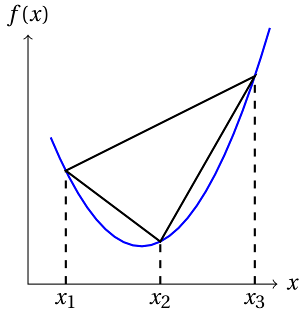
is polyhedral (it is the set of solutions to two linear inequalities), hence convex. ◁

Analogously, we may replace the weak inequality in the epigraph with a strict one:

$$f : C \to \mathbb{R} \text{ is convex} \qquad \Longleftrightarrow \qquad \{(x, t) \in C \times \mathbb{R} : t > f(x)\} \text{ is a convex set.} \tag{65}$$

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 17.2

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 $x_1 < x_2 < x_3$:

$$\frac{f(x_2) - f(x_1)}{x_2 - x_1} \le \frac{f(x_3) - f(x_1)}{x_3 - x_1} \le \frac{f(x_3) - f(x_2)}{x_3 - x_2}. \tag{66}$$

(b) Conversely, each of the inequalities in (66) 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:

$$\begin{aligned}\frac{f(x_2) - f(x_1)}{x_2 - x_1} &= \frac{f(\lambda x_1 + (1-\lambda)x_3) - f(x_1)}{\lambda x_1 + (1-\lambda)x_3 - x_1}\\ &\le \frac{\lambda f(x_1) + (1-\lambda) f(x_3) - f(x_1)}{\lambda x_1 + (1-\lambda)x_3 - x_1}\\ &= \frac{(1-\lambda)\big(f(x_3) - f(x_1)\big)}{(1-\lambda)(x_3 - x_1)}\\ &= \frac{f(x_3) - f(x_1)}{x_3 - x_1}.\end{aligned}$$

This proves the first inequality in (66); 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 (66) 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 (66) becomes

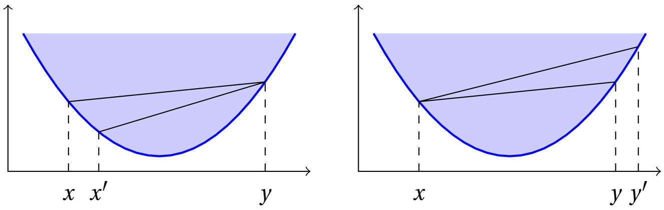
$$\frac{f\big(\lambda x + (1-\lambda) y\big) - f(x)}{\lambda x + (1-\lambda) y - x} = \frac{f\big(\lambda x + (1-\lambda) y\big) - f(x)}{(1-\lambda)(y - x)} \le \frac{f(y) - f(x)}{y - x}.$$

Rearranging terms gives (64). □

Rewriting the inequalities in (66), 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 17.3 (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

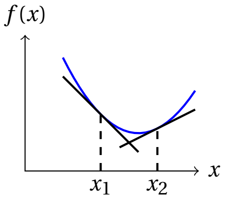
$$Q(x, y) = \frac{f(y) - f(x)}{y - x} \qquad (\text{with } x, y \in I, x \neq y)$$

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 $Q(x, y) \le Q(x', y)$. 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 (66) gives $Q(x, y) \le Q(x', y)$.

Conversely, if the difference quotient is weakly increasing in both variables, then $Q(x_1, x_2) \le Q(x_1, x_3)$ for all $x_1, x_2, x_3 \in I$ with $x_1 < x_2 < x_3$: the first inequality in (66) holds, so f is convex by Theorem 17.2. □

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 17.4 (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: $f'(x) \le f'(y)$. For all $h > 0$ with $x < x + h \le y - h < y$, Theorem 17.3 gives

$$Q(x, x + h) = \frac{f(x + h) - f(x)}{h} \le Q(y - h, y) = \frac{f(y) - f(y - h)}{h}.$$

Letting h tend to zero, these terms converge to $f'(x)$ and $f'(y)$, respectively, so $f'(x) \le f'(y)$.

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, $f'(\alpha) \le f'(\beta)$. By Theorem 17.2, 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 17.3 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 17.4 (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 17.4, convex functions lie above their tangents:

Theorem 17.5 (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 $\lambda \in (0,1]$:

$$f\left(\lambda x + (1-\lambda)x^*\right) \leq \lambda f(x) + (1-\lambda) f(x^*) \qquad \Longleftrightarrow \qquad f\left(x^* + \lambda(x - x^*)\right) \leq f(x^*) + \lambda\left(f(x) - f(x^*)\right).$$

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 17.6

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 $\mathrm{epi}(f) = \cap_{i\in I}\,\mathrm{epi}(f_i)$ 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:

$$f(\lambda x + (1-\lambda)y) \le \lambda f(x) + (1-\lambda) f(y).$$

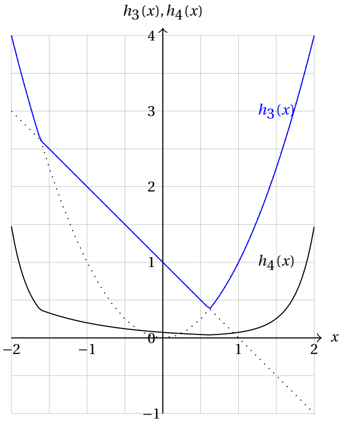
Using that g is nondecreasing to prove the first inequality and convex to prove the second, we find

$$g\big(f(\lambda x + (1-\lambda)y)\big) \le g\big(\lambda f(x) + (1-\lambda) f(y)\big) \le \lambda g(f(x)) + (1-\lambda) g(f(y)). \qquad \square$$

Example 17.5 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 17.6, 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. ◁



17.2 Variants of convex functions

There are a few common variants of convex functions that pop up in economics; here are some of them.

Definition 17.3 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 17.1, a function is concave if and only if the set of points *under* its graph is a convex set; and by Theorem 17.5, concave functions lie *below* their tangent lines.

Theorem 17.7

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 17.8 (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 $I \cap (x^*, \infty)$.

17.3 More on continuity and differentiability

Convex functions are continuous on the interior of their domain. Exercise 17.1 shows that they need not be continuous in boundary points.

Theorem 17.9 (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 17.5, 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 17.10

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

$$\text{for all } x \in C: \qquad f(x) \geq f(z) + a^\top (x - z). \tag{67}$$

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:[3]

Theorem 17.11 (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 17.4) extends to convex functions

$$(x_1, \dots, x_n) \mapsto f(x_1, \dots, x_n)$$

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 17.12

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.

17.4 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.

[3] This presumes that you know a little about differentiable functions of several variables. If you don't, Section 18 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 17.13 (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 17.11 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 17.14 (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. □

Quasiconcavity is a popular assumption in traditional results about economic and game-theoretic equilibria: Theorem 17.14 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.

17.5 Postponed proofs

17.5.1 Proof of Theorem 17.8

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.

17.5.2 Proof of Theorem 17.9

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 $c \in \operatorname{int}(C)$: 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, $\mathbf{0} \in \operatorname{int}(D)$, and $h(\mathbf{0}) = 0$. It suffices to prove that h is continuous at $\mathbf{0}$.

By construction:

$$\text{for all } i = 1, \ldots, n: \qquad \delta e_i \in D \text{ and } \delta(-e_i) \in D$$

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

$$\text{if } x \in D, \text{ then:} \qquad h(x) \ge -h(-x). \tag{68}$$

Let $x \in \mathbb{R}^n$ have $\|x\|_\infty < \frac{\delta}{n}$. Then x is a convex combination of $\delta e_1, \ldots, \delta e_n, \delta(-e_1), \ldots, \delta(-e_n), \mathbf{0} \in D$:

$$x = \sum_{i : x_i > 0} \frac{x_i}{\delta} \delta e_i + \sum_{i : x_i < 0} \frac{-x_i}{\delta} \delta(-e_i) + \left(1 - \sum_i \frac{|x_i|}{\delta}\right) \mathbf{0}.$$

Let $\beta = \frac{1}{\delta} \max\{h(\delta e_1), \ldots, h(\delta e_n), h(\delta(-e_1)), \ldots, h(\delta(-e_n))\}$. By (68), $\beta \ge 0$. 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 (68), 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.

17.5.3 Proof of Theorem 17.10

By expression (65), 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 16.4, there is a vector $(a, \tau) \neq \mathbf{0}$ such that

$$\text{for all } x \in C: \qquad a^\top x + \tau t \le a^\top z + \tau f(z). \tag{69}$$

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 (69) 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 (67).

17.5.4 Proof of Theorem 17.11

That $f'(z)$ is a subgradient follows as in Theorem 17.5: Let $x \in C$. By convexity of f, for each $\lambda \in (0,1]$:

$$f(\lambda x + (1-\lambda)z) \le \lambda f(x) + (1-\lambda) f(z) \quad \Longleftrightarrow \quad f(z + \lambda(x-z)) \le f(z) + \lambda(f(x) - f(z)).$$

Rearranging terms and dividing by λ gives:

$$\frac{f(z+\lambda(x-z)) - f(z)}{\lambda} \le f(x) - f(z).$$

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:

$$\text{for all } x \in C: \qquad f(x) \ge f(z) + a^\top (x-z).$$

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

$$f(z+tv) \ge f(z) + a^\top (tv).$$

So on a neighborhood of $t = 0$, the function

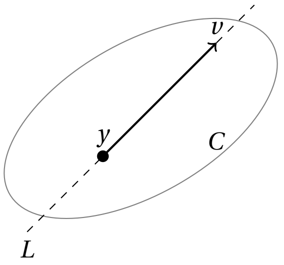
$$t \mapsto f(z+tv) - f(z) - a^\top (tv)$$

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 $f'(z) = a^\top$.

17.5.5 Proof of Theorem 17.12

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 17.4; this derivative is $v^\top H_f(y) v$. And $v^\top H_f(y) v \ge 0$ 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 17

17.1 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?

17.2 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.

Drawing graphs will help you with the intuition.

17.3 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.

17.4 Give an example of a quasiconcave function with a local maximum that is not a global maximum.

17.5 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?

17.6 **(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 17.6.

17.7 Prove expression (65).

17.8 **(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.

18 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

$$\lim_{h \to 0} \frac{f(x+h) - f(x)}{h} = f'(x). \tag{70}$$

The number $f'(x)$ is referred to as the derivative of f at x. We can rewrite the definition as

$$\lim_{h \to 0} \frac{f(x+h) - f(x) - f'(x)h}{h} = 0. \tag{71}$$

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 $x_1, \dots, x_n$:

- ☒ 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 18.1 (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:

$$\begin{gathered} f \text{ is differentiable} \\ \Downarrow \\ f \text{ has directional derivatives in each direction} \\ \Downarrow \\ f\text{'s partial derivatives with respect to each variable } x_1, \dots, x_n \text{ exist} \end{gathered}$$

and the implications in the other direction are false in general, but true under additional assumptions.

18.1 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 18.1** 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 18.1 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

$$\nabla f(x)=\left(\frac{\partial f(x)}{\partial x_1},\frac{\partial f(x)}{\partial x_2},\frac{\partial f(x)}{\partial x_3}\right)=\left(3x_3^2-8x_1^3x_2^6,\ -12x_1^4x_2^5,\ 6x_1x_3\right).$$

$\lhd$

18.2 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 18.2** 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 18.2 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. ◁

18.3 Differentiability

The definition of differentiability for a function of n variables closely mimics the one-variable case (71):

Definition 18.3

☒ 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

$$\lim_{h\to\mathbf{0}}\frac{|f(x+h)-f(x)-f'(x)h|}{\|h\|}=0. \tag{72}$$

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 (72), 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 18.1

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 18.3 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

$$D_d f(0,0) = \frac{d_1|d_2|}{\sqrt{d_1^2+d_2^2}}. \tag{73}$$

But f is not differentiable at $x = (0,0)$: if it were, then Theorem 18.1 says that the directional derivative must be a linear function of the direction,

$$D_d f(0,0) = f'(0,0)d.$$

However, (73) 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 18.2

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 18.1 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 18.4 Function $f:\mathbb{R}^2\to\mathbb{R}$ with $f(x_1,x_2)=2x_1+x_1^2x_2$ 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

$$\ell(x_1,x_2)=f(x^*)+\nabla f(x^*)(x-x^*)=3+4(x_1-1)+(x_2-1).$$ ◁

18.4 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 18.3 (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 (72), 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 $\|h\| < \delta$:

$$\begin{aligned}\left|f(x+h)-f(x)\right| &\le \left|f(x+h)-f(x)-f'(x)h\right| + \left|f'(x)h\right| \\ &< \|h\| + \left|f'(x)h\right| \le \|h\| + \|f'(x)\|\,\|h\| = \|h\|(1+\|f'(x)\|) \\ &< \delta(1+\|f'(x)\|) < \varepsilon.\end{aligned}$$ □

18.5 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 18.4 (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 $\|d\| \le 1$ is given by $\nabla f(x)/\|\nabla f(x)\|$.

Proof: Differentiability, the Cauchy-Schwarz inequality, and the assumption that $\|d\| \le 1$, give

$$D_d f(x) = \nabla f(x) d \le \|\nabla f(x)\|\,\|d\| \le \|\nabla f(x)\|,$$

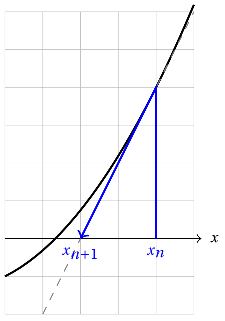
with equality throughout if and only if $d = \nabla f(x)/\|\nabla f(x)\|$. □

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.

18.6 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

$$x_{n+1} = x_n - f(x_n)/f'(x_n). \tag{74}$$



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 18.5 (Heron's formula) Consider the equation $f(x) = x^2 - a = 0$ for given $a > 0$. Then (74) becomes

$$x_{n+1} = x_n - \frac{x_n^2 - a}{2x_n} = \frac{2x_n^2 - x_n^2 + a}{2x_n} = \frac{1}{2}\left(x_n + \frac{a}{x_n}\right). \tag{75}$$

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 (75) converges to $\sqrt{a}$. Using standard algebra, we find that for all $n \in \mathbb{N}: \sqrt{a} \le x_{n+1} \le x_n$. 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 18

18.1 Consider $f: \mathbb{R}^2 \to \mathbb{R}$ with

$$f(x) = \begin{cases} \frac{x_1^2 x_2}{x_1^4 + x_2^2}\|x\| & \text{if } x \ne \mathbf{0}, \\ 0 & \text{if } x = \mathbf{0}. \end{cases}$$

(a) Show that f is continuous at $x = \mathbf{0}$. HINT: Use that $(x_1^2 - x_2)^2 \ge 0$.

(b) Show that at $x = \mathbf{0}$, function f has a directional derivative in each direction $d \ne \mathbf{0}$ 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).

19 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.

19.1 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 19.1

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 18.1 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 $\nabla f(x^*) = \mathbf{0}$. □

19.2 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:

$$\text{maximize } f(x) \text{ with } h_1(x) \leq 0, \ldots, h_p(x) \leq 0. \tag{76}$$

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. 15.2), 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 (76) and that the first r constraints are binding: $h_1(x^*) = \cdots = h_r(x^*) = 0$ and $h_{r+1}(x^*) < 0, \ldots, h_p(x^*) < 0$. Rearranging the constraints if necessary, this is without loss of generality. Recall that under suitable differentiability assumptions (e.g., Theorem 18.1) — 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

$$\nabla f(x^*)d > 0 \qquad \text{but} \qquad \nabla h_1(x^*)d < 0, \ldots, \nabla h_r(x^*)d < 0, \tag{77}$$

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 (77), 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

$$\mu_0 \nabla f(x^*) - \mu_1 \nabla h_1(x^*) - \cdots - \mu_r \nabla h_r(x^*) = \mathbf{0}. \tag{78}$$

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 19.2 (Fritz John conditions for problems with inequality constraints)

If x^* is a local maximum of optimization problem (76), 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)	$\mu_0 \nabla f(x^*) - \mu_1 \nabla h_1(x^*) - \cdots - \mu_p \nabla h_p(x^*) = \mathbf{0}$	(79)
(feasibility)	$h_j(x^*) \leq 0$ for all $j = 1, \ldots, p$	(80)
(complementary slackness)	$\mu_j \geq 0$ and $\mu_j h_j(x^*) = 0$ for all $j = 1, \ldots, p$	(81)

To see why conditions (81) 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 (78) 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 19.2 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 19.3 (Karush-Kuhn-Tucker conditions for problems with inequality constraints)

If x^* is a local maximum of optimization problem (76), 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 $\mu_1, \dots, \mu_p \geq 0$ with

$$\begin{aligned} &\text{(gradient condition)} && \nabla f(x^*) - \mu_1 \nabla h_1(x^*) - \cdots - \mu_p \nabla h_p(x^*) = \mathbf{0} && (82)\\ &\text{(feasibility)} && h_j(x^*) \leq 0 \text{ for all } j = 1, \dots, p && (83)\\ &\text{(complementary slackness)} && \mu_j \geq 0 \text{ and } \mu_j h_j(x^*) = 0 \text{ for all } j = 1, \dots, p && (84) \end{aligned}$$

Under additional assumptions, we can dispense with the linear dependence scenario in the previous theorem and each maximum must satisfy the KKT conditions:

Theorem 19.4

If x^* is a local maximum of optimization problem (76) 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: $h_1, \dots, h_p$ are convex functions and there is an x_0 with $h_1(x_0) < 0, \dots, h_p(x_0) < 0$.
CASE 3: $h_1, \dots, h_p$ 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 19.3. The proof for the other two cases is postponed to subsection 19.7. 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 19.5

If x^* satisfies the Karush-Kuhn-Tucker conditions of optimization problem (76) and

- ☒ goal function f is concave and differentiable,
- ☒ constraint functions $h_1, \dots, h_p$ 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 19.2 and 19.3 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 19.1 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 19.4, 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.

19.3 A worked example with inequality constraints

Consider the problem:

$$\text{maximize } x_1 + \ln(1 + x_2) \text{ with } 16x_1 + x_2 \le 495, \quad x_1 \ge 0, \quad x_2 \ge 0.$$

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 $\{x \in \mathbb{R}^2 : 16x_1 + x_2 \le 495, x_1 \ge 0, x_2 \ge 0\}$ is nonempty: it contains $(0,0)$, closed, since it is the intersection of three closed halfspaces, and bounded: $0 \le x_1 \le 495/16$ and $0 \le x_2 \le 495$. 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 (76): maximize $f(x) = x_1 + \ln(1 + x_2)$ subject to $h_1(x) = 16x_1 + x_2 - 495 \le 0, h_2(x) = -x_1 \le 0, h_3(x) = -x_2 \le 0$. The constraints h_1, h_2, h_3 are affine functions, so we are in Case 3 of Theorem 19.4: a maximum must satisfy the KKT conditions (82), (83), and (84). 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:

$$1 - 16\mu_1 + \mu_2 = 0 \tag{85}$$

$$\tfrac{1}{1+x_2} - \mu_1 + \mu_3 = 0 \tag{86}$$

☒ Feasibility:

$$16x_1 + x_2 \le 495 \tag{87}$$

$$x_1 \ge 0 \tag{88}$$

$$x_2 \ge 0 \tag{89}$$

⊠ Complementary slackness:

$$\mu_1, \mu_2, \mu_3 \ge 0 \tag{90}$$

$$\mu_1(16x_1 + x_2 - 495) = 0 \tag{91}$$

$$\mu_2 x_1 = 0 \tag{92}$$

$$\mu_3 x_2 = 0 \tag{93}$$

By (85) and (90): $16\mu_1 = 1 + \mu_2 \ge 1$, so $\mu_1 \ge 1/16$. By (91): $16x_1 + x_2 = 495$.

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 (93). By (86): $\mu_1 = 1/496$, contradicting $\mu_1 \ge 1/16$.
3. $x_1 > 0, x_2 = 0$. Then $\mu_2 = 0$ by (92) and $\mu_1 = 1/16$ by (85). But (86) gives $\mu_1 = 1 + \mu_3 \ge 1$, contradicting $\mu_1 = 1/16$.
4. $x_1 > 0, x_2 > 0$. Then $\mu_2 = \mu_3 = 0$ by complementary slackness. Substitution in (85) gives $\mu_1 = 1/16$. Substitution in (86) 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 $f(30, 15) = 30 + \ln(16)$.

19.4 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:

$$\text{maximize } f(x) \text{ with } g_1(x) = 0, \dots, g_m(x) = 0, h_1(x) \le 0, \dots, h_p(x) \le 0, \tag{94}$$

where $f, g_1, \dots, g_m, h_1, \dots, h_p$ 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 $\mu_1, \dots, \mu_p$ with the inequality constraints $h_1(x) \le 0, \dots, h_p(x) \le 0$. And we introduce new multipliers $\lambda_1, \dots, \lambda_m$ for the equality constraints $g_1(x) = 0, \dots, g_m(x) = 0$.

Theorem 19.6 (Fritz John conditions for problems with mixed constraints)

If x^* is a local maximum of optimization problem (94), then it satisfies these Fritz John conditions: there are numbers $\lambda_1, \dots, \lambda_m$ and $\mu_0, \mu_1, \dots, \mu_p \ge 0$, not all zero, satisfying gradient condition

$$\mu_0 \nabla f(x^*) - \lambda_1 \nabla g_1(x^*) - \cdots - \lambda_m \nabla g_m(x^*) - \mu_1 \nabla h_1(x^*) - \cdots - \mu_p \nabla h_p(x^*) = \mathbf{0}, \tag{95}$$

and

$$\text{(feasibility)} \qquad g_1(x^*) = 0, \dots, g_m(x^*) = 0, h_1(x^*) \le 0, \dots, h_p(x^*) \le 0 \tag{96}$$

$$\text{(compl. slackness)} \qquad \mu_j \ge 0 \text{ and } \mu_j h_j(x^*) = 0 \text{ for all } j = 1, \dots, p \tag{97}$$

The proofs for this subsection are less insightful, so I postpone them to subsection 19.7. Multipliers

μ_j for the inequality constraints $h_j(x) \le 0$ are nonnegative ($\mu_j \ge 0$), 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 19.1 (Problems with equality constraints) In a problem with only equality constraints,

$$\text{maximize } f(x) \text{ with } g_1(x) = 0, \dots, g_m(x) = 0, \tag{98}$$

these conditions simplify considerably, because there are no complementary slackness conditions: if x^* is a maximum of (98), then it is feasible and there are Lagrange multipliers $\mu_0 \ge 0$ and $\lambda_1, \dots, \lambda_m$, 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 19.6 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 19.3:

Theorem 19.7

If x^* is a local maximum of optimization problem (94), then the gradients of the binding constraints at x^* are linearly dependent or there are numbers $\lambda_1, \dots, \lambda_m$ and $\mu_1, \dots, \mu_p \ge 0$ such that

$$\nabla f(x^*) - \lambda_1 \nabla g_1(x^*) - \cdots - \lambda_m \nabla g_m(x^*) - \mu_1 \nabla h_1(x^*) - \cdots - \mu_p \nabla h_p(x^*) = \mathbf{0}, \tag{99}$$

and

$$\begin{aligned} &\text{(feasibility)} && g_1(x^*) = 0, \dots, g_m(x^*) = 0, h_1(x^*) \le 0, \dots, h_p(x^*) \le 0 && (100)\\ &\text{(compl. slackness)} && \mu_j \ge 0 \text{ and } \mu_j h_j(x^*) = 0 \text{ for all } j = 1, \dots, p && (101) \end{aligned}$$

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 (99), (100), (101) 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 (99), (100), (101) 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 19.8 (Fritz John conditions under concave/affine constraints: $\mu_0 = 1$)

Let x^* be a local maximum of (94). If the equality constraints $g_1, \dots, g_m$ are affine functions and the inequality constraints $h_1, \dots, h_p$ 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 $B(p,w) = \{x \in \mathbb{R}^n : p^\top x \le w, x \ge \mathbf{0}\}$. One final observation:

Theorem 19.9 (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.,

$$\mathscr{L}(x^*, \lambda, \mu) \ge \mathscr{L}(x, \lambda, \mu) \text{ for all feasible } x,$$

then x^* also solves the optimization problem (94).

19.5 A first worked example with mixed constraints

Consider the problem:

$$\text{maximize } x_1^2 - 3x_2^2 \text{ with } x_1^2 + x_2^2 = 17, \quad x_1 - x_2 \le 3.$$

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 $h(x) = x_1 - x_2 - 3 \le 0$. 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 103:

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 (99), (100), and (101) 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:

$$2x_1 - 2\lambda x_1 - \mu = 0 \qquad (102)$$

$$-6x_2 - 2\lambda x_2 + \mu = 0 \qquad (103)$$

☒ Feasibility:

$$x_1^2 + x_2^2 = 17 \qquad (104)$$

$$x_1 - x_2 \le 3 \qquad (105)$$

☒ Complementary slackness for the inequality constraint:

$$\mu \geq 0 \tag{106}$$

$$\mu(x_1 - x_2 - 3) = 0 \tag{107}$$

Case 1: $\mu = 0$. Equations (102) and (103) then give

$$\begin{aligned} 2x_1(1-\lambda) &= 0 & \iff && x_1 = 0 \text{ or } \lambda = 1 \\ -2x_2(3+\lambda) &= 0 & \iff && x_2 = 0 \text{ or } \lambda = -3. \end{aligned}$$

This gives four possibilities:

1. $x_1 = 0$ and $x_2 = 0$. This is not feasible: it contradicts (104).
2. $x_1 = 0$ and $\lambda = -3$. Then (104) gives $x_2^2 = 17$, so $x_2 = -\sqrt{17}$ or $x_2 = \sqrt{17}$. The first violates (105), 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 (104) gives $x_1^2 = 17$, so $x_1 = -\sqrt{17}$ or $x_1 = \sqrt{17}$. The second violates (105), 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 (107) gives $x_1 - x_2 = 3$, so $x_2 = x_1 - 3$. Substitution in (104) 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 (102) and (103) 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 (102) and (103) 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 (106).

STEP 3: Comparing all solution candidates, we find maximal value 17 in feasible point $(x_1, x_2) = (-\sqrt{17}, 0)$.

19.6 A second worked example with mixed constraints

Consider the problem:

$$\text{maximize } -x_1^2 - x_2^2 - \cdots - x_n^2 \qquad \text{subject to } x_1 \geq 0, \dots, x_n \geq 0, \quad x_1 + \cdots + x_n = 1.$$

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 $i = 1, \dots, n$. 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 19.8 says that a solution must satisfy a solution must satisfy the Fritz John conditions (99), (100), and (101) 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 $i = 1, \dots, n$:

$$-2x_i - \lambda + \mu_i = 0. \tag{108}$$

☒ Feasibility:

$$x_1 + \cdots + x_n = 1, \tag{109}$$

$$x_1 \geq 0, \dots, x_n \geq 0. \tag{110}$$

☒ Complementary slackness for the inequality constraints, i.e., for each $i = 1, \dots, n$:

$$\mu_i \geq 0 \qquad \text{and} \qquad \mu_i x_i = 0. \tag{111}$$

By (109) and (110), *some* coordinate i must have $x_i > 0$. By (111), $\mu_i = 0$. By (108), $\lambda = -2x_i < 0$. But then *all* coordinates must have $x_j > 0$: otherwise $x_j = 0$ and (108) would give $\mu_j = \lambda < 0$, contradicting (111). Since all coordinates of x are positive, all μ_i are zero by complementary slackness, so all coordinates of x are equal by (108). Since they add up to one, we find $x = (1/n, \dots, 1/n)$.

We found only one solution candidate, $x = (1/n, \dots, 1/n)$. We also argued that there must be a solution. So the function is maximized in $x = (1/n, \dots, 1/n)$. Its maximal value is $-(1/n)^2 - \cdots - (1/n)^2 = -1/n$.

19.7 Postponed proofs

19.7.1 Proof of Theorem 19.4

As mentioned, the first case just restates the preceding theorem. So it remains to prove:
CASE 2: For each constraint $j = 1, \dots, p$, Theorem 17.11 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 $h_{r+1}(x^*) < 0, \dots, h_p(x^*) < 0$. Then

$$\nabla h_j(x^*)(x_0 - x^*) < 0$$

for $j = 1, \dots, r$. If there is no nonnegative solution $\mu_1, \dots, \mu_r$ 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 15.1; 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 ($j = 1, \dots, r$) 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 $j = 1, \dots, r$ (regardless of ε) and $\nabla f(x^*) d > 0$. As in the proof of Theorem 19.2, 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 $\mu_1, \dots, \mu_r \geq 0$ 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 $\nabla f(x^*) y > 0, \nabla h_1(x^*) y \leq 0, \dots, \nabla h_r(x^*) y \leq 0$. 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.

19.7.2 Proof of Theorem 19.5

Let x^* satisfy the KKT conditions with nonnegative multipliers $\mu_1, \dots, \mu_p$. Let x be feasible. To see that x^* is a maximum, we show that $f(x) - f(x^*) \leq 0$:

$$\begin{aligned} f(x) - f(x^*) &\leq \nabla f(x^*)(x - x^*) && \text{(by concavity of } f\text{, Thm. 17.11)} \\ &= \sum_{j=1}^{p} \mu_j \nabla h_j(x^*)(x - x^*) && \text{(by the KKT conditions)} \\ &\leq \sum_{j=1}^{p} \mu_j \left(h_j(x) - h_j(x^*) \right) && \text{(by convexity of } h_j\text{, Thm. 17.11)} \\ &= \sum_{j=1}^{p} \mu_j h_j(x) && \text{(by complementary slackness)} \\ &\leq 0 && \text{(since } \mu_j \geq 0 \text{ and } h_j(x) \leq 0\text{)} \end{aligned}$$

19.7.3 Proof of Theorem 19.6

We derive the result as a consequence of the following more general theorem:

Theorem 19.10 (Fritz John conditions)

Let x^* be a local maximum of (94). Then there exist numbers $\lambda_1, \dots, \lambda_m$ and $\mu_0, \mu_1, \dots, \mu_p$, 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 $j = 0, 1, \dots, p$.

(c) In each open neighborhood N of x^*, there is an $x \in N$ with $\lambda_i g_i(x) > 0$ for all $i = 1, \dots, m$ 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 19.4 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

$$F_k(x) = f(x) - \frac{1}{2} k \sum_{i=1}^m \left(g_i(x)\right)^2 - \frac{1}{2} k \sum_{j=1}^p \left(h_j^+(x)\right)^2 - \frac{1}{2} \|x - x^*\|^2.$$

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 13.3).

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 $F_k(x^k) \geq F_k(x^*)$:

$$f(x^*) = F_k(x^*) \leq F_k(x^k) = f(x^k) - \frac{1}{2} k \sum_{i=1}^m \left(g_i(x^k)\right)^2 - \frac{1}{2} k \sum_{j=1}^p \left(h_j^+(x^k)\right)^2 - \frac{1}{2} \|x^k - x^*\|^2. \tag{112}$$

Rearranging terms and dividing both sides by k gives

$$\frac{f(x^*)}{k} \leq \frac{f(x^k)}{k} - \frac{1}{2} \sum_{i=1}^m \left(g_i(x^k)\right)^2 - \frac{1}{2} \sum_{j=1}^p \left(h_j^+(x^k)\right)^2 - \frac{\|x^k - x^*\|^2}{2k} \leq \frac{f(x^k)}{k}. \tag{113}$$

Since f is bounded on the compact set S and $\|x^k - x^*\| \leq \varepsilon$, the left- and right-hand side of (113) converge to zero, so that

$$\lim_{k \to \infty} \sum_{i=1}^m \left(g_i(x^k)\right)^2 = \lim_{k \to \infty} \sum_{j=1}^p \left(h_j^+(x^k)\right)^2 = 0. \tag{114}$$

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 (114), this limit is feasible in the original optimization problem, where x^* is optimal. Combining this with (112) 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

$$\nabla f(x^k) - \sum_{i=1}^{m} \lambda_i^k \nabla g_i(x^k) - \sum_{j=1}^{p} \mu_j^k \nabla h_j(x^k) - (x^k - x^*) = \mathbf{0}. \tag{115}$$

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 (115) 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 19.6 we only need to show that complementary slackness is implied by the conditions of Theorem 19.10. If $h_j(x^*) = 0$, this is trivial. If $h_j(x^*) < 0$, then $h_j^+(x^k) = 0$ for large k, so $\mu_j = 0$.

19.7.4 Proof of Theorem 19.8

By Theorem 19.10, 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 19.10(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.

19.7.5 Proof of Theorem 19.9

For all feasible x, writing out that $\mathscr{L}(x^*, \lambda, \mu) \geq \mathscr{L}(x, \lambda, \mu)$ and rearranging terms gives

$$f(x^*) - f(x) \geq \sum_{i=1}^{m} \lambda_i \left(g_i(x^*) - g_i(x)\right) + \sum_{j=1}^{p} \mu_j \left(h_j(x^*) - h_j(x)\right). \tag{116}$$

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 (116) is nonnegative. Hence, $f(x^*) \geq f(x)$ for all feasible x, making x^* the desired maximum.

Exercises section 19

19.1 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) $0 \leq x_2 \leq x_1^3, x_2 \geq 2x_1^3 - 6x_1^2 + 12x_1 - 8$

19.2 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 $0 \leq x_2 \leq x_1^3, x_2 \geq 2x_1^3 - 6x_1^2 + 12x_1 - 8$.

19.3 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?

19.4 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?

19.5 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.

19.6 Use the Fritz John conditions to find a rectangle with perimeter equal to one and with maximal area.

19.7 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 $x_1 - x_2 = -1, x_1 - 2x_2 + x_3 = -3$.

19.8 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 $x_2 x_3 = -2, x_1^2 + x_4^2 = 1$.

19.9 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 $x_1^2 + x_2^2 + x_3^2 \leq 5, x_1^2 + x_3^2 \leq 1$.

20 The dynamic programming algorithm

20.1 Introduction

Dynamic optimization is about making good decisions at several consecutive stages, often modeled as different moments in time. In addition to the direct effect of a decision, there is often an impact on future outcomes and feasible options and a decision maker has to be careful to find the appropriate combination of current and future benefits. Modeling decisions in the presence of time involve a number of considerations:

- ☒ Choice between deterministic or stochastic problems: are there random variables like income shocks or other variables whose outcomes are outside our control that should be taken into account?
- ☒ Choice of horizon: should one look finitely or infinitely far into the future? Keynes' famous quote "In the long run, we are all dead" could be an argument in favor of a finite horizon. Many economic models involve just two periods as an abstraction of "now" and "the future". On the other hand, many decisions have no clearly defined final period: you — or in an evolutionary sense as in overlapping generations models, your genes — may live to see another day. In such cases, an infinite horizon makes sense.
- ☒ Choice of time as a discrete or continuous variable: also here, common sense, the appropriate level of abstraction, and (not rarely) the modeler's choice of mathematical tools is decisive.

In this course, we concentrate on deterministic problems in discrete time. We start with a finite horizon and finish the course with lectures on infinite-horizon problems.

20.2 Problem formulation

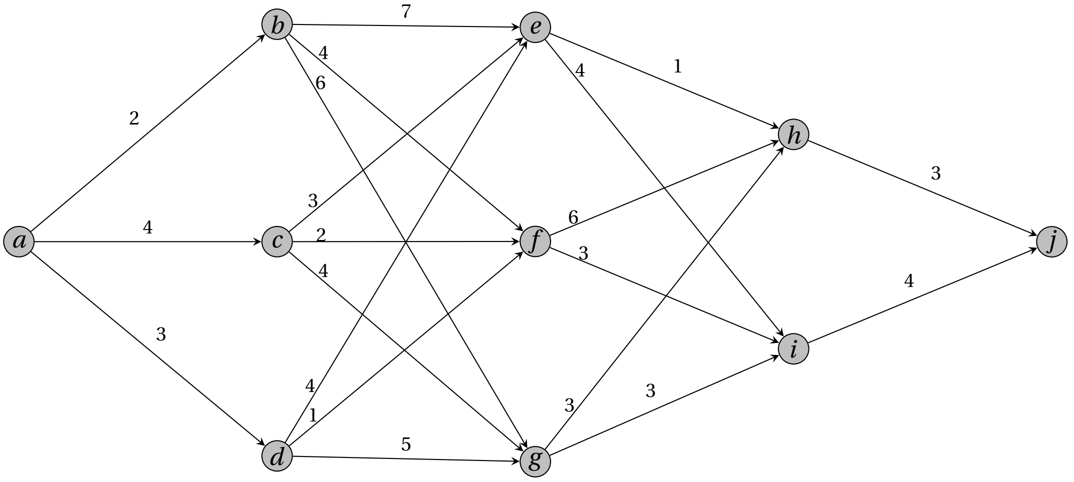
Let's start with two well-known examples from the dynamic optimization literature to motivate the different ingredients of our general problem formulation:

Example 20.1 (Shortest/longest path problems) You want to travel from city a to city j in Figure 1. An arrow from, say, a to b indicates that there is a road from a to b. Its label 2 indicates the length of that road. What are the longest/shortest paths from a to j?

Example 20.2 (Knapsack problems) Suppose you have an empty knapsack with weight capacity $W \in \mathbb{N}$. There are $n \in \mathbb{N}$ items you can pack; each item $i = 1, \ldots, n$ has a weight $w(i) \in \mathbb{N}$ and a value $v(i) \in \mathbb{N}$. Which items should you pack if you want to maximize value subject to the weight constraint? This doesn't sound like a dynamic optimization problem, but can fruitfully be modeled as such. For instance: in the first stage, decide whether or not to pack item 1, in the second, decide whether or not to pack item 2, and so on. ◁

Here are the ingredients that we have to take into account:

1. The problem has a number of stages, referred to as the horizon.
2. At the beginning of each stage, the decision to be taken is characterized by a state variable; in route planning, it described your current location and in the knapsack problem, it describes how much capacity you have left.
3. The states at the different time periods determine the feasible choices you can make; these choices are referred to as controls. In route planning, if you are at city c after the first decision, your controls are ("move to") e, f, and g. In the knapsack problem, if you have come to item i, you have at most two controls: pack item i (if there is enough capacity left) or don't pack item i.



◁

Figure 1: A road map

4. At each stage, your decision leads to some payoff. In route planning, this is the relevant distance. In the knapsack problem, you add value to the knapsack if you pack an item.
5. But it also leads to future constraints: the roads that are feasible at the next intersection or the remaining capacity of your knapsack.
6. The overall outcome is evaluated by the aggregate payoffs accumulated over the various stages: the total length of the route from a to j and the total value of the items in the knapsack.

The formal definition is as follows:

Definition 20.1 In its standard form, a discrete-time dynamic optimization problem (DP) with finite horizon is of the form

$$\textbf{\textit{DP:}} \quad \left\{ \begin{array}{lll} \sup_{u=(u(t))_{t=0}^{T}} & \sum_{t=0}^{T} f(t, x(t), u(t)) & \\ \text{subject to} & u(t) \in U(t, x(t)) & t = 0, \dots, T \\ & x(t+1) = g(t, x(t), u(t)) & t = 0, \dots, T-1 \\ & x(0) \text{ given} & \end{array} \right. \tag{117}$$

Here, $T \in \mathbb{N}$ is the ***horizon*** of the problem. At each time $t = 0, \dots, T$, there is

- ⊠ a variable $u(t)$ referred to as the ***control*** at time t. It lies in a set U referred to as the ***control space***. The controls $u(0), \dots, u(T)$ are the choice variables, i.e., the variables that the decision maker can choose to make the value of the goal function as large as possible;
- ⊠ a variable $x(t)$ referred to as the ***state*** at time t. It lies in a set X referred to as the ***state space***. Initial state $x(0)$ is assumed to be given.

These variables are related:

- ⊠ control $u(t)$ is constrained to be an element of a nonempty ***control region*** $U(t, x(t)) \subseteq U$ that may depend on the time t and state $x(t)$;

☒ having chosen control $u(t)$, next period's state $x(t+1)$ is determined by a given function $g : \{0,\dots,T-1\} \times X \times U \to X$ that may depend on the current time t, state $x(t)$, and control $u(t)$. The equation $x(t+1) = g(t, x(t), u(t))$ is referred to as the ***system equation*** or ***plant equation***.

The triplet $(t, x(t), u(t))$ gives an ***instantaneous value*** (payoff/utility) $f(t, x(t), u(t)) \in \mathbb{R}$ determined by a given function $f : \{0,\dots,T\} \times X \times U \to \mathbb{R}$. The goal is to make the ***aggregate value*** $\sum_{t=0}^{T} f(t, x(t), u(t))$ as large as possible.

Since we start counting time at $t = 0$, there are $T+1$ decisions to be taken. Of course, starting at time $t = 0$ is just a convention. It makes the formulas a bit easier later on when we do discounted dynamic programming. In some problems, on the other hand, it seems more convenient to say that the initial period is $t = 1$ — think of the knapsack problem where you decide whether to pack items 1 to n at different stages. Feel free to do so: all definitions remain sensible with the obvious adjustments. In principle, I could have indicated the initial time by some parameter t_0 of the problem. But notation is complicated enough already.

There is some confusion in the literature regarding the question over what feasible set the decision maker facing the problem (117) tries to optimize the goal function. There are essentially three options, which we discuss in turn, before defining them formally. Since controls affect states via the system equation and states affect which controls are feasible, you typically cannot choose $u = (u(0),\dots,u(T))$ and $x = (x(0),\dots,x(T))$ separately. Therefore, the first candidate consists of pairs (x, u). Such a pair is feasible if the constraints of the maximization problem are satisfied and optimal if you can't find a better feasible one.

On the other hand, this suggests more freedom than there really is. Already in the description of the decision problem in Definition 20.1 it was indicated that the decision maker chose the controls $u(t)$, not the states $x(t)$: these follow mechanically from the system equation. So the second candidate consists of control sequences $u = (u(0),\dots,u(T))$. Such a control sequence is feasible if there is a state sequence x such that (x, u) is a feasible pair. Indeed, given the initial state x_0 and the control sequence u, the optimization problem can be rewritten entirely in terms of the controls, because the state sequence follows uniquely from the equations $x(0) = x_0$ and $x(t+1) = g(t, x(t), u(t))$. Once you know $x(0)$ and control $u(0)$, you know that $x(1) = g(0, x_0, u(0))$ and — generally — $x(t+1) = g(t, x(t), u(t))$. In other words, the state sequence x can be written as a function of x_0 and u: state $x(t)$ is determined uniquely by x_0 and the previous controls $u(0),\dots,u(t-1)$.

Thirdly, it is common to consider policy functions π that assign to each time t and state x a feasible control $\pi(t, x) \in U(t, x)$. Comparing this with the other two approaches, which choose only $T+1$ controls, this seems to be overdoing things: there are typically a lot of states that are never even reached due to the earlier choices (for dramatic effect, choices like committing suicide or loosing your virginity at time t rule out future states where you are alive/virgin at later times), so why would you want a policy function that describes the choice behavior in states that are anyway incompatible with earlier choices? Well, one of the main reasons is that dynamic problems are often solved backwards or by comparing the outcomes of different controls by asking hypothetical questions like "what would you do if you ever were to reach state...?"

These three options are formally defined as follows:

Definition 20.2 Consider the problem (117).

☒ PAIRS (x, u): A pair $(x, u) = (x(t), u(t))_{t=0}^{T}$ is an ***admissible/feasible pair*** if it satisfies the restrictions of the optimization problem:

  ☒ $x(0) = x_0$;

  ☒ $x(t+1) = g(t, x(t), u(t))$ for all $t = 0,\dots,T-1$;

  ☒ $u(t) \in U(t, x(t))$ for all $t = 0,\dots,T$.

The pair (x,u) is ***optimal*** (and u an ***optimal control***) if it is feasible and there is no other feasible pair with a higher value of the goal function.

☒ CONTROLS u: A ***control (sequence)*** $u = (u(0),\dots,u(T))$ is ***admissible/feasible*** if there is a state sequence $x = (x(0),\dots,x(T))$ such that (x,u) is admissible.

If such a state sequence exists, it is unique and determined inductively as follows: $x(0) = x_0$ and for each $t = 0,\dots,T-1$: $x(t+1) = g(t,x(t),u(t))$. In particular, $x(t+1)$ can be written as a function of $x_0, u(0),\dots,u(t)$.

Control u is ***optimal*** if the corresponding pair (x,u) is optimal.

☒ POLICIES π: A policy π is a function π that assigns to each time t and state x a feasible control $\pi(t,x) \in U(t,x)$. The set of policies is denoted $\Pi = \{\pi : \{0,\dots,T\} \times X \to U \mid \pi(t,x) \in U(t,x) \text{ for all } (t,x)\}$.

Given initial state x_0, policy π gives rise to an admissible pair (x^π, u^π) with

$$\begin{aligned} x^\pi(0) &= x_0, \\ x^\pi(t+1) &= g(t,x^\pi(t),u^\pi(t)), \quad t = 0,\dots,T-1, \\ u^\pi(t) &= \pi(t,x^\pi(t)), \quad t = 0,\dots,T, \end{aligned}$$

and consequently to payoff $\sum_{t=0}^{T} f(t,x^\pi(t),u^\pi(t))$. Policy π is ***optimal*** if there is no other policy with a higher payoff.

But wait, if you optimize a goal function over three different domains, can't you get three different solutions? In this case, no! Optimization over admissible pairs (x,u) and controls u clearly give the same outcome. Moreover, every policy π gives rise to an admissible pair (x,u) by doing, at each moment in time, whatever π does. And conversely, every admissible pair (x,u) gives rise to a policy π defined as follows: for each $t = 0,\dots,T$, define $\pi(t,x(t)) = u(t)$. For other (t,x), define $\pi(t,x) \in U(t,x)$ arbitrarily.

Remark 20.1 There is a bit of a catch, though: this final step works perfectly in this simple class of models, but choosing $\pi(t,x)$ arbitrarily for pairs (t,x) that are anyway never reached does not work in more intricate models, where policy functions may need to satisfy additional restrictions like continuity and integrability. Fortunately, this need not concern us here. ◁

This leads to the important insight that the optimal value of the goal function is the same, regardless of whether optimization is over pairs (x,u), controls u, or policies π! One can choose the most convenient of the three and each of them will have a role to play in at least one of the methods we discuss for solving dynamic optimization problems.

There is a number of special cases that occur frequently in the economics literature:

Example 20.3 (Scrap values) Some authors (like Dimitri P. Bertsekas in his popular two-volume book *Dynamic Programming and Optimal Control*, 3rd edition, 2005 (Vol. 1) and 2007 (Vol. 2), Athena Scientific) specify the objective function as $\sum_{t=0}^{T-1} f(t,x(t),u(t)) + S(x(T))$, where $S(x(T))$ can be interpreted as the value of anything that remains — the scrap value — if the system is in state $x(T)$ in the final period. This is a special case of (117) with $f(T,x(T),u(T)) = S(x(T))$. ◁

Example 20.4 (Problems without separated controls and states) Some authors (like Nancy L. Stokey and Robert E. Lucas in *Recursive Methods in Economic Dynamics*, Harvard University Press, 1989, and Daron Acemoglu in *Introduction to Modern Economic Growth*, Princeton University Press, 2009) study problems without an explicit separation between control and state variables:

$$\left\{\begin{array}{lll} \sup_{x=(x(t))_{t=0}^T} & \sum_{t=0}^T F(t,x(t),x(t+1)) & \\ \text{subject to} & x(t+1)\in U(t,x(t)) & t=0,\dots,T \\ & x(0) \text{ given.} & \end{array}\right. \tag{118}$$

This is a special case of (117) if we introduce the control variable $u(t)$ via the system equation $u(t) = x(t+1)$ and instantaneous payoffs $f(t,x(t),u(t)) = F(t,x(t),x(t+1))$. ◁

Example 20.5 (Discounted stationary utility) By far the most common application of dynamic optimization in macroeconomic theory involves optimization of an exponentially discounted, but otherwise time-independent utility function. If also the feasible control region and the plant equation do not explicitly depend on time, the problem is called ***stationary*** and can be denoted as:

$$\left\{\begin{array}{lll} \sup_{u=(u(t))_{t=0}^T} & \sum_{t=0}^T \beta^t f(x(t),u(t)) & \\ \text{subject to} & u(t)\in U'(x(t)) & t=0,\dots,T \\ & x(t+1)=g'(x(t),u(t)) & t=0,\dots,T-1 \\ & x(0) \text{ given,} & \end{array}\right. \tag{119}$$

where $\beta\in(0,1)$ is a discount factor. This is a special case of (117) with $f(t,x(t),u(t)) = \beta^t U(x(t),u(t))$, $U(t,x(t)) = U'(x(t))$, and $g(t,x(t),u(t)) = g'(x(t),u(t))$. ◁

Before we actually start solving problems, let's take one more look at the knapsack problem. My purpose is to point out that there may be many mathematical formalizations of the same problem. Below, I give both a static and a dynamic model of the knapsack problem. But there are more, depending on how you choose the 'ingredients' like the control and state space in different stages. Some formalizations may be easier to solve than others and the modeling phase is as much of an art as actually solving it. Exercises 20.3 and 20.4 provide two alternative dynamic programming approaches to the knapsack problem.

Example 20.6 (Two formalizations of the knapsack problem) One can formulate the knapsack problem as a static optimization problem. Introduce decision variables $u(1),\dots,u(n)\in\{0,1\}$, where $u(i)=0$ means that you do *not* pack item i and $u(i)=1$ means that you do. The problem becomes:

$$\left\{\begin{array}{lll} \text{maximize} & \sum_{i=1}^n u(i)v(i) & \\ \text{subject to} & u(i)\in\{0,1\} \quad \text{for all } i=1,\dots,n, & \text{[pack } i \text{ or not]} \\ & \sum_{i=1}^n u(i)w(i)\le W. & \text{[packed weight cannot exceed capacity]} \end{array}\right.$$

Next, we formulate it as a dynamic programming problem. Decide in n periods: in period i you decide to pack item i or not. This is your control. What would be a good state variable? One candidate would be the remaining capacity. This leads to:

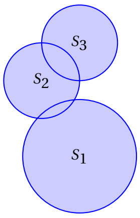
$$\left\{\begin{array}{lll} \text{maximize} & \sum_{i=1}^n u(i)v(i) & \\ \text{subject to} & x(1)=W & \\ & u(i)\in U(i,x(i)) = \left\{\begin{array}{ll} \{0,1\} & \text{if } w(i)\le x(i), \\ \{0\} & \text{otherwise.}\end{array}\right. & i=1,\dots,n \\ & x(i+1)=x(i)-u(i)w(i) & i=1,\dots,n-1 \end{array}\right.$$

The definition of the control region $U(i,x(i))$ indicates that you can pack item i only if the remaining capacity $x(i)$ is sufficient. Otherwise, you are left with only one feasible control: do not pack item i. The system equation $x(i+1) = x(i) - u(i)w(i)$ states that remaining capacity decreases by $w(i)$ if you pack i ($u(i)=1$) and remains unchanged otherwise ($u(i)=0$). ◁

20.3 Divide and conquer

A method that pops up over and over again in dynamic optimization consists in dividing a larger problem into a collection of smaller ones. What you gain is that smaller problems may be easier to solve, what you lose is that there are more of them. The fact that this method is used at all indicates that in dynamic optimization, the gains might outweigh the losses.

Let's illustrate the general principle. Suppose you want to maximize a goal function over a nonempty feasible set S. One way to do this is as follows: divide S into smaller subsets $(S_i)_{i\in I}$. The figure to the right has $S = S_1 \cup S_2 \cup S_3$. Solve the optimization problem over each of the smaller sets S_i separately; let $x_i^* \in S_i$ denote a maximum location. The maximum of the original problem with feasible set $S = \cup_{i\in I} S_i$ is obtained by simply taking the maximum over the optima x_i^* in the smaller problems.



To get the subtleties right: the maximum need not exist and we may need to replace maxima by suprema, but the idea remains the same:

Theorem 20.1 (Divide and conquer)

Suppose a nonempty feasible set can be written $S = \cup_{i\in I} S_i$ and let $f : S \to \mathbb{R}$ be bounded above. Then

$$\sup_{s\in S} f(s) = \sup_{i\in I} \sup_{s\in S_i} f(s). \tag{120}$$

Proof: Let $s \in S = \cup_{i\in I} S_i$. Then $s \in S_i$ for some $i \in I$. Hence,

$$f(s) \le \sup_{s\in S_i} f(s) \le \sup_{i\in I} \sup_{s\in S_i} f(s).$$

Since this holds for each $s \in S$, the right-hand side is an upper bound of f:

$$\sup_{s\in S} f(s) \le \sup_{i\in I} \sup_{s\in S_i} f(s). \tag{121}$$

Conversely, for each $i \in I$: $S_i \subseteq S$, so

$$\sup_{s\in S} f(s) \ge \sup_{s\in S_i} f(s).$$

Since this holds for each $i \in I$:

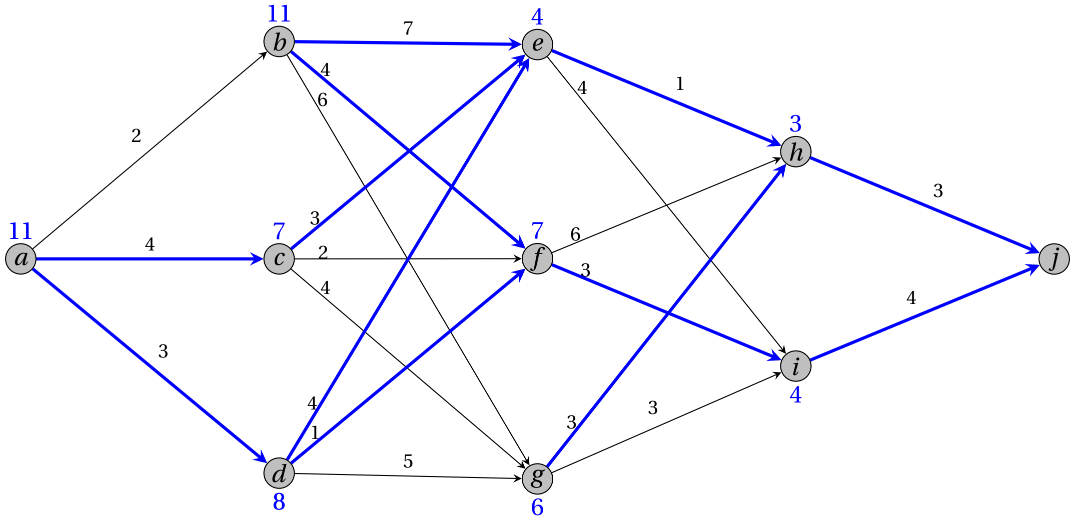
$$\sup_{s\in S} f(s) \ge \sup_{i\in I} \sup_{s\in S_i} f(s). \tag{122}$$

Inequalities (121) and (122) together prove (120). □

Of course, a similar result holds for minimization problems with infima instead of suprema.

How can the method described in Theorem 20.1 help us to solve the dynamic programming problem (117)? Solving the problem from front to back (present to future) starting at time $t = 0$, is difficult: consequences of control decisions on future states, feasible controls, and payoffs can be hard to foresee. So why not solve the problem backwards in time? We start with an example:

Example 20.7 In Figure 1, let us find the shortest path(s) from a to j. After 3 periods, we've arrived at $x(3) = h$ or $x(3) = i$ and the problem is really easy: there is only one feasible control. At h, you choose (to move to) j with corresponding length $J_3(h) = 3$; in the figure below, this optimal control is indicated with a blue line and the value $J_3(h) = 3$ with a blue label above h. At i, you choose j with corresponding distance $J_3(i) = 4$. In general, we will use $J_s(x)$ to denote the optimal value of the problem that starts at time s in state x; since this studies the problem from a certain point onward, it is often referred to as a tail problem.



Let’s move back one step. After 2 periods, we’ve arrived at $x(2) \in \{e, f, g\}$.

☒ at $x(2) = e$, choosing h leads to a trip of length $1 + J_3(h) = 1 + 3 = 4$, choosing i leads to a trip of length $4 + J_3(i) = 4 + 4 = 8$, so the shortest path from e to j is $J_2(e) = \min\{1 + J_3(h), 4 + J_3(i)\} = \{1+3, 4+4\} = 4$, with corresponding optimal control $u(2) = h$.

☒ at $x(2) = f$, choosing h leads to a trip of length $6 + J_3(h) = 6 + 3 = 9$, choosing i leads to a trip of length $3 + J_3(i) = 3 + 4 = 7$, so the shortest path from f to j has length $J_2(f) = \min\{9, 7\} = 7$, with corresponding optimal control $u(2) = i$.

☒ at $x(2) = g$, choosing h leads to a trip of length $3 + J_3(h) = 3 + 3 = 6$, choosing i leads to a trip of length $3 + J_3(i) = 3 + 4 = 7$, so the shortest path from g to j has length $J_2(g) = \min\{6, 7\} = 6$, with corresponding optimal control $u(2) = h$.

Again, these fact are indicated in the figure. We move back one step more, to $x(1) \in \{b, c, d\}$.

☒ at $x(1) = b$, the optimal control solves $J_1(b) = \min\{7 + J_2(e), 4 + J_2(f), 6 + J_2(g)\} = \min\{11, 11, 12\} = 11$, with two corresponding optimal controls, $u(1) \in \{e, f\}$.

☒ at $x(1) = c$, the optimal control solves $J_1(c) = \min\{3 + J_2(e), 2 + J_2(f), 4 + J_2(g)\} = \min\{7, 9, 10\} = 7$, with corresponding optimal control $u(1) = e$.

☒ at $x(1) = d$, the optimal control solves $J_1(d) = \min\{4 + J_2(e), 1 + J_2(f), 5 + J_2(g)\} = \min\{8, 8, 11\} = 8$, with two corresponding optimal controls, $u(1) \in \{e, f\}$.

We move back one step more, to $x(0) = a$. At $x(0) = a$, the optimal control solves $J_0(a) = \min\{2 + J_1(b), 4 + J_1(c), 3 + J_1(d)\} = \min\{13, 11, 11\} = 11$, with two corresponding optimal controls $u(0) \in \{c, d\}$.

We now found the length of the shortest path from a to j, namely 11, and the routes that achieve this, by starting at $x(0) = a$ and choosing an optimal control at each consecutive state. This leads to three shortest paths:

$$a \to c \to e \to h \to j$$
$$a \to d \to e \to h \to j$$
$$a \to d \to f \to i \to j. \qquad \lhd$$

The general idea behind the ***dynamic programming algorithm*** is to obtain the optimal solution to longer problems from those of shorter problems. If u^* is an optimal control, and along the way it takes you to state x at time t, then the remaining controls — from time t onward — must be optimal in the 'tail problem' starting at time t in state x. As a simplistic example: if the optimal route from Stockholm to Amsterdam takes you through Copenhagen, then it must also optimize the route from Copenhagen to Amsterdam.

This is how you solve the problem (117) backwards. More formally, for time $s \in \{0, \dots, T\}$ and state x, let $J_s(x)$ denote the ***(optimal) value function*** for the tail problem that starts in state x at time s:

$$J_s(x) = \begin{cases} \sup_{u(s),u(s+1),\dots,u(T)} & \sum_{t=s}^{T} f(t, x(t), u(t)) & \\ \text{subject to} & u(t) \in U(t, x(t)) & t = s, \dots, T \\ & x(t+1) = g(t, x(t), u(t)) & t = s, \dots, T-1 \\ & x(s) = x. & \end{cases}$$

Suppose you have come to the final period, $t = T$, and state $x(T) = x$. You don't have to worry about future consequences: the only part of the goal function you can still affect is the term $f(T, x, u)$ by choosing a control $u \in U(T, x)$ that makes it as large as possible. So $J_T(x)$ should satisfy

$$J_T(x) = \sup_{u \in U(T,x)} f(T, x, u). \tag{123}$$

Now let $s \in \{0, \dots, T-1\}$ and suppose you already figured out what is optimal in tail problems starting at time $s+1$. This helps you to decide what is optimal at time s. Suppose you find yourself in state $x(s) = x$ at time s. Your feasible controls are those in $U(s, x)$. So choosing $u \in U(s, x)$ leads to instantaneous payoff $f(s, x, u)$ and a next state $x(s+1) = g(s, x, u)$. But by assumption, you know that the optimal value from the next state $x(s+1)$ onward is $J_{s+1}(x(s+1)) = J_{s+1}(g(s, x, u))$. So the best thing you can do is to optimize the sum of these two expressions: the value functions at time s and $s+1$ are related via the equation

$$J_s(x) = \sup_{u \in U(s,x)} \big(f(s, x, u) + J_{s+1}(g(s, x, u)) \big).$$

Formally:

Theorem 20.2 (The dynamic programming algorithm)

The optimal value function satisfies:

$$\begin{aligned} J_T(x) &= \sup_{u \in U(T,x)} f(T, x, u), \\ J_s(x) &= \sup_{u \in U(s,x)} \big(f(s, x, u) + J_{s+1}(g(s, x, u)) \big) \qquad s = 0, \dots, T-1. \end{aligned}$$

In particular, $J_0(x(0))$ is the optimal value of the original problem (117).
Moreover, if $u = (u(t))_{t=0}^{T}$ and the corresponding states $x = (x(t))_{t=0}^{T}$ satisfy

$$J_T(x(T)) = f(T, x(T), u(T)), \tag{124}$$

$$J_s(x(s)) = f(s, x(s), u(s)) + J_{s+1}(g(s, x(s), u(s))), \qquad s = 0, \dots, T-1, \tag{125}$$

then u is an optimal control.

Proof: For time s and state x, let $F(s, x)$ denote the feasible control sequences for the tail problem starting at time s in state x:

$$F(s, x) = \left\{ (u(s), \dots, u(T)) : \begin{array}{ll} u(t) \in U(t, x(t)) & t = s, \dots, T \\ x(t+1) = g(t, x(t), u(t)) & t = s, \dots, T-1 \\ x(s) = s & \end{array} \right\}.$$

The first part of the theorem follows by partitioning the set of feasible controls $F(s,x(s))$ into components that differ in their choice at the current period s,

$$\cup_{u(s)\in U(s,x(s))}\{(u(s),u(s+1),\dots,u(T)):\quad (u(s+1),\dots,u(T))\in F(s+1,g(s,x(s),u(s)))\}$$

and applying Theorem 20.1. For the second part, note:

$$\begin{aligned}
J_0(x(0)) &\overset{(125)}{=} f(0,x(0),u(0))+J_1(x(1))\\
&\overset{(125)}{=} f(0,x(0),u(0))+f(1,x(1),u(1))+J_2(x(2))\\
&= \cdots\\
&\overset{(125)}{=} f(0,x(0),u(0))+f(1,x(1),u(1))+\cdots+f(T-1,x(T-1),u(T-1))+J_T(x(T))\\
&\overset{(124)}{=} f(0,x(0),u(0))+f(1,x(1),u(1))+\cdots+f(T-1,x(T-1),u(T-1))+f(T,x(T),u(T)).
\end{aligned}$$

Since feasible pair (x,u) generates the optimal payoff $J_0(x(0))$, it is optimal. □

Notice the generality of the dynamic programming algorithm: so far, we have made no assumptions whatsoever about the structure of the state space and control regions and we have seen examples where choices can be road segments or items to put in a knapsack. We finish this section by applying the dynamic programming algorithm to the knapsack problem from Example 20.2.

Example 20.8 Define $J_i(x)$ as the optimal solution of the knapsack problem with items $i,i+1,\dots,n$ and capacity x. The dynamic programming algorithm gives us the following relations:

$$J_n(x) \quad = \quad \begin{cases} v(n) & \text{if } w(n)\le x,\\ 0 & \text{otherwise.}\end{cases}$$

and for $i=1,\dots,n-1$:

$$J_i(x) \quad = \quad \begin{cases} \max\{J_{i+1}(x), v(i)+J_{i+1}(x-w(i))\} & \text{if } w(i)\le x,\\ J_{i+1}(x) & \text{otherwise.}\end{cases}$$

Deciding whether to pack the final item is simple: do if capacity allows it. And from then on, the recursive relation kicks in. We use it to solve the knapsack problem with capacity 10 and

item	1	2	3	4	5	6	7
weight	3	5	4	1	4	3	1
value	60	60	40	10	16	9	3

The corresponding table of optimal value functions $J_i(x)$ is:

	10	**133**	**110**	**69**	**38**	**28**	**12**	**3**
	9	**130**	**100**	**66**	**38**	**28**	**12**	**3**
	8	**120**	**73**	**59**	**35**	**28**	**12**	**3**
	7	**100**	**73**	**53**	**29**	**25**	**12**	**3**
	6	**73**	**70**	**53**	**29**	**19**	**12**	**3**
state x	5	**73**	**60**	**50**	**26**	**19**	**12**	**3**
	4	**70**	40	**40**	**19**	**16**	**12**	**3**
	3	**60**	13	13	**13**	9	**9**	**3**
	2	13	13	13	**13**	3	3	**3**
	1	10	10	10	**10**	3	3	**3**
	0	0	0	0	0	0	0	0
		1	2	3	4	5	6	7
					stage i			

Bold-faced numbers are associated with the decision to pack the item in the corresponding column. It follows by considering $J_1(10)$, the optimal value from starting at initial time 1 with capacity 10, that the optimal value is 133, obtained by packing items 1,2,4, and 7. To see the latter, note that $J_1(10) = \mathbf{133}$ is bold, so pack item 1 with weight 3. This leads to $J_2(10-3) = J_2(7) = \mathbf{73}$, which is bold, so pack item 2 with weight 5, leading to $J_3(7-5) = J_3(2) = 13$, etc.

We solve backwards, so we started with the 7-th column. It tells us to pack item 7 with value 3 as long as remaining capacity x suffices. Next, we move to the 6-th, where we need to decide, given remaining capacity x, whether to pack item 6 or not. Its weight is 3, so for capacities $x < 3$, we simply cannot do this and there is only one feasible control: do not pack item 6. Hence, the optimal value $J_6(x)$ is the same as the optimal value $J_7(x)$ for $x < 3$. Things are a bit more exciting for capacities $x \in \{3, \dots, 10\}$. Let's do one of them: what is $J_6(3)$? At remaining capacity $x = 3$, we have two controls:

1. Do not pack item 6: this leads to instantaneous payoff 0 and remaining capacity 3 in the next period, so the dynamic programming algorithm tells us that this leads to payoff $0 + J_7(3) = 3$;

2. Do pack item 6: this leads to instantaneous payoff $v(6) = 9$, but remaining capacity $3-3 = 0$ in the next period, so the dynamic programming algorithm tells us that this leads to payoff $9 + J_7(0) = 9$;

Hence, it is optimal to pack item 6. The table indicates that $J_6(3)$ equals $\max\{J_7(3), v(6) + J_7(0)\} = 9$, and it is printed in bold-face to stress that the optimal decision is to pack item 6. ◁

Exercises section 20

20.1 Find the longest path(s) from a to j in Figure 1.

20.2 (a) Solve the knapsack problem with capacity 7 and

item	1	2	3	4	5
weight	5	3	2	2	1
value	4	7	3	5	4

(b) Solve the knapsack problem with capacity 8 and

item	1	2	3	4	5	6
weight	7	3	1	2	5	1
value	6	1	4	5	6	3

(c) Solve the knapsack problem with capacity 6 and

item	1	2	3	4	5	6
weight	1	3	2	4	5	2
value	4	6	5	8	7	7

20.3 Consider the knapsack problem in Example 20.2. For $i = 1, \dots, n$ and $x = 0, \dots, W$, let $J_i(x)$ be the optimal value of the knapsack problem with items $\{1, \dots, i\}$ and capacity x. We are interested in finding $J_n(W)$, but the problem is easiest to solve for small i, i.e., front to back.

(a) Show that the value functions satisfy the following recurrence relation:

$$\text{for all } x = 0, \dots, W: \qquad J_1(x) = \begin{cases} 0 & \text{if } w(1) > x, \\ v(i) & \text{if } w(1) \le x, \end{cases}$$

and for all $i = 2, \dots, n$ and $x = 0, \dots, W$:

$$J_i(x) = \begin{cases} J_{i-1}(x) & \text{if } w(i) > x, \\ \max\{v(i) + J_{i-1}(x - w(i)), J_{i-1}(x)\} & \text{if } w(i) \le x. \end{cases}$$

(b) Use the dynamic programming algorithm in (a) to solve the instance of the knapsack problem at the end of Example 20.8. [Hint: make a table similar to the one there, but start with entries $J_1(x)$.]

20.4 Consider the knapsack problem in Example 20.2. Define $V = \sum_{i=1}^{n} v(i)$. For $i = 1, \dots, n$ and $x = 0, \dots, V$, let $J_i(x)$ be the minimal total weight of a subset $S \subseteq \{1, \dots, i\}$ of items for which $\sum_{j \in S} v(j) = x$, i.e., the minimal weight of a knapsack packed from items $\{1, \dots, i\}$ with value x. Take $J_i(x) = \infty$ if such a set does not exist.

(a) How can the optimal value of the knapsack problem be derived from $J_n(0), J_n(1), \dots, J_n(V)$?

(b) Show that the value functions satisfy the following recurrence relation:

$$\text{for all } x = 0, \dots, V: \qquad J_1(x) = \begin{cases} 0 & \text{if } x = 0 \text{ (take } S = \emptyset\text{)}, \\ w(1) & \text{if } x = v(1) \text{ (take } S = \{1\}\text{)}, \\ \infty & \text{otherwise,} \end{cases}$$

and for all $i = 2, \dots, n$ and $x = 0, \dots, V$:

$$J_i(x) = \begin{cases} J_{i-1}(x) & \text{if } v(i) > x, \\ \min\{w(i) + J_{i-1}(x - v(i)), J_{i-1}(x)\} & \text{if } v(i) \leq x. \end{cases}$$

(c) I won't ask you to solve Example 20.8 using this algorithm; if you were to make a table of $J_i(x)$ with $i = 1, \dots, n$ and $x = 0, \dots, V$, how large would it be?

20.5 Given horizon $T \in \mathbb{N}$ and initial state $x_0 \in \mathbb{R}$, use the dynamic programming algorithm to solve:

$$\begin{array}{lll} \text{maximize} & \sum_{t=0}^{T} x(t)^2 (2 - u(t)) & \\ \text{with} & u(t) \in [0, 1] & t = 0, \dots, T \\ & x(t+1) = x(t) u(t) & t = 0, \dots, T-1 \\ & x(0) = x_0 & \end{array}$$

HINT: show that J_t is of the form $J_t(x) = \alpha_t x^2$ for some $\alpha_t \in \mathbb{R}$ and find a formula for α_t. Or compute J_T, J_{T-1}, J_{T-2}, ... until you figure out the pattern.

20.6 Given horizon $T \in \mathbb{N}$ and initial state $x_0 \in \mathbb{R}$, use the dynamic programming algorithm to solve:

$$\begin{array}{lll} \text{maximize} & \sum_{t=0}^{T} (x(t) - u(t)) & \\ \text{with} & u(t) \in [0, 1] & t = 0, \dots, T \\ & x(t+1) = x(t) + u(t) & t = 0, \dots, T-1 \\ & x(0) = x_0 & \end{array}$$

HINT: show that J_t is of the form $J_t(x) = \alpha_t x + \beta_t$ for some $\alpha_t, \beta_t \in \mathbb{R}$ and express α_t and β_t in terms of α_{t+1} and β_{t+1}. Or compute J_T, J_{T-1}, J_{T-2}, ... until you figure out the pattern.

20.7 Given horizon $T \in \mathbb{N}$ and initial state $x_0 \in \mathbb{R}$, use the dynamic programming algorithm to solve:

$$\begin{array}{lll} \text{maximize} & \sum_{t=0}^{T} (x(t) + \ln u(t)) & \\ \text{with} & u(t) \in (0, 1] & t = 0, \dots, T \\ & x(t+1) = x(t) - u(t) & t = 0, \dots, T-1 \\ & x(0) = x_0. & \end{array}$$

21 The maximum principle

In this section, we impose some additional structure so that we optimize over real variables and the system equation and payoff function are sufficiently differentiable. In that case, we recognize the dynamic programming problem as a special case of the nonlinear programming problem from Section 19 and we can use the tools developed there. This leads to socalled maximum principles. These originate in the Russian school of dynamic optimization. A general version in continuous time was proved by Pontryagin in 1954; an early and very detailed treatment of optimal control of finite-horizon, discrete-time systems originating in this Russian school can be found in Boltjanski's book *Optimale Steuerung Diskreter Systeme* from 1976. The name 'maximum principle' comes from the fact that, under suitable conditions, the problem reduces to solving a sequence of separate maximization problems. We provide two versions of the maximum principle; the first using Theorem 19.1, the second using the Fritz-John conditions.

21.1 A first maximum principle

In this subsection, we consider the problem

$$\begin{array}{lll} \text{maximize} & \sum_{t=0}^{T-1} f(t, x(t), u(t)) & \\ \text{subject to} & u(t) \in U(t) & t = 0, \dots, T-1 \\ & x(t+1) = g(t, x(t), u(t)) & t = 0, \dots, T-2 \\ & x(0) = x_0 \text{ given.} & \end{array} \tag{126}$$

We assume that state vectors lie in $\mathbb{R}^n$ and controls in $\mathbb{R}^k$. Moreover, the control regions $U(t) \subseteq \mathbb{R}^k$ are assumed to be convex and independent of the state at time t.

Our first maximum principle derives necessary conditions for the optimal control by applying Theorem 19.1. In particular, the steps of the proof are:

1. write the problem as an optimization problem involving only the control variables, removing the state variables;

2. find necessary conditions for the optimum by applying Theorem 19.1;

3. simplify these conditions by introducing auxiliary functions/vectors.

Working out the details is tedious, since the derivative of the goal function is intricate.

STEP 1: Let u be a feasible control sequence. By substitution of the system equation, the state sequence x can be written in terms of the fixed initial state x_0 and the control sequence u:

$$\begin{array}{lllll} x(0) & = & x_0 & & \\ x(1) & = & g(0, x(0), u(0)) & = & g(0, x_0, u(0)), \\ x(2) & = & g(1, x(1), u(1)) & = & g(1, g(0, x_0, u(0)), u(1)), \\ x(3) & = & g(2, x(2), u(2)) & = & g(2, g(1, g(0, x_0, u(0)), u(1)), u(2)), \\ & \vdots & & & \end{array}$$

where the highlighted terms are substituted from the previous line. Define

$$X_0(u) = x_0 \quad \text{and for } t = 0, \dots, T-2: \quad X_{t+1}(u) = g(t, X_t(u), u(t)).$$

It follows that the state sequence associated with u is

$$x(t) = X_t(u), \qquad t = 0, \dots, T-1.$$

Note that $X_t(u)$ depends only on the controls $u(0),\dots,u(t-1)$ before time t, not on $u(t),\dots,u(T-1)$. The optimization problem becomes

$$\text{maximize } J(u) = \sum_{t=0}^{T-1} f(t, X_t(u), u(t)) \quad \text{with } u(t) \in U(t) \text{ for all } t = 0,\dots,T-1. \tag{127}$$

STEP 2: Since the control regions $U(t)$ are convex, Theorem 19.1 says that an optimal control sequence u^* must satisfy

$$\sum_{t=0}^{T-1} \nabla_{u(t)} J(u^*)(u(t) - u^*(t)) \leq 0 \quad \text{for all feasible } u = (u(0),\dots,u(T-1)).$$

Because we can set $u(t) = u^*(t)$ for all but one t, this can be decomposed into T separate conditions

$$\nabla_{u(t)} J(u^*)(u(t) - u^*(t)) \leq 0 \qquad \text{for all } t = 0,\dots,T-1, \text{ and all } u(t) \in U(t).$$

It remains to find an expression for the partial derivatives $\nabla_{u(t)} J$ of the goal function J with respect to the control $u(t)$ at each time t.

☒ The final control $u(T-1)$ appears only once in the goal function, in $f(T-1, x(T-1), u(T-1))$. Hence,

$$\nabla_{u(T-1)} J(u) = \nabla_u f(T-1, x(T-1), u(T-1)).$$

☒ For $u(t)$ at times $t < T-1$, this is a more intricate matter, as $u(t)$ not only affects the instantaneous payoff $f(t, x(t), u(t))$, but also all future payoffs via its effect on the later states. To make the structure as clear as possible, fix controls $u(s)$ at all other times $s \neq t$. In particular, we know state $x(t)$. Let $f_s : \mathbb{R}^n \to \mathbb{R}$ and $g_s : \mathbb{R}^n \to \mathbb{R}^n$ be abbreviations for $f(s, \cdot, u(s))$ and $g(s, \cdot, u(s))$, respectively. Then the terms of the goal function J involving $u(t)$ can be written as

$$\begin{aligned} f(t, x(t), u(t)) &+ \left(f_{t+1} \circ g(t, x(t), \cdot)\right)(u(t)) \\ &+ \left(f_{t+2} \circ g_{t+1} \circ g(t, x(t), \cdot)\right)(u(t)) \\ &+ \left(f_{t+3} \circ g_{t+2} \circ g_{t+1} \circ g(t, x(t), \cdot)\right)(u(t)) \\ &+ \cdots \\ &+ \left(f_{T-1} \circ g_{T-2} \circ \cdots \circ g_{t+1} \circ g(t, x(t), \cdot)\right)(u(t)). \end{aligned}$$

Computing its derivative using the Chain Rule gives

$$\begin{aligned} \nabla_{u(t)} J(u) = \partial_u f(t, x(t), u(t)) &+ Df_{t+1} \partial_u g(t, x(t), u(t)) \\ &+ Df_{t+2} Dg_{t+1} \partial_u g(t, x(t), u(t)) \\ &+ \cdots \\ &+ Df_{T-1} Dg_{T-2} \cdots Dg_{t+1} \partial_u g(t, x(t), u(t)). \end{aligned}$$

STEP 3: All terms above affecting payoffs at times $s > t$ end with $\partial_u g(t, x(t), u(t))$. Gathering the terms preceding it in an ***adjoint/costate*** vector $p_{t+1} \in \mathbb{R}^n$, the expression becomes

$$\nabla_{u(t)} J(u) = \partial_u f(t, x(t), u(t)) + p_{t+1}^\top \partial_u g(t, x(t), u(t)).$$

With $p_T = \mathbf{0}$, we see that the (adjoint/costate) vectors are related via the backward difference equation

$$p_t^\top = Df_t + p_{t+1}^\top Dg_t \quad \text{with terminal condition } p_T = \mathbf{0}.$$

Substituting the definitions of f_t and g_t, this becomes

$$\underbrace{p_t^\top}_{1\times n} = \underbrace{\nabla_x f(t,x(t),u(t))}_{1\times n} + \underbrace{p_{t+1}^\top}_{1\times n}\underbrace{\nabla_x g(t,x(t),u(t))}_{n\times n} \quad \text{with terminal condition } p_T = \mathbf{0}.$$

Introducing the Hamiltonian

$$H(t,x,u,p) = f(t,x,u) + p^\top g(t,x,u) \qquad \text{for } (t,x,u,p) \in \{0,\dots,T-1\} \times \mathbb{R}^n \times \mathbb{R}^k \times \mathbb{R}^n$$

we find that

$$\nabla_{u(t)} J(u) = \nabla_u H(t,x(t),u(t),p_{t+1}) \text{ and } p_t^\top = \nabla_x H(t,x(t),u(t),p_{t+1}), \quad p_T = \mathbf{0}.$$

This proves:

Theorem 21.1 (Maximum principle I)

Let $H(t,x,u,p) = f(t,x,u) + p^\top g(t,x,u)$ be the Hamiltonian of problem (126). If (x^*,u^*) is optimal, then there are costate vectors $p_t \in \mathbb{R}^n$ such that

$$\nabla_u H(t,x^*(t),u^*(t),p_{t+1})(u(t)-u^*(t)) \le 0 \qquad \text{for all } t = 0,\dots,T-1 \text{ and all } u(t) \in U(t). \tag{128}$$

The costate vectors p_t satisfy

$$p_t^\top = \nabla_x H(t,x^*(t),u^*(t),p_{t+1}) \qquad t = 1,\dots,T-1, \tag{129}$$

with terminal condition $p_T = \mathbf{0}$.

Remark 21.1 If the Hamiltonian is concave with respect to u, then $u^*(t)$ maximizes the Hamiltonian $H(t,x^*(t),\cdot,p_{t+1})$ over $U(t)$. This follows from Theorem 17.11, which says that for all $u(t) \in U(t)$:

$$\begin{aligned} H(t,x^*(t),u(t),p_{t+1}) &\le H(t,x^*(t),u^*(t),p_{t+1}) + \underbrace{\nabla_u H(t,x^*(t),u^*(t),p_{t+1})(u(t)-u^*(t))}_{\le 0} \\ &\le H(t,x^*(t),u^*(t),p_{t+1}). \end{aligned}$$

Under the presumed concavity condition, the dynamic programming problem then reduces to T separate maximization problems of the Hamiltonian, with the link between the different problems provided by the difference equation for the costates. ◁

Remark 21.2 If the optimal control $u^*(t)$ lies in the interior of the control region $U(t)$, it must be that

$$\nabla_u H(t,x^*(t),u^*(t),p_{t+1}) = \mathbf{0},$$

i.e. the partial derivative of the Hamiltonian with respect to the control vector must be zero. To see this, note that by assumption, for every coordinate i and $\varepsilon > 0$ sufficiently close to zero, the two vectors $u(t) = u^*(t) \pm \varepsilon e_i$ belong to the control region. Substituting this into (128) and noticing that this gives rise to expressions with an opposite sign, we find the desired conclusion. ◁

Without the concavity condition in Remark 21.1, optimal controls need *not* maximize the Hamiltonian.

Example 21.1 Consider the problem

$$\text{maximize } \sum_{t=0}^{2}(u(t)^2 - 2x(t)^2) \text{ with } u(t) \in [-1,1], x(0) = 0, x(t+1) = u(t).$$

Substituting the initial state and the system equation into the goal function, we need to maximize

$$(u(0)^2 - 2\cdot 0^2) + (u(1)^2 - 2u(0)^2) + (u(2)^2 - 2u(1)^2) = -u(0)^2 - u(1)^2 + u(2)^2$$

with $u(t) \in [-1,1]$. Choosing $u(0)^2$ and $u(1)^2$ as small and $u(2)^2$ as large as possible, we find two optimal control sequences, $u^* = (u^*(t))_{t=0,1,2} = (0,0,\pm 1)$, both with state sequence $x^* = (0,0,0)$.

Let us verify that the conditions of Maximum Principle I hold. The Hamiltonian is

$$H(t,x,u,p) = f(t,x,u) + pg(t,x,u) = (u^2 - 2x^2) + pu.$$

According to the theorem, there are $p_1, p_2, p_3 \in \mathbb{R}$ with

$$\begin{aligned} p_3 &= 0 \\ p_2 &= \partial_x H(2, x^*(2), u^*(2), p_3) = -4x^*(2) = 0 \\ p_1 &= \partial_x H(1, x^*(1), u^*(1), p_2) = -4x^*(1) = 0, \end{aligned}$$

and

$$\partial_u H(t, x^*(t), u^*(t), p_{t+1})(u - u^*(t)) \leq 0 \quad \text{for all } t = 0,1,2 \text{ and } u \in [-1,1].$$

For $t \in \{0,1\}$, this follows from $\partial_u H(t, x^*(t), u^*(t), p_{t+1}) = 2u^*(t) + p_{t+1} = -4\cdot 0 + 0 = 0$. For $t = 2$, the optimal controls are $u^*(2) = \pm 1$. Let us check the case $u^*(2) = 1$. The case $u^*(2) = -1$ is similar. Then

$$\partial_u H(2, x^*(2), u^*(2), p_3)(u - u^*(2)) = 2u^*(2)(u - u^*(2)) = 2(u-1) \leq 0,$$

since $u - 1 \leq 0$ for all $u \in [-1,1]$.

Note that the optimal controls $u^*(t)$ at $t = 0,1$ are zero and do *not* maximize the Hamiltonian

$$H(t, x^*(t), u(t), p_t) = u(t)^2 - 2x^*(t) + p_{t+1}u(t) = u(t)^2$$

over $u(t) \in [-1,1]$. ◁

Next, we provide sufficient conditions for an optimal solution.

Theorem 21.2

If (x^*, u^*) and costates p_t satisfy the conditions of Maximum Principle I and the Hamiltonian $H(t,x,u,p)$ is concave in (x,u), then (x^*, u^*) is an optimal pair.

Proof: Let (x,u) be a feasible pair. We need to show that

$$\Delta = \sum_{t=0}^{T-1} f(t, x(t), u(t)) - \sum_{t=0}^{T-1} f(t, x^*(t), u^*(t)) \leq 0.$$

By definition of the Hamiltonian, we can write

$$\begin{aligned} \Delta &= \sum_{t=0}^{T-1} \left[H(t, x(t), u(t), p_{t+1}) - p_{t+1}^\top g(t, x(t), u(t))\right] - \sum_{t=0}^{T-1} \left[H(t, x^*(t), u^*(t), p_{t+1}) - p_{t+1}^\top g(t, x^*(t), u^*(t))\right] \\ &= \sum_{t=0}^{T-1} \left[H(t, x(t), u(t), p_{t+1}) - H(t, x^*(t), u^*(t), p_{t+1})\right] - \sum_{t=0}^{T-1} \left[p_{t+1}^\top \left(g(t, x(t), u(t)) - g(t, x^*(t), u^*(t))\right)\right]. \end{aligned}$$

By concavity of the Hamiltonian in (x, u) and Theorem 17.11:

$$\begin{aligned} H(t,x(t),u(t),p_{t+1}) - H(t,x^*(t),u^*(t),p_{t+1}) &\leq \partial_x H(t,x^*(t),u^*(t),p_{t+1})(x(t)-x^*(t)) \\ &\quad + \partial_u H(t,x^*(t),u^*(t),p_{t+1})(u(t)-u^*(t)) \\ &\overset{(128)}{\leq} \partial_x H(t,x^*(t),u^*(t),p_{t+1})(x(t)-x^*(t)). \end{aligned} \tag{130}$$

For $t = 0$, $x(0) = x^*(0) = x_0$ (given), so (130) vanishes. For $t = 1, \ldots, T-1$, condition (129) applies and we can write

$$\Delta \leq \sum_{t=1}^{T-1} p_t^\top \left(x(t) - x^*(t)\right) - \sum_{t=0}^{T-1} p_{t+1}^\top \left(g(t,x(t),u(t)) - g(t,x^*(t),u^*(t))\right)$$

With the system equation and $p_T = \mathbf{0}$, we see that the righthand side of this inequality is zero. □

Example 21.2 Maximize $\sum_{t=0}^{T} \left(x(t) - \frac{1}{2}u(t)^2\right)$ subject to the constraints

$$\begin{aligned} &u(t) \in \mathbb{R} && t = 0, \ldots, T \\ &x(t+1) = x(t) - u(t) && t = 0, \ldots, T-1 \\ &x(0) = x_0 && \text{(given)} \end{aligned}$$

SOLUTION: The Hamiltonian

$$H(t,x,u,p) = f(t,x,u) + pg(t,x,u) = x - \tfrac{1}{2}u^2 + p(x-u),$$

is concave in (x, u), so by Theorem 21.2 the maximum principle gives us the desired solution. Note:

$$\frac{\partial}{\partial u} H(t,x,u,p) = -u - p \qquad \text{and} \qquad \frac{\partial}{\partial x} H(t,x,u,p) = 1 + p.$$

By Maximum Principle I (Thm. 21.1) and Remark 21.2, the following must hold in an optimum:[4]

$$\frac{\partial}{\partial u} H(t,x^*(t),u^*(t),p_{t+1}) = -u^*(t) - p_{t+1} = 0 \qquad t = 0, \ldots, T \tag{131}$$

$$\frac{\partial}{\partial x} H(t,x^*(t),u^*(t),p_{t+1}) = 1 + p_{t+1} = p_t \qquad t = 1, \ldots, T \tag{132}$$

$$p_{T+1} = 0 \tag{133}$$

Since $p_{T+1} = 0$ and, by (132), the costates increase by one for each step backward in time, taking $k = 0, \ldots, T$ steps backward gives $p_{T+1-k} = k$. To find the controls using (131), we want an expression for p_{t+1} at times $t = 0, \ldots, T$. Substituting $k = T - t$ in our formula for the costates, $p_{t+1} = T - t$. So the decision maker's optimal choices, the optimal controls, by (131), are $u^*(t) = -p_{t+1} = t - T$ at each time $t = 0, \ldots, T$.

(Also finding the sequence of states in the optimum is a bit trickier: $x^*(0) = x_0$ is given and $x^*(t+1) = x^*(t) - u^*(t) = x^*(t) - (t - T)$ by the system equation. Writing out $x^*(1), x^*(2), x^*(3), \ldots$, you will probably see the pattern,

$$x^*(t) = x_0 + tT - (1 + 2 + \cdots + (t-1)),$$

which you can formally verify by induction.) ◁

[4] Our horizon is T, whereas it is $T-1$ in Theorem 21.1. This means that all conditions involving the final period go up by one. In particular, the final costate is p_{T+1}, not p_T.

21.2 A second maximum principle

In this subsection, we consider the problem

$$
\begin{aligned}
\text{maximize} \sum_{t=0}^{T-1} f(t, x(t), u(t)) & & \qquad (134)\\
\text{with} \quad x(t+1) = g(t, x(t), u(t)) & \qquad t = 0, \ldots, T-1 & \qquad (135)\\
v_0(x(0)) = \mathbf{0} & & \qquad (136)\\
v_T(x(T)) = \mathbf{0} & & \qquad (137)\\
h_t(u(t)) \leq \mathbf{0} & \qquad t = 0, \ldots, T-1 & \qquad (138)
\end{aligned}
$$

State vectors lie in $\mathbb{R}^n$, controls in $\mathbb{R}^k$. Functions $v_0 : \mathbb{R}^n \to \mathbb{R}^{n_0}$ and $v_T : \mathbb{R}^n \to \mathbb{R}^{n_T}$ impose restrictions on the initial and final state, respectively. The functions $h_t : \mathbb{R}^k \to \mathbb{R}^{k_1}$ restrict the feasible controls at time $t = 0, \ldots, T-1$.

We assume that all functions are continuously differentiable on an open set containing the feasible arguments. This means that to this problem, the Fritz-John conditions from Theorem 19.6 apply. The Lagrangian is

$$
\begin{aligned}
&\mathscr{L}(x(0), \ldots, x(T), u(0), \ldots, u(T-1), p_1, \ldots, p_T, \alpha_0, \alpha_T, \gamma_0, \ldots, \gamma_{T-1}) = \\
&\qquad q_0 \sum_{t=0}^{T-1} f(t, x(t), u(t)) - \sum_{t=0}^{T-1} p_{t+1}^\top (x(t+1) - g(t, x(t), u(t))) \\
&\qquad\qquad - \alpha_0^\top v_0(x(0)) - \alpha_T^\top v_T(x(T)) \\
&\qquad\qquad - \sum_{t=0}^{T-1} \gamma_t^\top h_t(u(t)),
\end{aligned}
$$

where we introduced adjoint/costate variables:

⊠ $q_0 \in \mathbb{R}$ for the goal function,

⊠ $p_1, \ldots, p_T \in \mathbb{R}^n$ for the system equation (135),

⊠ $\alpha_0 \in \mathbb{R}^{n_0}$ and $\alpha_T \in \mathbb{R}^{n_T}$ for the constraints (136) and (137) on the initial and final state, respectively,

⊠ $\gamma_0, \ldots, \gamma_{T-1} \in \mathbb{R}^{k_1}$ for the constraints (138) on the controls.

According to the Fritz-John conditions, if (x^*, u^*) is a local maximum, there are corresponding vectors $q_0, p_1, \ldots, p_T, \alpha_0, \alpha_T, \gamma_0, \ldots, \gamma_{T-1}$, not all zero, and with $q_0 \in \{0, 1\}$ such that:

(a) The partial derivatives of the Lagrangian with respect to the states $x(t)$ and controls $u(t)$ for $t = 0, \ldots, T$ are zero,

(b) $\gamma_0, \ldots, \gamma_{T-1} \geq \mathbf{0}$,

(c) complementary slackness: $\gamma_t^\top h_t(u^*(t)) = 0$ for all $t = 0, \ldots, T$.

Let us compute the partial derivatives with respect to the controls and states:

⊠ First with respect to the controls $u(t)$, $t = 0, \ldots, T-1$:

$$\underbrace{q_0}_{1\times 1}\,\underbrace{\partial_u f(t, x^*(t), u^*(t))}_{1\times k} + \underbrace{p_{t+1}^\top}_{1\times n}\underbrace{\partial_u g(t, x^*(t), u^*(t))}_{n\times k} - \underbrace{\gamma_t^\top}_{1\times k_1}\,\underbrace{Dh_t(u^*(t))}_{k_1\times k} = \underbrace{\mathbf{0}^\top}_{1\times k} \qquad (139)$$

☒ Now with respect to the initial state $x(0)$:

$$\underbrace{q_0}_{1\times 1}\ \underbrace{\partial_x f(0,x^*(0),u^*(0))}_{1\times n} + \underbrace{p_1^\top}_{1\times n}\ \underbrace{\partial_x g(0,x^*(0),u^*(0))}_{n\times n} - \underbrace{\alpha_0^\top}_{1\times n_0}\ \underbrace{Dv_0(x^*(0))}_{n_0\times n} = \underbrace{\mathbf{0}^\top}_{1\times n} \tag{140}$$

☒ Then with respect to states $x(t), t = 1,\ldots,T-1$:

$$\underbrace{q_0}_{1\times 1}\ \underbrace{\partial_x f(t,x^*(t),u^*(t))}_{1\times n} - \underbrace{p_t^\top}_{1\times n} + \underbrace{p_{t+1}^\top}_{1\times n}\underbrace{\partial_x g(t,x^*(t),u^*(t))}_{n\times n} = \underbrace{\mathbf{0}^\top}_{1\times n} \tag{141}$$

☒ And finally, with respect to the terminal state $x(T)$:

$$-\underbrace{p_T^\top}_{1\times n} - \underbrace{\alpha_T^\top}_{1\times n_T}\ \underbrace{Dv_T(x^*(T))}_{n_T\times n} = \underbrace{\mathbf{0}^\top}_{1\times n} \tag{142}$$

If we introduce the Hamiltonian H from $\mathbb{R}^{k+2n+1}$ to $\mathbb{R}$ by

$$H(t,x,u,p) = q_0 f(t,x,u) + p^\top g(t,x,u), \qquad t = 0,\ldots,T-1,$$

and

$$p_0 = Dv_0(x^*(0))^\top \alpha_0,$$

the Fritz John conditions reduce to the ones stated in our second Maximum Principle:

Theorem 21.3 (Maximum Principle II)

If (x^*, u^*) is optimal in the problem (134) – (138), there are vectors $p_0,\ldots,p_T \in \mathbb{R}^n, \alpha_0 \in \mathbb{R}^{n_0}, \alpha_T \in \mathbb{R}^{n_T}, \gamma_0,\ldots,\gamma_{T-1} \in \mathbb{R}^{k_1}$, and $q_0 \in \{0,1\}$ such that

$$\gamma_t \geq \mathbf{0} \quad \text{and} \quad \gamma_t^\top h_t(u^*(t)) = 0 \qquad t = 0,\ldots,T-1. \tag{143}$$

The control variables satisfy

$$\partial_u H(t,x^*(t),u^*(t),p_{t+1}) = \gamma_t^\top Dh_t(u^*(t)) \qquad t = 0,\ldots,T-1. \tag{144}$$

The costate variables p_t satisfy

$$\partial_x H(t,x^*(t),u^*(t),p_{t+1}) = p_t^\top \qquad t = 0,\ldots,T-1, \tag{145}$$

with initial and terminal conditions

$$p_0^\top = \alpha_0^\top Dv_0(x^*(0)) \tag{146}$$
$$p_T^\top = -\alpha_T^\top Dv_T(x^*(T)). \tag{147}$$

There are a few things to take notice of here. Firstly, in the Hamiltonian, the costate is p_{t+1}, but the other time indices are t. Secondly, if there are no control restrictions, h_t can be taken to be a constant function and the derivative of the Hamiltonian with respect to the controls has to be zero.

The Fritz John conditions allow numerous other restrictions on the problem's variables, with obvious modifications to the results.

21.3 The linear quadratic (LQ) optimal control problem

The linear quadratic (LQ) optimal control problem has a linear system equation and quadratic instantaneous payoffs. A fairly general formulation of the problem is:

$$\begin{aligned} \text{maximize} \quad & -\frac{1}{2}\sum_{t=0}^{T-1}\left[x(t)^\top Q(t)x(t) + u(t)^\top R(t)u(t)\right] - \frac{1}{2}x(T)^\top Q(T)x(T) \\ \text{with} \quad & x(t+1) = A(t)x(t) + B(t)u(t) \qquad t = 0,\dots,T-1 \\ & x(0) = x_0 \qquad \text{given} \end{aligned} \tag{148}$$

State vectors lie in $\mathbb{R}^n$, controls in $\mathbb{R}^k$. The $n \times n$ matrices $Q(t)$ and $k \times k$ matrices $R(t)$ are symmetric and positive semidefinite; matrices $A(t)$ are $n \times n$, matrices $B(t)$ are $n \times k$.

The LQ control problem is often formulated as a minimization problem; I put a minus sign in front of the goal function to make it a maximization problem. Factor $\frac{1}{2}$ simplifies some derivatives.

We solve the problem using the Fritz John conditions. The constraints are linear, so Theorem 19.8 allows us to assign weight 1 to the goal function. With multiplier $p_t \in \mathbb{R}^n$ for the system equation defining $x(t)$, the Lagrangian is

$$\begin{aligned} \mathscr{L}(\cdot) = & -\frac{1}{2}\sum_{t=0}^{T-1}\left[x(t)^\top Q(t)x(t) + u(t)^\top R(t)u(t)\right] - \frac{1}{2}x(T)^\top Q(T)x(T) \\ & - \sum_{t=0}^{T-1} p_{t+1}^\top \left(x(t+1) - A(t)x(t) - B(t)u(t)\right) - p_0^\top \left(x(0) - x_0\right) \end{aligned}$$

Recall that a quadratic form $y \mapsto y^\top M y$ has derivative $y \mapsto y^\top (M + M^\top)$, which reduces to $y \mapsto 2y^\top M$ if M is symmetric. In an optimum, the FJ conditions assure that there are costate vectors p_t such that the derivatives of $\mathscr{L}$ with respect to all states $x(t)$ and all controls $u(t)$ are zero. For the states $x(t)$ this gives

$$-x(t)^\top Q(t) + p_{t+1}^\top A(t) - p_t^\top = \mathbf{0}^\top \qquad t = 0,\dots,T-1 \tag{149}$$

$$-x(T)^\top Q(T) \qquad - p_T^\top = \mathbf{0}^\top \tag{150}$$

Taking transposes and rearranging terms, this becomes

$$p_t = -Q(t)x(t) + A(t)^\top p_{t+1} \qquad t = 0,\dots,T-1 \tag{151}$$

$$p_T = -Q(T)x(T) \tag{152}$$

Similarly, the derivatives w.r.t. $u(t)$ give

$$-R(t)u(t) + B(t)^\top p_{t+1} = \mathbf{0} \qquad t = 0,\dots,T-1 \tag{153}$$

From (151) and (152), a reasonable guess is that p_t is linear in $x(t)$, i.e., that for some matrix $P(t)$:

$$p_t = -P(t)x(t) \qquad t = 0,\dots,T. \tag{154}$$

Substituting conditions (154) and the system equation in (153) and solving for $u(t)$ gives

$$u(t) = -\left[R(t) + B(t)^\top P(t+1)B(t)\right]^{-1} B(t)^\top P(t+1)A(t)x(t) \qquad t = 0,\dots,T-1, \tag{155}$$

assuming that the inverse of the matrix in square brackets exists. Starting from adjoint equation (151) and substituting (154), we find

$$\begin{aligned}
-P(t)x(t) &= -A(t)^\top P(t+1)x(t+1) - Q(t)x(t) \\
&\overset{(148)}{=} -A(t)^\top P(t+1)\left(A(t)x(t) + B(t)u(t)\right) - Q(t)x(t) \\
&\overset{(155)}{=} -A(t)^\top P(t+1)A(t)x(t) \\
&+ \ A(t)^\top P(t+1)B(t)\left[R(t) + B(t)^\top P(t+1)B(t)\right]^{-1} B(t)^\top P(t+1)A(t)x(t) - Q(t)x(t).
\end{aligned}$$

Consequently, our guess that the costate is linear in the state is justified if we solve the sequence of matrices $P(t)$ from the following recursion, known as the discrete-time ***Riccati equation***:

$$\begin{aligned}
&P(t) = A(t)^\top P(t+1)A(t) - A(t)^\top P(t+1)B(t)\left[R(t) + B(t)^\top P(t+1)B(t)\right]^{-1} B(t)^\top P(t+1)A(t) + Q(t) \\
&P(T) = Q(T).
\end{aligned}$$

Exercises section 21

21.1 Solve the optimization problem of Exercise 20.7 using the maximum principle.

21.2 Solve the optimization problem of Exercise 20.6 using the maximum principle.

22 Problem formulation for an infinite horizon

The transition from a finite to an infinite horizon introduces a range of complications. The first is that the goal function becomes a sum of infinitely many terms, so we need to make sure that it is well-defined. This issue is addressed in the present section.

Definition 22.1 In its standard form, a discrete-time dynamic optimization problem (DP) with infinite horizon is of the form

$$\textbf{\textit{DP:}} \quad \begin{cases} \text{maximize} & \sum_{t=0}^{\infty} f(t,x(t),u(t)) & \\ \text{with} & u(t) \in U(t,x(t)) & t \in \mathbb{Z}_+ = \{0,1,2,\dots\} \\ & x(t+1) = g(t,x(t),u(t)) & t \in \mathbb{Z}_+ \\ & x(0) = x_0 \text{ (given)} & \end{cases} \tag{156}$$

Its ingredients are familiar from Definition 20.1 and feasible/admissible pairs (x,u), etc. are defined analogously to Definition 20.2. The only relevant change is that we replaced the finite horizon T by ∞: we consider an infinite number of time periods $t \in \mathbb{Z}_+ = \{0,1,2,\dots\}$.

An infinite sum like $\sum_{t=0}^{\infty} f(t,x(t),u(t))$ or, more generally, $\sum_{t=0}^{\infty} x_t$ for a sequence $(x_t)_{t\in\mathbb{Z}_+}$ of real numbers is often called a ***series***; it is defined as the limit

$$\sum_{t=0}^{\infty} x_t = \lim_{T\to\infty} \sum_{t=0}^{T} x_t$$

of the ***partial sums*** $\sum_{t=0}^{T} x_t$, if this limit exists. In that case, we call the sequence ***summable***.

Note that we index the terms of our sequence $(x_t)_{t\in\mathbb{Z}_+}$ by the nonnegative integers $\mathbb{Z}_+ = \{0,1,2,\dots\}$, while we earlier mostly denoted a sequence by $(x_t)_{t\in\mathbb{N}}$, using positive integers $\mathbb{N} = \{1,2,\dots\}$. That is, we now start at 0 instead of 1. This is just a notational convenience. It happens to be nice for our purposes, since time in most applied models is assumed to start at zero, rather than at one.

If we want the sequence of partial sums $\sum_{t=0}^{T} x_t$ to converge, the numbers x_t must converge to zero. The sequence $(1,-1,1,-1,1,-1,\dots)$, for instance, is not summable: the partial sum of an even number of terms is zero, the partial sum of an odd number of terms is one, so the partial sums do not converge. But we need more: some sequences that converge to zero are still not summable. The sequence $(1/2,1/3,1/4,1/5,\dots)$, for instance, converges to zero, but is not summable:

$$\frac{1}{2} + \underbrace{\frac{1}{3} + \frac{1}{4}}_{>2\cdot\frac{1}{4}=\frac{1}{2}} + \underbrace{\frac{1}{5} + \frac{1}{6} + \frac{1}{7} + \frac{1}{8}}_{>4\cdot\frac{1}{8}=\frac{1}{2}} + \cdots > \frac{1}{2} + \frac{1}{2} + \frac{1}{2} + \frac{1}{2} + \cdots$$

has 2^1 terms exceeding $\frac{1}{4}$, then 2^2 terms exceeding $\frac{1}{8}$, then 2^3 terms exceeding $\frac{1}{16}$, etc.: the partial sums diverge to infinity.

Consequently, we need to assure that the infinite sum in problem (156) is well-defined. A rather tolerant assumption would be:

> **(Nontriviality)** The feasible set is nonempty. The infinite sum in the goal function lies in $\mathbb{R} \cup \{-\infty\}$ for each feasible candidate and in $\mathbb{R}$ for some feasible candidate(s).

The nonemptiness requirement needs little motivation. The rest says that partial sums diverging to minus infinity are okay, as long as some feasible sequences of partial sums have a finite limit: infinitely bad outcomes in our *maximization* problem pose no threat as long as we can avoid them. The assumption rules out two kinds of behavior: (1) partial sums diverging to plus infinity, the ‘infinitely good' outcomes of our maximization problem, since most of our mathematical tools break down in such cases, and (2) partial sums that do not converge at all, in which case the infinite sum makes no sense whatsoever. A fairly general theorem that assures the right kind of summability is:

Theorem 22.1

If sequences $x_0, x_1, x_2, \ldots$ and $y_0, y_1, y_2, \ldots$ in $\mathbb{R}$ satisfy $x_t \leq y_t$ for all t and sequence $(y_t)_{t \in \mathbb{Z}_+}$ is summable, then

$$\lim_{T \to \infty} \sum_{t=0}^{T} x_t = -\infty \qquad \text{or} \qquad \lim_{T \to \infty} \sum_{t=0}^{T} x_t \in \mathbb{R}.$$

Proof: Let $\sum_{t=0}^{\infty} y_t = y \in \mathbb{R}$ and write $x_t = (x_t - y_t) + y_t$. The sequence of partial sums $\sum_{t=0}^{T}(x_t - y_t)$ is weakly decreasing by assumption.

☒ If it is not bounded from below, then for each $r \in \mathbb{R}$, we can find an $N \in \mathbb{N}$ such that $T \geq N$ implies

$$\sum_{t=0}^{T}(x_t - y_t) < r - y - 1 \qquad \text{and} \qquad \left|\sum_{t=0}^{T} y_t - y\right| < 1,$$

so

$$\sum_{t=0}^{T} x_t = \sum_{t=0}^{T}(x_t - y_t) + \sum_{t=0}^{T} y_t < (r - y - 1) + (y + 1) = r,$$

showing that $\sum_{t=0}^{T} x_t$ diverges to $-\infty$.

☒ If it is bounded from below, then it converges to its infimum, say $z \in \mathbb{R}$. Hence, for each $\varepsilon > 0$ we can find an $N \in \mathbb{N}$ such that $T \geq N$ implies

$$\left|\sum_{t=0}^{T}(x_t - y_t) - z\right| < \varepsilon/2 \qquad \text{and} \qquad \left|\sum_{t=0}^{T} y_t - y\right| < \varepsilon/2.$$

By the triangle inequality,

$$\left|\sum_{t=0}^{T} x_t - (z + y)\right| \leq \left|\sum_{t=0}^{T}(x_t - y_t) - z\right| + \left|\sum_{t=0}^{T} y_t - y\right| < \varepsilon/2 + \varepsilon/2 = \varepsilon,$$

showing that $\sum_{t=0}^{T} x_t$ has a well-defined limit $z + y$ in $\mathbb{R}$. □

Remark 22.1 (Sandwich property) Analogously, if there is a summable sequence $z_0, z_1, z_2, \ldots$ with $z_t \leq x_t$ for all t, the partial sums $\sum_{t=0}^{T} x_t$ either diverge to $+\infty$ or converge. A forteriori, if $z_t \leq x_t \leq y_t$ for summable sequences y and z, then the sequence x that is 'sandwiched' in between them is summable as well. This holds in particular if $|x_t| \leq |y_t|$ for some summable sequence y. We apply the sandwich property to a common class of discounted infinite-horizon problems in the next section. ◅

Macro-economists seem to prefer a reduced form problem

$$\left\{\begin{array}{lll} \text{maximize} & \sum_{t=0}^{\infty} f(t, x(t), x(t+1)) & \\ \text{with} & (x(t), x(t+1)) \in X(t) & t \in \mathbb{Z}_+ \\ & x(0) = x_0 \text{ (given)}, & \end{array}\right. \tag{157}$$

an infinite-horizon version of Example 20.4, after rewriting the restriction $x(t+1) \in U(t, x(t))$ to a restriction $(x(t), x(t+1)) \in X(t)$ on consecutive pairs of controls. For instance, a typical problem in economic growth is of the form

$$\begin{array}{lll} \sup_{(k(t),c(t))_{t\in\mathbb{Z}_+}} & \sum_{t=0}^{\infty}\beta^t u(c(t)) & \\ \text{subject to} & k(t+1)=f(k(t))+(1-\delta)k(t)-c(t) & t\in\mathbb{Z}_+ \\ & k(t),c(t)\geq 0 & t\in\mathbb{Z}_+ \\ & k(0)\text{ given} & \end{array}$$

Here, $u(c(t))$ is the instantaneous utility of consuming $c(t)$ at time t. The capital-labor ratio $k(t+1)$ at time $t+1$ is determined by total output $f(k(t))$ and the fraction $(1-\delta)k(t)$ of capital-labor ratio $k(t)$ that remains after fraction δ has depreciated, minus consumption $c(t)$.

To write this in the reduced form, note that $c(t)=f(k(t))+(1-\delta)k(t)-k(t+1)$, so if we substitute this into the function u and introduce

$$F(k(t),k(t+1))=u(c(t))=u\left[f(k(t))+(1-\delta)k(t)-k(t+1)\right], \tag{158}$$

this problem reduces to

$$\begin{array}{lll} \sup_{(k(t))_{t\in\mathbb{Z}_+}} & \sum_{t=0}^{\infty}\beta^t F(k(t),k(t+1)) & \\ \text{subject to} & 0\leq k(t+1)\leq f(k(t))+(1-\delta)k(t) & t\in\mathbb{Z}_+ \\ & k(0)\text{ given,} & \end{array}$$

a special instance of (157) with $f(t,k(t),k(t+1))=\beta^t F(k(t),k(t+1))$ and

$$X(t)=\{(k(t),k(t+1))\mid 0\leq k(t+1)\leq f(k(t))+(1-\delta)k(t)\}.$$

Applying Theorem 22.1, one can find reasonable assumptions under which the goal function is well-defined. Given (158), for instance, it seems reasonable to assume that $F(k(t),k(t+1))$ increases with $k(t)$ and decreases with $k(t+1)$. Moreover, each $k(t)$ is assumed to be nonnegative. Assumptions like these are spelled out in a slightly more general setting in Exercise 22.2.

Exercises section 22

22.1 In each of the following cases, show that the real sequence $x_0,x_1,x_2,\ldots$ is summable:

(a) There are $\alpha\in[0,1)$ and $N\in\mathbb{N}$ such that $t\geq N$ implies $|x_{t+1}|\leq\alpha|x_t|$.

(b) There are $\alpha>0,\beta>1$, and $N\in\mathbb{N}$ such that $t\geq N$ implies $|x_t|\leq\frac{\alpha}{t^\beta}$.

22.2 Consider a reduced form model with payoffs discounted by a given $\beta\in(0,1)$:

$$\left\{\begin{array}{lll} \text{maximize} & \sum_{t=0}^{\infty}\beta^t f(x(t),x(t+1)) & \\ \text{with} & (x(t),x(t+1))\in X(t) & t\in\mathbb{Z}_+ \\ & x(0)=x_0\text{ (given),} & \end{array}\right.$$

We provide three sets of sufficient conditions to assure that the goal function is well-defined.

(a) In this part of the exercise, assume:

(i) Vectors $x(t)$ are nonnegative vectors in $\mathbb{R}^k$, i.e., $x_0\in\mathbb{R}^k_+$ and $X(t)\subseteq\mathbb{R}^k_+\times\mathbb{R}^k_+$ for each $t\in\mathbb{Z}_+$.

(ii) $f:\mathbb{R}^k_+\times\mathbb{R}^k_+\to\mathbb{R}$ is increasing in its first k coordinates and decreasing in its final k coordinates.

(iii) There is a $\theta\in(0,1]$ such that for all feasible $(x(t))_{t\in\mathbb{Z}_+}$ and $t\in\mathbb{Z}_+$: $\|x(t+1)\|\leq\theta\|x(t)\|$.

We prove that for each feasible sequence $(x(t))_{t\in\mathbb{Z}_+}$ and each $t\in\mathbb{Z}_+$:

$$\beta^t f(x(t),x(t+1))\leq\beta^t f(\|x_0\|\mathbf{1},\mathbf{0}), \tag{159}$$

where $\mathbf{1}=(1,\ldots,1)\in\mathbb{R}^k$ is a k-dimensional vector of ones. The sequence of terms on the right-hand side is summable, so the goal function is well-defined by Theorem 22.1.

(a1) Let $(x(t))_{t\in\mathbb{Z}_+}$ be a feasible sequence and let $t \in \mathbb{Z}_+$. Prove:

$$\|x(t)\| \leq \theta^t \|x_0\| \qquad \text{and} \qquad x(t) \leq \theta^t \|x_0\| \mathbf{1}.$$

(a2) For each $t \in \mathbb{Z}_+$, establish the following chain of inequalities:

$$f(x(t), x(t+1)) \leq f(x(t), \mathbf{0}) \leq f(\theta^t \|x_0\| \mathbf{1}, \mathbf{0}) \leq f(\|x_0\| \mathbf{1}, \mathbf{0}),$$

which implies the desired result (159)!

(b) In this part of the exercise, assume (i), (ii), and

(iv) There is a $\theta \in (1, 1/\beta)$ such that for all feasible $(x(t))_{t\in\mathbb{Z}_+}$ and $t \in \mathbb{Z}_+$: $\|x(t+1)\| \leq \theta \|x(t)\|$.

(v) $f(\mathbf{0}, \mathbf{0}) = \mathbf{0}$.

(vi) $f(\cdot, \mathbf{0})$ is concave.

Let $(x(t))_{t\in\mathbb{Z}_+}$ be a feasible sequence and let $t \in \mathbb{Z}_+$. Write $\|x_0\|\mathbf{1} = \frac{\theta^t}{\theta^t}\|x_0\|\mathbf{1} + \left(1 - \frac{1}{\theta^t}\right)\mathbf{0}$ and use concavity to show that

$$f(\theta^t \|x_0\| \mathbf{1}, \mathbf{0}) \leq \theta^t f(\|x_0\| \mathbf{1}, \mathbf{0}).$$

Deduce that

$$\beta^t f(x(t), x(t+1)) \leq (\beta\theta)^t f(\|x_0\| \mathbf{1}, \mathbf{0}). \tag{160}$$

Since $\beta\theta \in (0, 1)$ by (iv), the sequence of terms on the right-hand side is summable: the goal function is well-defined by Theorem 22.1.

(c) In this part of the exercise, assume (i), (ii), and

(vii) There is a $\theta \in (0, 1/\beta)$ such that for all feasible $(x(t))_{t\in\mathbb{Z}_+}$ and $t \in \mathbb{Z}_+$: $f(x(t+1), \mathbf{0}) \leq \theta f(x(t), \mathbf{0})$.

Prove that for each feasible $(x(t))_{t\in\mathbb{Z}_+}$ and each $t \in \mathbb{Z}_+$:

$$\beta^t f(x(t), x(t+1)) \leq \left(\beta\theta\right)^t f(x_0, \mathbf{0}).$$

Since the sequence of terms on the right-hand side is summable, the goal function is well-defined by Theorem 22.1.

23 Discounted infinite-horizon problems

As we saw in the previous section, one complication in infinite horizon problems arises since we need to make sure that the goal function, as a sum of infinitely many terms, is well-defined. Another complication is that our earlier methods of solving finite-horizon problems no longer work:

☒ The dynamic programming algorithm works backwards from the final period. There is no such period in an infinite-horizon problem.

☒ The variant of the maximum principle we discussed relies on applications of constrained optimization techniques that we developed only for vectors of finite dimension.

Fortunately, the archetypical infinite-horizon problem, the extension of Example 20.5, has a sufficiently simple structure that an approach much like the dynamic programming algorithm still works. This approach uses the famous Bellman equation.

23.1 Problem formulation

We restrict attention to a simple infinite-horizon version of the discrete-time dynamic programming problem (DP):

> **Definition 23.1** A ***stationary (infinite-horizon, dynamic optimization) problem*** with discount factor $\beta \in (0,1)$ is of the form
>
> $$\textbf{\textit{DP:}} \quad \begin{cases} \sup_{u=(u(t))_{t=0}^{\infty}} & \sum_{t=0}^{\infty} \beta^t f(x(t), u(t)) & \\ \text{subject to} & u(t) \in U(x(t)) & t \in \mathbb{Z}_+ = \{0,1,2,\ldots\} \\ & x(t+1) = g(x(t), u(t)) & t \in \mathbb{Z}_+ \\ & x(0) \text{ given} & \end{cases} \tag{161}$$
>
> The problem is stationary, since neither f nor g depends explicitly on time.

Feasible/admissible pairs (x, u), etc. are defined analogously to Definition 20.2. Let $\Phi(x_0)$ denote the set of feasible pairs (x, u) if the initial state is x_0:

$$\Phi(x_0) = \{(x,u) = (x(t), u(t))_{t \in \mathbb{Z}_+} : x(0) = x_0 \text{ and } u(t) \in U(x(t)), x(t+1) = g(x(t), u(t)) \text{ for all } t \in \mathbb{Z}_+\}.$$

Throughout most of this section, we assume:

(B) The function f is bounded.

By Remark 22.1, this makes the goal function well-defined: if $(x(t))_{t \in \mathbb{Z}_+}$ is feasible and f is bounded by $M \geq 0$, then

$$|\beta^t f(x(t), x(t+1))| \leq \beta^t M \text{ for all } t \qquad \text{and} \qquad \sum_{t=0}^{T} \beta^t M = \frac{1-\beta^{T+1}}{1-\beta} M \to \frac{1}{1-\beta} M,$$

making the latter summable. In applications, the boundedness assumption (B) may not hold. Then one of the following insights may come in handy:

☒ Of course, the boundedness condition needs to hold only over pairs of states and controls that can occur in combination: we need that the set of images $f(x(t), u(t))$ is bounded by looking at $(x(t), u(t))$ that occur in feasible pairs; it is perfectly okay that the function f behaves weirdly in points of its domain that are not reached by feasible solutions to the optimization problem anyway!

☒ Sometimes, a subtle insight or modeling choice may allow you to restrict attention to sets where the boundedness condition does hold.

As examples of the latter point, it is often more elegant to model — in growth theory — consumption choices in each period as a choice of what *fraction* $u(t) \in [0,1]$ of your income you want to spend at time t, rather than a choice of what *amount*, since the latter may change over time without being bounded a priori. Moreover, sometimes the domain over which you optimize may be restricted — like we did earlier in Theorem 16.1 — by discarding feasible pairs that turn out for some reason to be suboptimal. The function f may then be bounded if one restricts attention to feasible pairs among the remaining candidates.

23.2 Optimal values, optimal solutions, and the Bellman equation

Just like we did in the finite-horizon case, we're going to look at the (optimal) value function and show that it has a nice property. That nice property was the dynamic programming algorithm in Theorem 20.2 that allowed us to solve problems with a finite horizon backwards. In the infinite-horizon case, the nice property is going to be the Bellman equation in Theorem 23.1, which essentially says that the value function is a fixed point of a certain mapping. But let's not run too fast ahead...

Definition 23.2 The ***value function*** $J : X \to \mathbb{R}$ assigns to each initial state $x_0 \in X$ the optimal value

$$J(x_0) = \sup_{(x,u)\in\Phi(x_0)} \sum_{t=0}^{\infty} \beta^t f(x(t), u(t))$$

of the goal function.

This definition, of course, is fine under our boundedness/nontriviality assumptions. Below, the formulas may look messy, but just try to get the central idea: we apply our divide-and-conquer strategy to divide optimization over the entire set of feasible pairs into different subproblems, one for each initial choice $u(0)$. The tail problem that we face in the next period $t = 1$ will be in state $g(x(0), u(0))$, but looks suspiciously like the initial problem at time $t = 0$: it still has an infinite horizon, we still know the initial state $g(x(0), u(0))$: the only real difference is that all payoffs we get occur with a delay of one period, i.e., are discounted by a factor β. So if J tells us what the optimal value is given any initial state, after $u(0)$, we run into a problem with value $\beta J(g(x(0), u(0)))$.

Theorem 23.1 (Bellman equation)

The value function satisfies the Bellman equation:

$$\text{for each } x \in X: \qquad J(x) = \sup_{u\in U(x)} \{f(x,u) + \beta J(g(x,u))\}. \tag{162}$$

Proof: For each $x \in X$:

$$(x(t), u(t))_{t\in\mathbb{Z}_+} \in \Phi(x) \Leftrightarrow \begin{cases} x(0) & = & x, \\ u(0) & \in & U(x), \\ (x(t+1), u(t+1))_{t\in\mathbb{Z}_+} & \in & \Phi(g(x(0), u(0))). \end{cases} \tag{163}$$

By Theorem 20.1 and (163), for each $x \in X$:

$$\begin{aligned} J(x) &= \sup_{\Phi(x)} \sum_{t=0}^{\infty} \beta^t f(x(t), u(t)) \\ &= \sup_{u \in U(x)} \sup_{\Phi(g(x,u))} \left(f(x,u) + \sum_{t=1}^{\infty} \beta^t f(x(t), u(t)) \right) \\ &= \sup_{u \in U(x)} \left(f(x,u) + \sup_{\Phi(g(x,u))} \sum_{t=1}^{\infty} \beta^t f(x(t), u(t)) \right) \\ &= \sup_{u \in U(x)} \left(f(x,u) + \beta \sup_{\Phi(g(x,u))} \sum_{t=0}^{\infty} \beta^t f(x(t+1), u(t+1)) \right) \\ &= \sup_{u \in U(x)} \big(f(x,u) + \beta J(g(x,u)) \big). \end{aligned}$$

□

There may, however, be other solutions. For $f(x,u) = 0$ and $g(x,u) = x/\beta$, we have the trivial problem

$$\text{maximize} \sum_{t=0}^{\infty} \beta^t \cdot 0 \qquad \text{with} \qquad u(t) \in \mathbb{R}, x(t+1) = x(t)/\beta, \text{ and } x(0) = x_0 \text{ given.}$$

The goal function is zero, no matter what you do: every feasible control sequence is optimal and the value function is the zero function. But *every* linear function $V : x \mapsto cx$ satisfies the Bellman equation:

$$\max_{u \in \mathbb{R}} f(x,u) + \beta V(g(x,u)) = \max_{u \in \mathbb{R}} 0 + \beta c \left(\frac{x}{\beta} \right) = cx = V(x).$$

So on the one hand, the Bellman equation tells us something about the solution to our infinite-horizon optimization problem; on the other hand, there may be a whole bunch of other solutions to make our life miserable. A powerful tool brings together a lot of our earlier work into one very important and useful conclusion:

Theorem 23.2

There is exactly one *bounded* solution to the Bellman equation: the value function J.

Proof: By Theorem 23.1, the value function J satisfies the Bellman equation and is bounded by boundedness assumption (B). So if there is only one solution, it must equal J.

Let us now look (here comes the heavy artillery) at the metric space $((B(X,\mathbb{R}), d_\infty)$ of bounded real-valued functions on the state space X, endowed with the supremum metric. As we concluded above, value function J lies in this space. Moreover, since $\mathbb{R}$ with its usual distance is a complete space, so is $((B(X,\mathbb{R}), d_\infty)$ by Theorem 10.3. The mapping $T : B(X,\mathbb{R}) \to B(X,\mathbb{R})$ with

$$T(J)(x) = \sup_{u \in U(x)} \{ f(x,u) + \beta J(g(x,u)) \}$$

is a contraction: if J and J' are bounded, then

$$\begin{aligned} T(J)(x) &= \sup_{u \in U(x)} \{ f(x,u) + \beta J'(g(x,u)) + \beta (J(g(x,u)) - J'(g(x,u))) \} \\ &\leq \sup_{u \in U(x)} \{ f(x,u) + \beta J'(g(x,u)) + \beta d_\infty(J, J') \} \\ &= \sup_{u \in U(x)} \{ f(x,u) + \beta J'(g(x,u)) \} + \beta d_\infty(J, J') \\ &= T(J')(x) + \beta d_\infty(J, J'). \end{aligned}$$

Similarly, $T(J')(x) \leq T(J)(x) + \beta d_\infty(J, J')$. Hence

$$d_\infty(T(J), T(J')) = \sup_{x \in X} |T(J)(x) - T(J')(x)| \leq \beta d_\infty(J, J'),$$

so T is a contraction with modulus β. By the Banach contraction theorem, it has a unique fixed point, which, by definition of T, is a solution to the Bellman equation. □

Isn't she pretty, Theorem 23.2? Moreover, it is a key ingredient in the value and policy iteration algorithms of Section 23.3 to approach optimal solutions. Just as before, anything that solves the optimization problem has to generate an optimal solution in its tail problems as well.

Theorem 23.3

Let $(x^*, u^*) \in \Phi(x_0)$ solve the optimization problem (161) with initial state x_0.

(a) For each $t \in \mathbb{Z}_+$, $(x^*(s+t), u^*(s+t))_{s \in \mathbb{Z}_+}$ solves the problem with initial state $x^*(t)$.

(b) For each $t \in \mathbb{Z}_+$, define $J(x^*(t)) = \sup_{(x,u) \in \Phi(x^*(t))} \sum_{s=0}^\infty \beta^s f(x(s), u(s))$. Then

$$J(x^*(t)) = f(x^*(t), u^*(t)) + \beta J(x^*(t+1)).$$

(c) For each $t \in \mathbb{Z}_+$:

$$J(x^*(t)) = \sup_{u \in U(x^*(t))} \left\{ f(x^*(t), u) + \beta J(g(x^*(t), u)) \right\}.$$

Proof: **(a)** The proof is by induction on t. The result is true by assumption if $t = 0$. Now assume the result is true for $t \in \mathbb{Z}_+$; let's prove it for $t+1$. By the induction hypothesis:

$$\begin{aligned} \sup_{(x,u) \in \Phi(x^*(t))} \sum_{s=0}^\infty \beta^s f(x(s), u(s)) &= \sum_{s=0}^\infty \beta^s f(x^*(s+t), u^*(s+t)) \\ &= f(x^*(t), u^*(t)) + \beta \sum_{s=0}^\infty \beta^s f(x^*(t+1+s), u^*(t+1+s)). \quad (164) \end{aligned}$$

Let $(\hat{x}, \hat{u}) \in \Phi(x^*(t+1))$. Appending $x^*(t)$ and $u^*(t)$ at the beginning yields an element of $\Phi(x^*(t))$, so

$$\sup_{(x,u) \in \Phi(x^*(t))} \sum_{s=0}^\infty \beta^s f(x(s), u(s)) \geq f(x^*(t), u^*(t)) + \beta \sum_{s=0}^\infty \beta^s f(\hat{x}(s), \hat{u}(s)). \tag{165}$$

Then (164) and (165) show that

$$\sum_{s=0}^\infty \beta^s f(x^*(t+1+s), u^*(t+1+s)) \geq \sum_{s=0}^\infty \beta^s f(\hat{x}(s), \hat{u}(s)).$$

This inequality holds for arbitrary $(\hat{x}, \hat{u}) \in \Phi(x^*(t+1))$, so tail sequence $(x^*(s+t+1), u^*(s+t+1))_{s \in \mathbb{Z}_+}$ is optimal for initial state $x^*(t+1)$.

(b) Using (a) twice yields

$$\begin{aligned} J(x^*(t)) &= f(x^*(t), u^*(t)) + \beta \sum_{s=0}^\infty \beta^s f(x^*(t+1+s), u^*(t+1+s)) \\ &= f(x^*(t), u^*(t)) + \beta J(x^*(t+1)). \end{aligned}$$

(c) By Theorem 23.1, the value function satisfies the Bellman equation. □

Conversely, suppose you find a feasible control satisfying the Bellman equation with equality. Then it has to be optimal!

Theorem 23.4

Let $(x^*, u^*) \in \Phi(x_0)$ be a feasible pair that yields a maximum in the Bellman equation for the value function J at each time $t \in \mathbb{Z}_+$. That is, for each $t \in \mathbb{Z}_+$:

$$J(x^*(t)) = \sup_{u \in U(x)} \left\{ f(x^*(t), u) + \beta J(g(x^*(t), u)) \right\} = f(x^*(t), u^*(t)) + \beta J(g(x^*(t), u^*(t))).$$

Then (x^*, u^*) solves the DP with initial state x_0: $J(x_0) = \sum_{s=0}^{\infty} \beta^s f(x^*(s), u^*(s))$.

Proof: By induction on t, we have for each $t \in \mathbb{N}$ that

$$J(x_0) = J(x^*(0)) = \sum_{s=0}^{t} \beta^s f(x^*(s), u^*(s)) + \beta^{t+1} J(x^*(t+1)).$$

Since J is bounded by assumption, the right-hand side converges to $\sum_{s=0}^{\infty} \beta^s f(x^*(s), u^*(s))$. □

Make sure to read this correctly: *if* you know the value function J and feasible pair (x^*, u^*) achieves the maximum value in the Bellman equation, *then* it must generate the optimal value in the original problem. Recall, however, that there may be many functions J^B and corresponding maximizers (x^*, u^*) satisfying the Bellman equation.

Many interesting properties like concavity, continuity, differentiability, etc., of the optimal value function can be derived by introducing additional assumptions in the dynamic optimization problem. For now, however, I want to concentrate on two other issues:

1. The main ideas behind two algorithms to approximate the optimal value function: value and policy iteration. This is discussed in Section 23.3.

2. Deriving a set of necessary and sufficient conditions, often referred to as the Euler equations and the transversality condition, for a common class of dynamic optimization problems in macroeconomic growth theory. This is discussed in Section 24.

23.3 Algorithms for the value function: value and policy iteration

In large, complicated models that fall outside a relatively restricted class for which explicit solutions are easily found — like the LQ control problem — one often has to rely on approximations of optimal solutions. Theorem 23.2 tells us that the optimal value function is a fixed point of the contraction $T : B(X, \mathbb{R}) \to B(X, \mathbb{R})$ with

$$T(J)(x) = \sup_{u \in U(x)} \left\{ f(x, u) + \beta J(g(x, u)) \right\}$$

on the complete metric space $(B(X, \mathbb{R}), d_\infty)$ of bounded functions on the state space. By Banach's fixed point theorem (Theorem 11.1) the following ***value iteration*** algorithm generates a sequence $V_0, V_1, V_2, \ldots$ of bounded functions $V_k : X \to \mathbb{R}$ that converges to the optimal value function J:

1. Let $V_0 : X \to \mathbb{R}$ be any bounded function. For instance, the zero function: $V_0(x) = 0$ for all $x \in X$.

2. At each iteration k, calculate $V_{k+1} = T(V_k)$, that is, let

$$V_{k+1}(x) = \sup_{u \in U(x)} f(x, u) + \beta V_k(g(x, u)).$$

Another method, ***policy iteration***, works like this:

1. Let $\pi_0 : X \to U$ be a feasible policy: $\pi_0(x) \in U(x)$ for all $x \in X$.
2. At each iteration k, given the current policy π_k, do the following:
   (a) policy evaluation: compute the value from using that policy to choose controls at all times:
   $$V_k(x) = \sum_{t=0}^{\infty} \beta^t f(x(t), \pi_k(x(t))) \qquad \text{with } x(t+1) = g(x(t), \pi_k(x(t))) \text{ and } x(0) = x. \tag{166}$$
   (b) policy improvement: compute a new policy π_{k+1} by defining $\pi_{k+1}(x) \in U(x)$ to be a control that solves the two-period problem
   $$\max_{u \in U(x)} f(x,u) + \beta V_k(g(x,u)), \tag{167}$$
   assuming such a control exists.

This algorithm generates better and better policies: $V_{k+1} \geq V_k$ for all k. In case of equality, π_k is optimal. Indeed, for each policy π_k, the function $T_k : B(X,\mathbb{R}) \to B(X,\mathbb{R})$ with

$$T_k(V)(x) = f(x, \pi_k(x)) + \beta V(g(x, \pi_k(x)))$$

satisfies Blackwell's conditions (Exercise 11.1): it is a contraction and has a unique fixed point $V_k \in B(X,\mathbb{R})$. With $x(t)$ as in (166), induction gives, for each $T \in \mathbb{N}$:

$$V_k(x) = \sum_{t=0}^{T} \beta^t f(x(t), \pi_k(x(t))) + \beta^{T+1} V_k(x(T+1)).$$

Taking the limit, we see that $V_k(x) = \sum_{t=0}^{\infty} \beta^t f(x(t), \pi_k(x(t)))$ is the requested policy evaluation. In the policy improvement step, we choose policy π_{k+1} optimally in the 2-period problem (167), so $T_{k+1}(V_k) \geq T_k(V_k) = V_k$. Monotonicity of T_{k+1} implies $T_{k+1}^n(V_k) \geq T_k(V_k) = V_k$ for all iterates n. In the limit, $V_{k+1} = \lim_{n\to\infty} T_{k+1}^n(V_k) \geq V_k$. Moreover, if $V_{k+1} = V_k$, we see that $V_k \in B(X,\mathbb{R})$ satisfies the Bellman equation: it is optimal.

Apart from being useful in complicated models, value or policy iteration is useful in simple models if you don't really know where to start to solve them. Just do some steps of this value iteration process. In simple examples, you might after a few steps start to recognize a pattern. For instance, if you start with a quadratic function and you get a quadratic function back, it seems an educated guess that the value function is quadratic. Then all of a sudden, you've almost solved the problem!

Exercises section 23

23.1 On page 138 we considered the trivial optimization problem

$$\begin{array}{lll} \text{maximize} & \sum_{t=0}^{\infty} \beta^t \cdot 0 & \\ \text{with} & u(t) \in \mathbb{R} & t \in \mathbb{Z}_+ \\ & x(t+1) = x(t)/\beta & t \in \mathbb{Z}_+ \\ & x(0) = x_0 & \end{array}$$

for a given initial state x_0 and discount factor $\beta \in (0,1)$: all choices are optimal and the optimal value function $J : X \to \mathbb{R}$ is everywhere equal to zero. But we found many solutions to the Bellman equation. Here are some more.

(a) Write down the Bellman equation for this problem.

(b) Starting from a constant function V_0, what is the k-th function V_k in the value iteration algorithm? Do these functions converge to the optimal value function?

(c) Show that all functions of the form $V(x) = ax + b|x|$ for constants a and b in $\mathbb{R}$ are fixed points of the Bellman equation.

23.2 Given initial state $x_0 \in [0,1]$ and discount factor $\beta \in (0,1)$, consider the problem

$$\begin{array}{lll} \text{maximize} & \sum_{t=0}^{\infty} \beta^t \sqrt{x(t)u(t)} & \\ \text{with} & u(t) \in [0,1] & t \in \mathbb{Z}_+ \\ & x(t+1) = 1 - u(t) & t \in \mathbb{Z}_+ \\ & x(0) = x_0 & \end{array}$$

(a) Verify that the instantaneous payoff function $f(x,u) = \sqrt{xu}$ is a bounded function of the states x and controls u in $[0,1]$.

(b) Write down the Bellman equation for this problem.

(c) Starting from the constant function $V_0 : [0,1] \to \mathbb{R}$ with $V_0(x) = 0$ for all x, use the value iteration algorithm to compute V_1 and V_2.

(d) Verify that $V : [0,1] \to \mathbb{R}$ with $V(x) = \frac{\sqrt{x+\beta^2(1-x)}}{1-\beta^2}$ is a fixed point of the Bellman equation.

(e) This V is the optimal value function; why? And what is the optimal policy function?

24 Euler equations and transversality

In this section, we derive optimality conditions in a reduced-form model that is common in economic growth theory. It is partly based on T. Kamihigashi (2002) "A simple proof of the necessity of the transversality condition", Econ. Theory 20, 427–433. Consider the problem

$$\begin{cases} \text{maximize} & \sum_{t=0}^{\infty} f(t, x(t), x(t+1)) & \\ \text{with} & (x(t), x(t+1)) \in X(t) & t \in \mathbb{Z}_+ \\ & x(0) = x_0 \text{ (given),} & \end{cases}$$

an infinite-horizon version of Example 20.4, after rewriting the restriction $x(t+1) \in U(t, x(t))$ to a restriction $(x(t), x(t+1)) \in X(t)$ on consecutive pairs of controls. Assume:

(A1) For each feasible $(x(t))_{t\in\mathbb{Z}_+}$, payoff $\sum_{t=0}^{\infty} f(t, x(t), x(t+1))$ is well-defined in $\mathbb{R}$.

(A2) Vectors $x(t)$ are nonnegative vectors in $\mathbb{R}^k$, i.e., $x_0 \in \mathbb{R}^k_+$ and $X(t) \subseteq \mathbb{R}^k_+ \times \mathbb{R}^k_+$ for each $t \in \mathbb{Z}_+$.

(A3) For each $t \in \mathbb{Z}_+$, region $X(t)$ is convex and contains $(\mathbf{0}, \mathbf{0})$.

(A4) For each $t \in \mathbb{Z}_+$, reward $f(t, \cdot, \cdot)$ is continuously differentiable on the interior of $X(t)$ and concave.

(A5) For each $t \in \mathbb{Z}_+$ and each (y, z) in the interior of $X(t)$, the vector $f_3'(t, y, z)$ of partial derivatives with respect to z satisfies $f_3'(t, y, z) \leq \mathbf{0}$.

Theorem 24.1

Assume $(x^*(t))_{t\in\mathbb{Z}_+}$ is optimal and interior: $(x^*(t), x^*(t+1))$ lies in the interior of $X(t)$ for each $t \in \mathbb{Z}_+$. Then it satisfies the ***Euler equations***:

$$\text{for each } t \in \mathbb{Z}_+: \qquad f_3'(t, x^*(t), x^*(t+1)) + f_2'(t+1, x^*(t+1), x^*(t+2)) = \mathbf{0}^\top, \tag{168}$$

and the ***transversality condition***:

$$\lim_{t\to\infty} -f_3'(t, x^*(t), x^*(t+1))x^*(t+1) = 0. \tag{169}$$

Conversely, if $(x^*(t))_{t\in\mathbb{Z}_+}$ is feasible, interior, and satisfies (168) and (169), then it is optimal.

Proof: NECESSITY OF (168) AND (169): For $t \in \mathbb{Z}_+$, control $x(t+1)$ appears twice in the goal function:

$$\cdots + f(t, x(t), x(t+1)) + f(t+1, x(t+1), x(t+2)) + \cdots \tag{170}$$

The optimal x^* is interior to the control regions, so we can make small changes in the coordinates of $x^*(t+1)$ without affecting the variables $x^*(s)$ at times $s \neq t+1$. Condition (168) is the familiar condition that in an interior solution, the derivative of (170) with respect to $x(t+1)$ must be zero.

To establish (169), let $T \in \mathbb{Z}_+$. For each $\lambda \in [0, 1)$ sufficiently close to 1, the sequence

$$(\underbrace{x^*(0), \ldots, x^*(T)}_{\text{first } T \text{ terms as in } x^*}, \underbrace{\lambda x^*(T+1), \lambda x^*(T+2), \ldots}_{\text{remaining terms times } \lambda}) \tag{171}$$

is feasible:

☒ for $0 \leq t \leq T-1$, feasibility of x^* gives $(x^*(t), x^*(t+1)) \in X(t)$.

- ⊠ for $t \geq T+1$, feasibility of x^* gives $(x^*(t), x^*(t+1)) \in X(t)$. Assumption (A3) gives $(\mathbf{0}, \mathbf{0}) \in X(t)$ and by convexity that $(\lambda x^*(t), \lambda x^*(t+1)) \in X(t)$ for *all* $\lambda \in [0,1)$.
- ⊠ for $t = T$, since $(x^*(T), x^*(T+1))$ lies in the interior of $X(T)$, so does every vector sufficiently nearby. In particular, $(x^*(T), \lambda x^*(T+1))$ lies in $X(T)$ for $\lambda \in [0,1)$ sufficiently close to 1.

Since x^* is optimal and the sequence (171) is feasible, their payoff difference satisfies

$$\begin{aligned}
&f(T, x^*(T), \lambda x^*(T+1)) - f(T, x^*(T), x^*(T+1)) \\
&\qquad\qquad + \sum_{t=T+1}^{\infty} \big(f(t, \lambda x^*(t), \lambda x^*(t+1)) - f(t, x^*(t), x^*(t+1))\big) \leq 0.
\end{aligned}$$

By concavity (A4) of $f(t, \cdot, \cdot)$ on the convex control region (A3), we have

$$f(t, \lambda x^*(t), \lambda x^*(t+1)) \geq \lambda f(t, x^*(t), x^*(t+1)) + (1-\lambda) f^*(t, \mathbf{0}, \mathbf{0}),$$

so that

$$\frac{f(t, x^*(t), x^*(t+1)) - f(t, \lambda x^*(t), \lambda x^*(t+1))}{1-\lambda} \leq f(t, x^*(t), x^*(t+1)) - f(t, \mathbf{0}, \mathbf{0}).$$

Dividing the payoff difference by $1-\lambda$ and rearranging terms, we find

$$\begin{aligned}
&\frac{f(T, x^*(T), \lambda x^*(T+1)) - f(T, x^*(T), x^*(T+1))}{1-\lambda} \\
&\qquad \leq \frac{\sum_{t=T+1}^{\infty} \big(f(t, x^*(t), x^*(t+1)) - f(t, \lambda x^*(t), \lambda x^*(t+1))\big)}{1-\lambda} \\
&\qquad \leq \sum_{t=T+1}^{\infty} \big(f(t, x^*(t), x^*(t+1)) - f(t, \mathbf{0}, \mathbf{0})\big).
\end{aligned}$$

If we let λ increase to 1 in the first term, we recognize it as the directional derivative of $f^*(T, x^*(T), \cdot)$ at $x^*(T+1)$ in direction $-x^*(T+1)$, which exists by (A4). Using that $x^*(T+1) \geq \mathbf{0}$ by (A2) and $f_3'(T, x^*(T), x^*(T+1)) \leq \mathbf{0}$ by (A5), we find

$$0 \leq -f_3'(T, x^*(T), x^*(T+1)) x^*(T+1) \leq \sum_{t=T+1}^{\infty} \big(f(t, x^*(t), x^*(t+1)) - f(t, \mathbf{0}, \mathbf{0})\big).$$

Since the final term goes to zero as $T \to \infty$, we must have (169).

SUFFICIENCY (168) AND (169): We show that the objective function at x^* is at least as high as at any feasible $x = (x(t))_{t \in \mathbb{Z}}$. For the first $T \in \mathbb{N}$ terms, a linear approximation using Theorem 17.11 gives

$$\begin{aligned}
&\sum_{t=0}^{T} \big[f(t, x^*(t), x^*(t+1)) - f(t, x(t), x(t+1))\big] \\
&\quad \geq \sum_{t=0}^{T} \big[f_2'(t, x^*(t), x^*(t+1))(x^*(t) - x(t)) + f_3'(t, x^*(t), x^*(t+1))(x^*(t+1) - x(t+1))\big] \\
&\quad = f_2'(0, x^*(0), x^*(1)) \underbrace{(x^*(0) - x(0))}_{=0 \text{ since } x^*(0) = x(0) = x_0} \\
&\qquad + \sum_{t=0}^{T-1} \underbrace{\big[f_3'(t, x^*(t), x^*(t+1)) + f_2'(t+1, x^*(t+1), x^*(t+2))\big]}_{=0 \text{ by (168)}} (x^*(t+1) - x(t+1)) \\
&\qquad + \underbrace{f_3'(T, x^*(T), x^*(T+1))}_{\leq \mathbf{0} \text{ by (A5)}} (x^*(T+1) - \underbrace{x(T+1)}_{\geq \mathbf{0} \text{ by (A2)}}) \\
&\quad \geq f_3'(T, x^*(T), x^*(T+1)) x^*(T+1).
\end{aligned}$$

Let $T \to \infty$ and use (169):

$$\sum_{t=0}^{\infty} \left[f(t, x^*(t), x^*(t+1)) - f(t, x(t), x(t+1)) \right] \geq \lim_{T\to\infty} f_3'(T, x^*(T), x^*(T+1)) x^*(T+1) = 0,$$

showing that x^* is optimal. □

By the Euler equation, the transversality condition can be rewritten as

$$\lim_{t\to\infty} f_2'(t+1, x^*(t+1), x^*(t+2)) x^*(t+1) = 0.$$

Changing the time index, this reduces to

$$\lim_{t\to\infty} f_2'(t, x^*(t), x^*(t+1)) x^*(t) = 0. \tag{172}$$

In the important special case of payoffs discounted by some $\beta \in (0,1]$, the goal function is of the form

$$\sum_{t=0}^{\infty} \beta^t f(t, x(t), x(t+1)),$$

and the Euler equations become:

$$\text{for each } t \in \mathbb{Z}_+: \qquad f_3'(t, x^*(t), x^*(t+1)) + \beta f_2'(t+1, x^*(t+1), x^*(t+2)) = \mathbf{0}^\top, \tag{173}$$

and the transversality condition becomes:

$$\lim_{t\to\infty} -\beta^t f_3'(t, x^*(t), x^*(t+1)) x^*(t+1) = 0,$$

or, using (172),

$$\lim_{t\to\infty} \beta^t f_2'(t, x^*(t), x^*(t+1)) x^*(t) = 0. \tag{174}$$

Example 24.1 A standard example in macroeconomic growth theory concerns a one-sector growth model of the form

$$\begin{array}{lll} \text{maximize} & \sum_{t=0}^{\infty} \beta^t \ln(x(t)^\alpha - x(t+1)) & \\ \text{with} & 0 \leq x(t+1) \leq x(t)^\alpha & t \in \mathbb{Z}_+ \\ & x(0) = x_0 > 0 \text{ (given)}, & \end{array}$$

with $\alpha, \beta \in (0,1)$ and $\ln 0 = -\infty$. For each $t \in \mathbb{Z}_+$, the constraints imply that $\ln x(t+1) \leq \alpha \ln x(t)$. By induction, $\ln x(t) \leq \alpha^t \ln x_0$. Hence,

$$\begin{aligned} \sum_{t=0}^{\infty} \beta^t \ln\left(x(t)^\alpha - x(t+1)\right) &\leq \sum_{t=0}^{\infty} \beta^t \ln x(t)^\alpha \\ &\leq \sum_{t=0}^{\infty} \alpha\beta^t \ln x(t) \\ &\leq \sum_{t=0}^{\infty} \alpha\left(\alpha\beta\right)^t \ln x_0 \to \frac{\alpha \ln x_0}{1-\alpha\beta}, \end{aligned}$$

making the goal function well-defined in $\mathbb{R} \cup \{-\infty\}$ by Theorem 22.1. Not all assumptions preceding Theorem 24.1 are satisfied, but it can be shown with minor adjustments that there is an interior optimum satisfying the Euler equations and the transversality condition. With $f(t, x(t), x(t+1)) = \ln(x(t)^\alpha - x(t+1))$, the Euler equation (173) requires that for each $t \in \mathbb{Z}_+$:

$$\frac{-1}{x^*(t)^\alpha - x^*(t+1)} + \frac{\alpha\beta x^*(t+1)^{\alpha-1}}{x^*(t+1)^\alpha - x^*(t+2)} = 0,$$

and the transversality condition (174) is

$$\lim_{t\to\infty}\beta^t\frac{\alpha x^*(t)^{\alpha-1}}{x^*(t)^\alpha - x^*(t+1)}x^*(t) = 0.$$

This optimum is given by the feasible sequence x^* with $x^*(0) = x_0$ and $x^*(t+1) = \alpha\beta x^*(t)^\alpha$. The Euler equation is satisfied since

$$\begin{aligned}\frac{-1}{x^*(t)^\alpha - x^*(t+1)} + \frac{\alpha\beta x^*(t+1)^{\alpha-1}}{x^*(t+1)^\alpha - x^*(t+2)} &= \frac{-1}{x^*(t)^\alpha - \alpha\beta x^*(t)^\alpha} + \frac{\alpha\beta x^*(t+1)^{\alpha-1}}{x^*(t+1)^\alpha - \alpha\beta x^*(t+1)^\alpha}\\ &= \frac{-1}{(1-\alpha\beta)x^*(t)^\alpha} + \frac{\alpha\beta}{(1-\alpha\beta)x^*(t+1)}\\ &= \frac{-1}{(1-\alpha\beta)x^*(t)^\alpha} + \frac{\alpha\beta}{(1-\alpha\beta)\alpha\beta x^*(t)^\alpha}\\ &= 0.\end{aligned}$$

The transversality condition is satisfied, since

$$\begin{aligned}\lim_{t\to\infty}\beta^t\frac{\alpha x^*(t)^{\alpha-1}}{x^*(t)^\alpha - x^*(t+1)}x^*(t) &= \lim_{t\to\infty}\beta^t\frac{\alpha x^*(t)^\alpha}{x^*(t)^\alpha - \alpha\beta x^*(t)^\alpha}\\ &= \lim_{t\to\infty}\beta^t\frac{\alpha}{1-\alpha\beta}\\ &= 0.\end{aligned}$$ ◁

A Some prerequisites

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 $x + y = 0$. Indeed, it can be shown (along the lines of the proof of Theorem 1.1) 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 $y \in F$ such that $xy = 1$. We denote $y = x^{-1}$. Division is now defined 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

$$(a + bi)(c + di) = (ac - bd) + (ad + bc)i. \qquad \triangleleft$$

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:

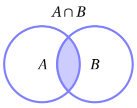
- ⊠ by explicitly listing its elements: $S = \{1, 2, 3, 4, 5, 6\}$ or $S = \{1, \ldots, 6\}$, where '…' is used if it ought to be clear from the context what the other elements are;
- ⊠ by some characterizing property P, like $S = \{x : x \text{ satisfies } P\}$. For instance,

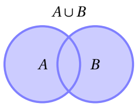
$$S = \{x : x \text{ is a positive integer}, x \leq 6\}.$$

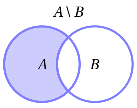
We write $x \in S$ to denote that x belongs to/is an element of set S and $x \notin S$ to denote that x does not belong to S. Other standard notation we will use:

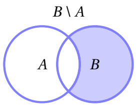
notation	name	characterizing property
$\emptyset$	the empty set	the set that contains no elements
$A \subseteq B$	A is a subset of B	each element of A belongs to B : $a \in A \Rightarrow a \in B$
$A = B$	sets A and B are equal	$A \subseteq B$ and $B \subseteq A$
$A \subset B$	A is a proper subset of 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:

$$A^c = \{x \in X : x \notin A\} = X \setminus A.$$

Notice that the notation $X \setminus A$ makes it apparent with respect to which set the complement is taken, whereas this is supposed to be clear from the context if we write A^c. If one considers the union/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 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 Ordered pairs and Cartesian products

Let a and b be distinct real numbers. The sets $\{a,b\}$ and $\{b,a\}$ are equal, because they contain the same elements; we just wrote them in a different order. Often, however, it is convenient to consider distinct elements in a particular order: in planar geometry, for instance, the coordinates (x,y) of a point represent an ordered pair of numbers: the point $(1,2)$ is different from the point $(2,1)$, whereas the set $\{1,2\}$ is the same as the set $\{2,1\}$. Whenever we wish to stress the order of two elements, we use round parentheses (and) and write (a,b) for the ordered pair with a as its first and b as its second element.

Given two sets A and B, we define the Cartesian product $A \times B = \{(a,b) : a \in A, b \in B\}$ as the set of ordered pairs (a,b) whose first element is from A and whose second element is from B.

Example A.6 If $A = \{x,y\}$ and $B = \{1,2\}$, then

$$A \times B = \{(x,1),(x,2),(y,1),(y,2)\} \qquad \text{but} \qquad B \times A = \{(1,x),(1,y),(2,x),(2,y)\}. \qquad \triangleleft$$

A.4 Functions

Let X, Y be nonempty sets. A function $f : X \to Y$ assigns to each $x \in X$ a unique element $f(x) \in Y$. Functions are also called maps, mappings, or transformations. The graph of f 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 x on the horizontal and $y = f(x)$ on the vertical axis. If the 'name' f of the function is not important, but its description is, we sometimes use notation like $x \mapsto \sin x$ to describe the function. X is the domain of f, Y the codomain, and $f(X) = \{f(x) : x \in X\}$ the range of f. The set of all functions from X to Y is sometimes denoted as Y^X.

A.5 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.1 (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 Suggested solutions

These are (sometimes short) solutions to exercises in the lecture notes. In solutions to the home assignments and exam questions, you are expected to start from relevant definitions and clearly deduce and motivate your answers. Suggestions for improvements (and corrections of potential mistakes) are welcome!

1.1 Looking at Definition 1.1 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 1.1. 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.

1.2 (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 1.2, 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 1.2, 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 1.2, 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 1.2, W is a subspace of $\mathbb{R}^{\mathbb{N}}$.

1.3 (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 1.2, 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 1.2, it cannot be a subspace.

1.4

- ☒ $\{\mathbf{0}\}$ 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 1.1(e).
- ☒ By Theorem 1.2, $\{\mathbf{0}\}$ is a subspace of V.

1.5 (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

$$(\alpha(x+y))_i = \alpha(x+y)_i = \alpha(x_i + y_i) \overset{(V7)}{=} \alpha x_i + \alpha y_i = (\alpha x)_i + (\alpha y)_i = (\alpha x + \alpha y)_i.$$

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

$$((\alpha+\beta)x)_i = (\alpha+\beta)x_i \overset{(V8)}{=} \alpha x_i + \beta x_i = (\alpha x)_i + (\beta x)_i = (\alpha x + \beta x)_i.$$

Hence, $(\alpha+\beta)x = \alpha x + \beta x$.

(b)

Example	I	V_i
1.1	$\{(i,j) : i \in \{1,2\}, j \in \{1,2,3\}\}$	$\mathbb{R}$
1.2	$\mathbb{R}$	$\mathbb{R}$
1.3	$\{1,\ldots,n\}$	$\mathbb{R}$
1.4	$\{(i,j) : i \in \{1,\ldots,m\}, j \in \{1,\ldots,n\}\}$	$\mathbb{R}$
1.5	$\mathbb{N}$	$\mathbb{R}$

1.6 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

$$\begin{aligned} p(x) &= a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0, \\ q(x) &= b_n x^n + b_{n-1}x^{n-1} + \cdots + b_1 x + b_0, \\ r(x) &= c_n x^n + c_{n-1}x^{n-1} + \cdots + c_1 x + c_0 \end{aligned}$$

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

$$\begin{aligned}(p+\mathbf{0})(x) &= (a_n+0)x^n + (a_{n-1}+0)x^{n-1} + \cdots + (a_1+0)x + (a_0+0)\\ &= a_nx^n + a_{n-1}x^{n-1} + \cdots + a_1x + a_0\\ &= p(x),\end{aligned}$$

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

$$\begin{aligned}((\alpha\beta)p)(x) &= (\alpha\beta)a_nx^n + (\alpha\beta)a_{n-1}x^{n-1} + \cdots + (\alpha\beta)a_1x + (\alpha\beta)a_0\\ &= \alpha(\beta a_n)x^n + \alpha(\beta a_{n-1})x^{n-1} + \cdots + \alpha(\beta a_1)x + \alpha(\beta a_0)\\ &= (\alpha(\beta p))(x),\end{aligned}$$

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

$$\begin{aligned}(1p)(x) &= (1a_n)x^n + (1a_{n-1})x^{n-1} + \cdots + (1a_1)x + (1a_0)\\ &= a_nx^n + a_{n-1}x^{n-1} + \cdots + a_1x + a_0\\ &= p(x),\end{aligned}$$

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

$$\begin{aligned}((\alpha+\beta)p)(x) &= (\alpha+\beta)a_nx^n + (\alpha+\beta)a_{n-1}x^{n-1} + \cdots + (\alpha+\beta)a_1x + (\alpha+\beta)a_0\\ &= (\alpha a_n+\beta a_n)x^n + (\alpha a_{n-1}+\beta a_{n-1})x^{n-1} + \cdots + (\alpha a_1+\beta a_1)x + (\alpha a_0+\beta a_0)\\ &= (\alpha p+\beta p)(x),\end{aligned}$$

showing that $(\alpha+\beta)p = \alpha p + \beta p$.

1.7 (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 1.2. 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 1.1(f), $-x = (-1)x \in W$: the additive inverse of x lies in W.

(b) Let W be a subspace of V. By Theorem 1.2, 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 $\beta = 0$).

2.1 Recall from the text following Definition 2.2 that a finite set $W = \{w_1, \ldots, w_n\}$ of different vectors is linearly independent if and only if

$$\alpha_1 w_1 + \cdots + \alpha_n w_n = \mathbf{0} \tag{175}$$

implies that $\alpha_1 = \cdots = \alpha_n = 0$. So we solve equation (175): 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 (175) 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 (175) 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 (175) 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 (175) 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.

2.2 (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 1.2, 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 $\operatorname{span}(W)$.

2.3 ⊠ Assume W is linearly dependent. Then W is nonempty by Definition 2.2. 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 $w, w_1, \dots, w_n$ in W such that w is a linear combination of $w_1, \dots, w_n$. In the first case, W is linearly dependent, since $\alpha\mathbf{0} = \mathbf{0}$ for each nonzero scalar α by Theorem 1.1(e). In the second case, there are scalars $\alpha_1, \dots, \alpha_n$ 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.

3.1 (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 $\{1, \dots, n\}$ 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 $\{1, \dots, n\}$. 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 $\{1, \dots, k, k+1\}$ 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 $\mathscr{A}$.

4.1 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

$$\begin{aligned} T(e_1) &= T(1,1) - \frac{1}{2}T(0,2) = (2,0,-3) - \frac{1}{2}(-1,4,2) = \left(\frac{5}{2}, -2, -4\right), \\ T(e_2) &= \frac{1}{2}T(0,2) = \frac{1}{2}(-1,4,2) = \left(-\frac{1}{2}, 2, 1\right). \end{aligned}$$

Conclude:

$$A = \begin{bmatrix} T(e_1) & T(e_2) \end{bmatrix} = \begin{bmatrix} \frac{5}{2} & -\frac{1}{2} \\ -2 & 2 \\ -4 & 1 \end{bmatrix}$$

4.2 (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.):

$$\left[\begin{array}{rrrrr|r} 1 & 4 & 5 & 6 & 9 & 0 \\ 3 & -2 & 1 & 4 & -1 & 0 \\ 1 & 0 & -1 & -2 & -1 & 0 \\ 2 & 3 & 5 & 7 & 8 & 0 \end{array}\right] \sim \left[\begin{array}{rrrrr|r} 1 & 4 & 5 & 6 & 9 & 0 \\ 0 & -14 & -14 & -14 & -28 & 0 \\ 0 & -4 & -6 & -8 & -10 & 0 \\ 0 & -5 & -5 & -5 & -10 & 0 \end{array}\right] \sim \left[\begin{array}{rrrrr|r} 1 & 4 & 5 & 6 & 9 & 0 \\ 0 & 1 & 1 & 1 & 2 & 0 \\ 0 & 0 & -2 & -4 & -2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right]$$

$$\sim \left[\begin{array}{rrrrr|r} 1 & 4 & 0 & -4 & 4 & 0 \\ 0 & 1 & 0 & -1 & 1 & 0 \\ 0 & 0 & 1 & 2 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right] \sim \left[\begin{array}{rrrrr|r} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & -1 & 1 & 0 \\ 0 & 0 & 1 & 2 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right]$$

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 2.2). 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 2.1, 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.

4.4 ☒ To see that $L(V,W)$ is a subspace of the functions from V to W, we apply Theorem 1.2:

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:

$$\begin{aligned} (T_1 + T_2)(x+y) &= T_1(x+y) + T_2(x+y) && \text{(by def. of } T_1 + T_2\text{)} \\ &= T_1(x) + T_1(y) + T_2(x) + T_2(y) && \text{(by linearity of } T_1 \text{ and } T_2\text{)} \\ &= T_1(x) + T_2(x) + T_1(y) + T_2(y) && \text{(after rearranging terms)} \\ &= (T_1 + T_2)(x) + (T_1 + T_2)(y) && \text{(by def. of } T_1 + T_2\text{)} \end{aligned}$$

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 1.2, $L(V,W)$ is a subspace of the set of functions from V to W.

☒ Recall that $0x = \mathbf{0}$ for every vector x (Theorem 1.1(d)). Since T is linear:

$$T\mathbf{0} = T0\mathbf{0} = 0T\mathbf{0} = \mathbf{0}.$$

☒ We use Theorem 1.2 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 1.2 that $T(U)$ is a subspace of W.

☒ We use Theorem 1.2 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 1.2 that the null space of T is a subspace of V.

5.1 (a) For each coordinate j:

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

Hence,

$$\|x\|_2 \geq \max\{|x_1|,\dots,|x_n|\} = \|x\|_\infty.$$

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)

$$\|x\|_\infty = \max\{|x_1|,\ldots,|x_n|\} \le \sum_{i=1}^n |x_i| = \|x\|_1,$$

$$\|x\|_1 = \sum_{i=1}^n |x_i| \le \sum_{i=1}^n \max\{|x_1|,\ldots,|x_n|\} = n\|x\|_\infty.$$

(c) Write $x = \sum_{i=1}^n x_i e_i$ as a linear combination of standard basis vectors and use the triangle inequality:

$$\|x\|_2 = \left\|\sum_{i=1}^n x_i e_i\right\| \le \sum_{i=1}^n \|x_i e_i\| = \sum_{i=1}^n |x_1|\underbrace{\|e_i\|}_{=1} = \sum_{i=1}^n |x_i| = \|x\|_1.$$

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:

$$\|x\|_1 = \sum_{i=1}^n |x_i| = \langle(|x_1|,\ldots,|x_n|),(1,\ldots,1)\rangle \le \|(|x_1|,\ldots,|x_n|)\|\|(1,\ldots,1)\| = \sqrt{n}\|x\|_2.$$

5.2 Let $x, y \in V$. Rewriting (note: $|x| \le \varepsilon \iff -\varepsilon \le x \le \varepsilon$), we need to prove

$$-\|x-y\| \le \|x\| - \|y\| \le \|x-y\|.$$

The first inequality follows from

$$\|y\| = \|x-(x-y)\| \overset{\text{(N4)}}{\le} \|x\| + \|-(x-y)\| \overset{\text{(N3)}}{=} \|x\| + |-1|\|x-y\| = \|x\| + \|x-y\|.$$

Similarly, for the second inequality:

$$\|x\| = \|y+(x-y)\| \le \|y\| + \|x-y\|.$$

5.3 Proof that Example 5.2 is a normed vector space: the norm is generated by an inner product, so this follows from Theorem 5.1 and the discussion preceding it.

Proof that Example 5.3 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

$$\|x+y\|_1 = \sum_{i=1}^n |x_i + y_i| \le \sum_{i=1}^n \big(|x_i| + |y_i|\big) = \sum_{i=1}^n |x_i| + \sum_{i=1}^n |y_i| = \|x\|_1 + \|y\|_1.$$

Proof that Example 5.4 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 5.5 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 5.6(c).

PROOF THAT EXAMPLE 5.6 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 5.4.

PROOF THAT EXAMPLE 5.7 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 13.3.) 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 5.8 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

$$\|f+g\|_1 = \int_a^b |f(x) + g(x)| \, \mathrm{d}x \le \int_a^b |f(x)| + |g(x)| \, \mathrm{d}x = \int_a^b |f(x)| \, \mathrm{d}x + \int_a^b |g(x)| \, \mathrm{d}x = \|f\|_1 + \|g\|_1.$$

5.4 We have

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

Subtracting the second equation from the first gives the polarization identity; adding them gives the parallelogram law.

5.5 (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.

5.6 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

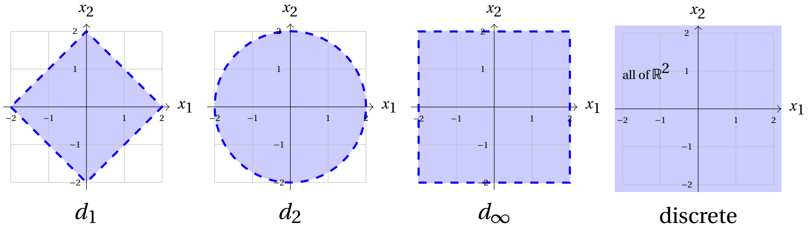
$$\sum_{i=1}^{n} (|x_i| + |y_i|)^p \leq \left(\left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p} + \left(\sum_{i=1}^{n} |y_i|^p\right)^{1/p}\right)^p.$$

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 $\|x+y\|_p \leq \|x\|_p + \|y\|_p$.

6.1 The requested pictures are:

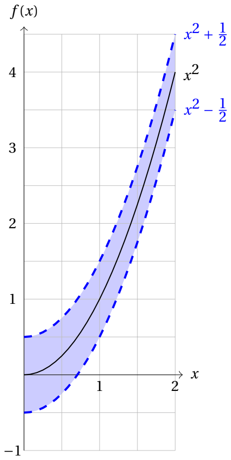


d_1 d_2 d_∞ discrete

6.2

$$\begin{aligned} d_1(x,y) &= |x_1-y_1|+|x_2-y_2|+|x_3-y_3| = |1-2|+|0-6|+|4-2| = 1+6+2=9,\\ d_2(x,y) &= \sqrt{(x_1-y_1)^2+(x_2-y_2)^2+(x_3-y_3)^2} = \sqrt{(1-2)^2+(0-6)^2+(4-2)^2} = \sqrt{1+36+4} = \sqrt{41},\\ d_\infty(x,y) &= \max\{|x_1-y_1|,|x_2-y_2|,|x_3-y_3|\} = \max\{|1-2|,|0-6|,|4-2|\} = \max\{1,6,2\} = 6,\\ d_{\text{discrete}}(x,y) &= 1 \text{ since } x \neq y. \end{aligned}$$

6.3 All functions with a graph in the shaded area:



6.4 (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:

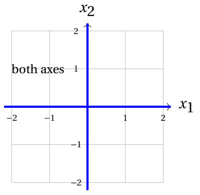
$$\{i : x_i \neq z_i\} \subseteq \{i : x_i \neq y_i\} \cup \{i : y_i \neq z_i\}.$$

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 $d(\alpha x, \alpha y) = \|\alpha x - \alpha y\| = \|\alpha(x-y)\| \overset{(N3)}{=} |\alpha|\|x-y\| = \alpha d(x,y)$.

(d)



Hamming

6.5 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 $d(x,y) \geq 0$.

6.6 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 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.

6.7 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,

$$\begin{aligned}\frac{d(x,z)}{1+d(x,z)} &= 1-\frac{1}{1+d(x,z)} \leq 1-\frac{1}{1+d(x,y)+d(y,z)}\\ &= \frac{d(x,y)+d(y,z)}{1+d(x,y)+d(y,z)} = \frac{d(x,y)}{1+d(x,y)+d(y,z)} + \frac{d(y,z)}{1+d(x,y)+d(y,z)}\\ &\leq \frac{d(x,y)}{1+d(x,y)} + \frac{d(y,z)}{1+d(y,z)}.\end{aligned}$$

6.8 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

$$d(y,z) \overset{\text{(I4)}}{\leq} d(y,x)+d(x,z) \overset{\text{(I3)}}{=} d(x,y)+d(x,z),$$

and the second from

$$d(x,y) \overset{\text{(I4)}}{\leq} d(x,z)+d(z,y) \overset{\text{(I3)}}{=} d(x,z)+d(y,z).$$

6.9 (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 5.1:

(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 $x,y \in V$: $\|x-y\| = d(x-y,\mathbf{0}) \overset{(a)}{=} d((x-y)+y, \mathbf{0}+y) = d(x,y)$.

6.10 (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

$$\begin{aligned}
d(A,B) + d(B,C) &= \frac{|A\cup B| - |A\cap B|}{|A\cup B|} + \frac{|B\cup C| - |B\cap C|}{|B\cup C|} && (176)\\
&\geq \frac{|T_1| + |T_2|}{|A\cup B\cup C|} + \frac{|T_2| + |T_3|}{|A\cup B\cup C|} && (177)\\
&\geq \frac{|T_1| + |T_2| + |T_3|}{|A\cup B\cup C|} && (178)\\
&= 1 - \frac{|V|}{|A\cup B\cup C|} && (179)\\
&\geq 1 - \frac{|A\cap C|}{|A\cup C|} && (180)\\
&= d(A,C). && (181)
\end{aligned}$$

☒ Equalities (176) and (181) 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 (176) to (177), 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.

☒ (177) is at least as large as (178) because

$$\frac{|T_1| + |T_2|}{|A\cup B\cup C|} + \frac{|T_2| + |T_3|}{|A\cup B\cup C|} = \frac{|T_1| + 2|T_2| + |T_3|}{|A\cup B\cup C|} \geq \frac{|T_1| + |T_2| + |T_3|}{|A\cup B\cup C|}.$$

☒ (178) equals (179) 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 $|T_1| + |T_2| + |T_3| = |A\cup B\cup C| - |V|$.

☒ (179) is at least as large as (180), 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.

7.1

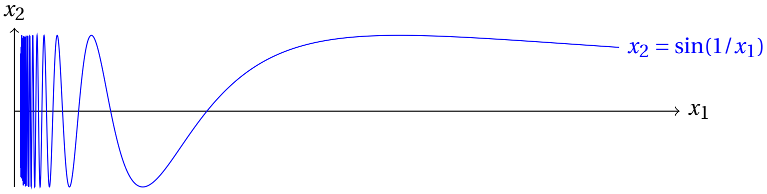
U	int(U)	cl(U)	bd(U)	acc(U)
$\{x : x_1 > 0\}$	$\{x : x_1 > 0\}$	$\{x : x_1 \geq 0\}$	$\{x : x_1 = 0\}$	$\{x : x_1 \geq 0\}$
$\{x : x_1^2 + x_2^2 = 4\}$	$\emptyset$	U	U	U
$\{x : x_1 \leq x_2\}$	$\{x : x_1 < x_2\}$	$\{x : x_1 \leq x_2\}$	$\{x : x_1 = x_2\}$	$\{x : x_1 \leq x_2\}$
$\{x : x_1 > 0, x_2 = \sin(1/x_1)\}$	$\emptyset$	$U \cup \{(0, x_2) : -1 \leq x_2 \leq 1\}$	$=$ cl(U)	$=$ 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

$$\left[\frac{1}{2(2\pi)}, \frac{1}{2\pi}\right] \text{ and } \left[\frac{1}{3(2\pi)}, \frac{1}{2(2\pi)}\right] \text{ and } \left[\frac{1}{4(2\pi)}, \frac{1}{3(2\pi)}\right] \text{ and}\ldots$$

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.

7.2 (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 7.5.

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 7.5.

(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.

7.4 (a) METHOD 1: Using Theorem 7.4:

$$\begin{aligned} x \in \mathrm{bd}(U) &\overset{\text{Thm 7.4(e)}}{\Leftrightarrow} && \forall \varepsilon > 0: \quad B(x,\varepsilon) \cap U \neq \emptyset \text{ and } B(x,\varepsilon) \cap U^c \neq \emptyset, \\ &\overset{\text{Thm 7.4(a),(b)}}{\Leftrightarrow} && x \in \mathrm{cl}(U), x \notin \mathrm{int}(U) \\ &\overset{\text{def}}{\Leftrightarrow} && x \in \mathrm{cl}(U) \setminus \mathrm{int}(U). \end{aligned}$$

METHOD 2: Using Exercise 7.9:

$$\mathrm{bd}(U) \overset{\text{def}}{=} \mathrm{cl}(U) \cap \mathrm{cl}(U^c) \overset{\text{Exc 7.9(b)}}{=} \mathrm{cl}(U) \cap \mathrm{int}(U)^c \overset{\text{def}}{=} \mathrm{cl}(U) \setminus \mathrm{int}(U).$$

(b) We show that $\operatorname{acc}(U)^c$ is open. Let $v \in \operatorname{acc}(U)^c$. Then there is an $\varepsilon > 0$ with $(B(v,\varepsilon) \setminus \{v\}) \cap U = \emptyset$. But then v is an interior point of $\operatorname{acc}(U)^c$, because all points in $B(v,\varepsilon)$ lie in $\operatorname{acc}(U)^c$: for each $w \in B(v,\varepsilon)$ with $w \neq v$, choose $0 < \delta < \min\{\varepsilon - d(v,w), d(v,w)\}$. By example 7.2, $B(w,\delta) \subseteq B(v,\varepsilon)$. 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., $w \in \operatorname{acc}(U)^c$.

7.4 Call the metric space (X,d). The empty set has zero elements; it is closed by Theorem 7.3. Next, consider a subset $\{x_1,\ldots,x_n\}$ of $n \in \mathbb{N}$ elements. Why is it closed?
METHOD 1: By Example 7.3, a set of one element is closed. Since the union of finitely many closed sets is closed (Theorem 7.3), 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 $\{x_1,\ldots,x_n\}^c$.

7.5 (a) By Theorem 7.1, 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 7.1.

(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 7.1.

7.6 (a) $\emptyset$ is a closed set and $\operatorname{cl}(\emptyset)$ is the smallest closed set containing $\emptyset$, so $\operatorname{cl}(\emptyset) = \emptyset$.

(b) $\operatorname{cl}(U)$ is a closed set containing U, so $U \subseteq \operatorname{cl}(U)$.

(c) Since $\operatorname{cl}(U)$ is closed, it is the smallest closed set containing itself: $\operatorname{cl}(\operatorname{cl}(U)) = \operatorname{cl}(U)$.

(d) By (b), $A \subseteq B \subseteq \operatorname{cl}(B)$, so $\operatorname{cl}(B)$ is a closed set containing A. Set $\operatorname{cl}(A)$ is the smallest closed set containing A, so $\operatorname{cl}(A) \subseteq \operatorname{cl}(B)$.

(e) Since $A \subseteq A \cup B$, (d) gives $\operatorname{cl}(A) \subseteq \operatorname{cl}(A \cup B)$. Likewise, $\operatorname{cl}(B) \subseteq \operatorname{cl}(A \cup B)$. So $\operatorname{cl}(A) \cup \operatorname{cl}(B) \subseteq \operatorname{cl}(A \cup B)$. For the converse inclusion, (b) gives $A \subseteq \operatorname{cl}(A)$ and $B \subseteq \operatorname{cl}(B)$, so $A \cup B \subseteq \operatorname{cl}(A) \cup \operatorname{cl}(B)$. As the union of two closed sets, $\operatorname{cl}(A) \cup \operatorname{cl}(B)$ is a closed set containing $A \cup B$. Set $\operatorname{cl}(A \cup B)$ is the smallest closed set containing $A \cup B$, so $\operatorname{cl}(A \cup B) \subseteq \operatorname{cl}(A) \cup \operatorname{cl}(B)$.

7.7 The proofs proceed along the same lines as those in exercise 7.6:

(a) $\emptyset$ is an open set and $\operatorname{int}(\emptyset)$ is the largest open set contained in $\emptyset$, so $\operatorname{int}(\emptyset) = \emptyset$.

(b) $\operatorname{int}(U)$ is an open set contained in U, so $\operatorname{int}(U) \subseteq U$.

(c) Since $\operatorname{int}(U)$ is open, it is the largest open set containing itself: $\operatorname{int}(\operatorname{int}(U)) = \operatorname{int}(U)$.

(d) By (b), $\operatorname{int}(A) \subseteq A \subseteq B$, so $\operatorname{int}(A)$ is an open set contained in B. Set $\operatorname{int}(B)$ is the largest open set contained in B, so $\operatorname{int}(A) \subseteq \operatorname{int}(B)$.

(e) Since $U \cap V \subseteq U$, (d) gives $\operatorname{int}(U \cap V) \subseteq \operatorname{int}(U)$. Likewise, $\operatorname{int}(U \cap V) \subseteq \operatorname{int}(V)$. So $\operatorname{int}(U \cap V) \subseteq \operatorname{int}(U) \cap \operatorname{int}(V)$. For the converse inclusion, (b) gives $\operatorname{int}(U) \subseteq U$ and $\operatorname{int}(V) \subseteq V$, so $\operatorname{int}(U) \cap \operatorname{int}(V) \subseteq U \cap V$. As the intersection of two open sets, $\operatorname{int}(U) \cap \operatorname{int}(V)$ is an open set contained in $U \cap V$. Set $\operatorname{int}(U \cap V)$ is the largest open set contained in $U \cap V$, so $\operatorname{int}(U) \cap \operatorname{int}(V) \subseteq \operatorname{int}(U \cap V)$.

7.8 (a) Take $U = (0,1), V = (1,2)$. Then $\operatorname{cl}(U \cap V) = \operatorname{cl}(\emptyset) = \emptyset$, but $\operatorname{cl}(U) \cap \operatorname{cl}(V) = [0,1] \cap [1,2] = \{1\}$.

(b) Take $U = (-\infty,0)$. Then $\operatorname{cl}(U^c) = \operatorname{cl}[0,\infty) = [0,\infty)$, but $\operatorname{cl}(U)^c = (-\infty,0]^c = (0,\infty)$.

(c) Take $U = (0,1), V = [1,2)$. Then $\operatorname{int}(U \cup V) = \operatorname{int}((0,2)) = (0,2)$, but $\operatorname{int}(U) \cup \operatorname{int}(V) = (0,1) \cup (1,2)$.

(d) Take $U = (-\infty,0)$. Then $\operatorname{int}(U^c) = (0,\infty)$, but $\operatorname{int}(U)^c = [0,\infty)$.

7.9 (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:

$$\begin{aligned}
\operatorname{cl}(U)^c &= \left(\cap_{W \supseteq U, W \text{ is closed}} W\right)^c && \text{(by definition)}\\
&= \cup_{W \supseteq U, W \text{ is closed}} W^c && \text{(De Morgan's Law)}\\
&= \cup_{W^c \subseteq U^c, W^c \text{ is open}} W^c && \\
&= \cup_{V \subseteq U^c, V \text{ is open}} V && \text{(writing } V = W^c\text{)}\\
&= \operatorname{int}(U^c) && \text{(by definition)}
\end{aligned}$$

(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:

$$\begin{aligned}
\operatorname{int}(U)^c &= \left(\cup_{W \subseteq U, W \text{ is open}} W\right)^c && \text{(by definition)} \\
&= \cap_{W \subseteq U, W \text{ is open}} W^c && \text{(De Morgan's Law)} \\
&= \cup_{W^c \supseteq U^c, W^c \text{ is closed}} W^c && \\
&= \cup_{V \supseteq U^c, V \text{ is closed}} V && \text{(writing } V = W^c\text{)} \\
&= \operatorname{cl}(U^c) && \text{(by definition)}
\end{aligned}$$

7.10 It suffices to prove the chain of implications (a) $\Rightarrow$ (b) $\Rightarrow$ (c) $\Rightarrow$ (d) $\Rightarrow$ (a). In principle, I prove things straight from their definitions, but I will also provide some alternative proofs by invoking Theorem 7.4.

(a) $\Rightarrow$ (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) $\Rightarrow$ (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 7.4, $U = \operatorname{cl}(U) = U \cup \operatorname{bd}(U)$. Equality $U = U \cup \operatorname{bd}(U)$ implies $\operatorname{bd}(U) \subseteq U$.

(c) $\Rightarrow$ (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 7.4, $U = U \cup \operatorname{bd}(U) = \operatorname{cl}(U) = U \cup \operatorname{acc}(U)$. Equality $U = U \cup \operatorname{acc}(U)$ implies $\operatorname{acc}(U) \subseteq U$.

(d) $\Rightarrow$ (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 7.4, $U = U \cup \operatorname{acc}(U) = \operatorname{cl}(U)$. Set $\operatorname{cl}(U)$ is closed, so U is closed.

7.11

☒ 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!

8.1 ☒ 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.

8.2 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. 8.3), it suffices to prove that p and h are continuous. We did so for p in Example 8.4. For h we can use the coordinate criterion (Example 8.6) 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 8.3. 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 8.5 respectively.

8.3 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 8.6.

(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 8.3). The rescaled function cf is the composition $m \circ f$ of two continuous functions, hence continuous (Thm. 8.3).

(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 8.3). The sum function $f + g$ is the composition $s \circ h$ of two continuous functions, hence continuous (Thm. 8.3).

(c), (d) Same as (b), but taking the composition with the function $(x_1,x_2) \mapsto x_1 x_2$ from Example 8.4 and the function $(x_1,x_2) \mapsto x_1/x_2$ from Exercise 8.2.

8.4 Let $\varepsilon > 0$. Choose $\delta = \varepsilon/2$. Then for all $x, y \in (0,1)$ with $|x - y| < \delta$, we have

$$|x^2 - y^2| = |(x+y)(x-y)| = |x+y||x-y| \leq 2|x-y| < 2\delta = \varepsilon.$$

9.1 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 $k \geq N$:

$$d(x_k, 0) = \left|\frac{(-1)^k}{3k-1} - 0\right| = \left|\frac{(-1)^k}{3k-1}\right| = \frac{\left|(-1)^k\right|}{|3k-1|} = \frac{1}{3k-1} \leq \frac{1}{3k-k} = \frac{1}{2k} \leq \frac{1}{2N} < \varepsilon.$$

9.2 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 $k \geq \max\{N_1, \dots, N_n\}$:

$$\|x_k - x\| \leq \sum_{i=1}^{n} |x_{ki} - x_i| \leq \sum_{i=1}^{n} (\varepsilon/n) = \varepsilon,$$

showing that $\lim_{k\to\infty} x_k = x$.

9.3 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 $k \geq N$.

9.4 (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. 9.1). 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. 9.1). 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:

$$\begin{aligned}
\left|f(x_1, x_2) - 0\right| &= \left|\frac{\sin(x_1 x_2)}{\sqrt{x_1^2 + x_2^2}} - 0\right| = \left|\frac{\sin(x_1 x_2)}{\sqrt{x_1^2 + x_2^2}}\right| \\
&\leq \frac{|x_1 x_2|}{\sqrt{x_1^2 + x_2^2}} = \frac{|x_1||x_2|}{\sqrt{x_1^2 + x_2^2}} \\
&\leq \frac{\|x\|\,\|x\|}{\|x\|} = \|x\| < \delta = \varepsilon.
\end{aligned}$$

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 9.5: 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

$$\begin{aligned}\left|f(x_1,x_2) - 0\right| = \left|\frac{x_1^4 + x_2^4}{x_1^2 + x_2^2} - 0\right| &= \frac{|x_1|^4 + |x_2|^4}{x_1^2 + x_2^2} \leq \frac{\|x\|^4 + \|x\|^4}{\|x\|^2} \\ &= \frac{2\|x\|^4}{\|x\|^2} = 2\|x\|^2 < 2\delta^2 = 2(\sqrt{\varepsilon/2})^2 = \varepsilon.\end{aligned}$$

(e) Intuition: writing

$$f(x_1,x_2) = \frac{x_1^2 + 3x_1^2 x_2 + x_2^2}{x_1^2 + x_2^2} = \frac{x_1^2 + x_2^2}{x_1^2 + x_2^2} + \frac{3x_1^2 x_2}{x_1^2 + x_2^2} = 1 + \frac{3x_1^2 x_2}{x_1^2 + x_2^2},$$

and realizing that the third-degree polynomial $3x_1^2 x_2$ 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 9.5: 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

$$\left|f(x_1,x_2) - 0\right| = \left|\frac{x_1^2 + 3x_1^2 x_2 + x_2^2}{x_1^2 + x_2^2} - 1\right| = \left|\frac{3x_1^2 x_2}{x_1^2 + x_2^2}\right| \leq \frac{3|x_1|^2\,|x_2|}{\|x\|^2} \leq \frac{3\|x\|^2\,\|x\|}{\|x\|^2} = 3\|x\| < 3\delta = \varepsilon.$$

(f) We prove using Definition 9.5 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

$$\left|f(x_1,x_2) - 0\right| = \left|\frac{4x_1 x_2}{\sqrt{x_1^2 + x_2^2}} - 0\right| = \frac{4|x_1|\,|x_2|}{\sqrt{x_1^2 + x_2^2}} \leq \frac{4\|x\|\,\|x\|}{\|x\|} = 4\|x\| < 4\delta = \varepsilon.$$

(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 (an in Theorem 9.1). Since we have two different candidates, the limit can't exist.

11.1 Let $f, g \in B(U, \mathbb{R})$. Then $f \leq g + d_\infty(f,g)$, so

$$T(f) \overset{(a)}{\leq} T(g + d_\infty(f,g)) \overset{(b)}{\leq} T(g) + \beta d_\infty(f,g).$$

Reversing the roles of f and g gives $T(g) \leq T(f) + \beta d_\infty(f,g)$. Combining the two inequalities gives $d_\infty(T(f), T(g)) \leq \beta d_\infty(f,g)$.

11.2 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 $d(x,y) \overset{\text{fixed}}{=} d(T(x), T(y)) \overset{(39)}{<} d(x,y)$, 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 (39) 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 $k \in \mathbb{N}$: $0 \le d(T^{k+1}(v_0), v) \overset{\text{fixed}}{=} d(T^{k+1}(v_0), T(v)) \overset{(39)}{\le} d(T^k(v_0), v)$, 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 (39), 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.

11.3 (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.

12.1 (a) The sequence converges uniformly and consequently pointwise to the zero function, i.e., to the function f with $f(x) = 0$ for all $x \in \mathbb{R}$. Let $\varepsilon > 0$. Choose $N \in \mathbb{N}, N > \frac{1}{\varepsilon^2}$. Then for all $k \ge N$ and all $x \in \mathbb{R}$:

$$\left|f_k(x) - f(x)\right| = \left|\frac{\sin kx}{\sqrt{k}} - 0\right| \le \frac{1}{\sqrt{k}} \le \frac{1}{\sqrt{N}} < \varepsilon.$$

(b) If $x = 0$, then $\lim_{k\to\infty} f_k(x) = \lim_{k\to\infty} e^0 = 1$. If $x \in (0,\infty)$, then $\lim_{k\to\infty} f_k(x) = \lim_{k\to\infty} e^{-kx^2} = 0$. So the sequence converges pointwise to $f : D \to \mathbb{R}$ with $f(0) = 1$ and $f(x) = 0$ for all $x > 0$.

The sequence does not converge uniformly; otherwise, its restriction to, for instance, the interval $[0,1]$ would be continuous by Theorem 12.1.

We can also prove the latter from the definition of uniform convergence. For instance, let $0 < \varepsilon < e^{-1}$. For each $k \in \mathbb{N}$, if $x = 1/\sqrt{k}$, then

$$\left|f_k(x) - f(x)\right| = \left|e^{-1} - 0\right| = e^{-1} > \varepsilon.$$

(c) For each $x \in \mathbb{R}_+$,

$$\lim_{k\to\infty} f_k(x) = \lim_{k\to\infty} \frac{k}{1+k\sqrt{k}} x = \lim_{k\to\infty} \frac{1}{\frac{1}{k} + \sqrt{k}} x = 0x = 0,$$

so the sequence converges pointwise to the zero function.

The sequence does not converge uniformly. For instance, let $0 < \varepsilon < \frac{1}{2}$. For each $k \in \mathbb{N}$, if $x = \sqrt{k}$, then

$$\left|f_k(x) - 0\right| = \left|\frac{k\sqrt{k}}{1+k\sqrt{k}} - 0\right| \ge \frac{k\sqrt{k}}{k\sqrt{k} + k\sqrt{k}} = \frac{1}{2} > \varepsilon.$$

(d) The sequence converges uniformly and consequently pointwise to the linear function $f : [0,1] \to \mathbb{R}$ with $f(x) = x$.

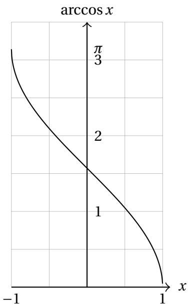
Let $\varepsilon > 0$. Choose $N \in \mathbb{N}$ such that

$$1 - \cos\frac{1}{1+N} < \varepsilon,$$

i.e., such that

$$1 - \varepsilon < \cos\frac{1}{1+N}.$$

To see that this can be done, note that the inequality is automatically true if $\varepsilon > 1$. If $0 < \varepsilon \leq 1$, recall that the cosine function $x \mapsto \cos x$ is invertible on domain $[0,\pi]$ and that its inverse is the arccosine function $x \mapsto \arccos x$ from $[-1,1]$ to $[0,\pi]$:



Therefore,

$$\begin{aligned} 1 - \varepsilon < \cos\frac{1}{1+N} &\Leftrightarrow \frac{1}{1+N} < \arccos(1-\varepsilon) \\ &\Leftrightarrow \frac{1}{\arccos(1-\varepsilon)} < 1 + N \\ &\Leftrightarrow N > \frac{1}{\arccos(1-\varepsilon)} - 1. \end{aligned}$$

So we need to choose $N > \frac{1}{\arccos(1-\varepsilon)} - 1$.

Then for all $k \geq N$ and all $x \in [0,1]$:

$$\begin{aligned} \left|f_k(x) - f(x)\right| = \left|x\cos\frac{x}{x+k} - x\right| = |x|\left|\cos\frac{x}{x+k} - 1\right| \leq \left|\cos\frac{x}{x+k} - 1\right| \\ = 1 - \cos\frac{x}{x+k} \leq 1 - \cos\frac{1}{1+k} \leq 1 - \cos\frac{1}{1+N} < \varepsilon. \end{aligned}$$

(e) Idea: In $x = 0$, the sequence of function values $f_k(0) = \frac{k \cdot 0^2 + 1}{k \cdot 0 + 1} = 1$ converges trivially to $f(0) = 1$. If $x \in (0,1]$, then

$$f_k(x) = \frac{kx^2+1}{kx+1} = \frac{x^2 + 1/k}{x + 1/k} \to x \qquad \text{as } k \to \infty.$$

So our intuition is that the sequence converges pointwise to the function

$$f : [0,1] \to \mathbb{R} \text{ with } f(0) = 1 \text{ and } f(x) = x \text{ otherwise.} \tag{182}$$

Since this function is not continuous, the convergence cannot be uniform by Theorem 10.1. Let's make that precise:

Pointwise convergence to (182):

☒ If $x = 0$, the sequence of function values $f_k(0) = \frac{k \cdot 0^2 + 1}{k \cdot 0 + 1} = 1$ converges trivially to $f(0) = 1$.

☒ If $x \in (0,1]$, let $\varepsilon > 0$ and choose $N > 2/(\varepsilon x)$. For each $k \geq N$:

$$|f_k(x) - f(x)| = \left|\frac{kx^2+1}{kx+1} - x\right| = \left|\frac{kx^2+1-x(kx+1)}{kx+1}\right| = \left|\frac{1-x}{kx+1}\right|$$
$$\leq \frac{1+|x|}{kx+1} < \frac{1+1}{kx+0} = \frac{2}{kx} < \varepsilon.$$

Uniform convergence: Having established pointwise convergence to the function f in (182), which is not continuous in $x = 0$, it follows from Theorem 10.1 that the convergence cannot be uniform.

(f) On domain $D = [1,2]$, the sequence converges uniformly and consequently pointwise to $f : D \to \mathbb{R}$ with $f(x) = x$.
Let $\varepsilon > 0$. Choose $N > 3/\varepsilon$. Then for all $k \geq N$ and $x \in [1,2]$:

$$|f_k(x) - f(x)| \overset{\text{as above}}{=\cdots} \leq \frac{1+|x|}{kx+1} < \frac{1+2}{k\cdot 1+0} = \frac{3}{k} < \varepsilon.$$

12.2 **(a) ⇒ (b):** Assume that $(f_k)_{k\in\mathbb{N}}$ converges uniformly to $f : D \to \mathbb{R}$. Let $\varepsilon > 0$. By uniform convergence, there is an $N \in \mathbb{N}$ such that for all $k \geq N$ and all $x \in D$: $|f_k(x) - f(x)| < \frac{\varepsilon}{2}$. Hence, for all $k, \ell \geq N$ and all $x \in D$:

$$|f_k(x) - f_\ell(x)| = |f_k(x) - f(x) + f(x) - f_\ell(x)| \leq |f_k(x) - f(x)| + |f(x) - f_\ell(x)| < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

(b) ⇒ (a): Assume that for each $\varepsilon > 0$ there is an $N \in \mathbb{N}$ such that for all $k, \ell \geq N$ and all $x \in D$: $|f_k(x) - f_\ell(x)| < \varepsilon$. In particular, for each $x \in D$, the sequence $(f_k(x))_{k\in\mathbb{N}}$ is a Cauchy sequence in $\mathbb{R}$ and consequently has a limit $f(x)$. The function $f : D \to \mathbb{R}$ defined this way is the pointwise limit of $(f_k)_{k\in\mathbb{N}}$. Now, let's show that the convergence is uniform.
SHORT ARGUMENT: Let $\varepsilon > 0$. There is an $N \in \mathbb{N}$ such that for all $k \geq N, n \in \mathbb{N}, x \in D$:

$$|f_k(x) - f_{k+n}(x)| < \frac{\varepsilon}{2}.$$

Taking the limit over n gives

$$|f_k(x) - f(x)| = \lim_{n\to\infty} |f_k(x) - f_{k+n}(x)| \leq \frac{\varepsilon}{2} < \varepsilon.$$

So the convergence is uniform.
DETAILED ARGUMENT: Let $\varepsilon > 0$. By the Cauchy criterion, there is an $N \in \mathbb{N}$ such that

$$\text{for all } k, \ell \geq N, x \in D: \qquad |f_k(x) - f_\ell(x)| < \varepsilon/2.$$

Now fix $k \geq N$ and $x \in D$. Since $\lim_{n\to\infty} f_n(x) = f(x)$, there is an $M \geq N$ such that

$$\text{for all } \ell \geq M: \qquad |f_\ell(x) - f(x)| < \varepsilon/2.$$

In particular, $|f_M(x) - f(x)| < \varepsilon/2$. Therefore,

$$\text{for all } k \geq N: \qquad |f_k(x) - f(x)| = |f_k(x) - f_M(x) + f_M(x) - f(x)| \leq |f_k(x) - f_M(x)| + |f_M(x) - f(x)| < \varepsilon/2 + \varepsilon/2 = \varepsilon.$$

Since $\varepsilon > 0, k \geq N, x \in D$ were arbitrary, it follows that for all $\varepsilon > 0$ there is an $N \in \mathbb{N}$ such that

$$\text{for all } k \geq N, x \in D: \qquad |f_k(x) - f(x)| < \varepsilon.$$

So $f_n \to f$ uniformly.
REMARK: This was a bit tricky. In the proof, we did an intermediate step using an M that was allowed to depend on both ε and x. But we obtained a conclusing that was independent of x. What the proof did was illustrate a process that could be used repeatedly for each $x \in D$.

13.1 According to Definition 13.1, 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

$$0 \leq x_1 = x_1 + 0 \leq x_1 + x_2 \leq 3,$$

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:

$$\{x \in \mathbb{R}^2 : 3x_1 - 2x_2 \leq 6\}, \quad \{x \in \mathbb{R}^2 : x_1 + x_2 \leq 3\}, \quad \{x \in \mathbb{R}^2 : x_1 \geq 0\}, \quad \{x \in \mathbb{R}^2 : x_2 \geq 0\}.$$

As the pre-image of a closed set under a continuous function, each of these four sets is closed (recall Theorem 8.2). 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 $x_1 \leq 3$.

13.2 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.

13.3 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:

$$d(c, c_1) \leq d(c, c_i) + d(c_i, c_1) < 1 + d(c_i, c_1) \leq \delta.$$

13.4 Each totally bounded set is bounded (Exc. 13.3). 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 9.2.

13.5 The closure of U is closed (Definition 7.3). If we can show that it is bounded as well, it is compact by Definition 13.1 (or Heine-Borel, Theorem 13.5). We establish boundedness in two ways:

METHOD 1: By definition (Def. 6.3), 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 ≤, 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 7.2) containing U and cl(U) is the *smallest* closed set containing U:

$$\text{cl}(U) \subseteq \{y \in \mathbb{R}^n : d(x,y) \leq r\}.$$

This shows that cl(U) is bounded: it is contained in any open ball around x with radius larger than r.

METHOD 2: By definition (Def. 6.3), 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 7.4(b), which says that

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

to show that

$$\text{cl}(U) \subseteq B(x, r+1),$$

which makes cl(U) bounded. So let $y \in \text{cl}(U)$. 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!

13.6 (a) For each $x \in [0,1]$, the triangle inequality gives:

$$\begin{aligned}\left|f_b(x) - f_a(x)\right| &= |(b_1 x + b_2) - (a_1 x + a_2)| = |(b_1 - a_1)x + (b_2 - a_2)| \\ &\leq |(b_1 - a_1)x| + |b_2 - a_2| = |b_1 - a_1|\,|x| + |b_2 - a_2| \\ &\leq |b_1 - a_1| \cdot 1 + |b_2 - a_2| = |b_1 - a_1| + |b_2 - a_2|.\end{aligned}$$

(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 13.6). By Tychonoff's theorem, their Cartesian product $[-1,4]\times[2,3]$ is compact in $\mathbb{R}^2$. By Theorem 13.6(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.

13.7 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!

14.1 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!

14.2 We solve the system of linear inequalities:

$$\begin{aligned} x_1 - x_2 \phantom{{}+2x_3} &\leq 0 && (183)\\ x_1 \phantom{{}+x_2} - x_3 &\leq 0 && (184)\\ -x_1 + x_2 + 2x_3 &\leq 2 && (185)\\ - x_3 &\leq -1 && (186) \end{aligned}$$

We first eliminate x_1. In inequality (185), x_1 has a negative coefficient. It imposes a lower bound on x_1:

$$x_2 + 2x_3 - 2 \leq x_1.$$

Inequalities (183) and (184), where x_1 has a positive coefficient, impose upper bounds on x_1:

$$\begin{aligned} x_1 &\le x_2 \\ x_1 &\le x_3 \end{aligned}$$

or, equivalently: $x_1 \le \min\{x_2, x_3\}$. Inequality (186), where x_1 has coefficient zero (i.e., it does not occur in (186)) imposes no bounds on x_1.

So x_1 must satisfy

$$x_2 + 2x_3 - 2 \le x_1 \le \min\{x_2, x_3\}$$

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 (186). So there is a solution if and only if

$$\begin{aligned} x_2 + 2x_3 - 2 &\le x_2 \\ x_2 + 2x_3 - 2 &\le x_3 \\ -\ x_3 \quad &\le -1 \end{aligned}$$

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:

$$\begin{aligned} x_3 &\le 1 \\ x_2 + x_3 &\le 2 \\ -x_3 &\le -1 \end{aligned}$$

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:

$$x_2 \le 2 - x_3.$$

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 $x_3 = 1$, $x_2 \le 2 - x_3 = 1$ and

$$x_2 + 2x_3 - 2 \le x_1 \le \min\{x_2, x_3\} \iff x_2 + 2\cdot 1 - 2 = x_2 \le x_1 \le \min\{x_2, 1\}.$$

Since we know that $x_2 \le 1$, it follows that $\min\{x_2, 1\} = x_2$, so

$$x_2 \le x_1 \le x_2 \qquad \text{or simply} \qquad x_1 = x_2.$$

14.3 If $P = \{x \in \mathbb{R}^n : Ax \le \mathbf{0}\}$ for some matrix A, then P is obviously a polyhedron. It is also a convex cone: if $x, y \in P$ and $\lambda \ge 0$, then

☒ $Ax \le \mathbf{0}$ and $Ay \le \mathbf{0}$ imply $A(x + y) = Ax + Ay \le \mathbf{0} + \mathbf{0} = \mathbf{0}$, so $x + y \in P$.

☒ $Ax \le \mathbf{0}$ implies $A(\lambda x) = \lambda(Ax) \le \lambda\mathbf{0} = \mathbf{0}$, 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 $P = \{x \in \mathbb{R}^n : Ax \le b\}$. Let us prove that $P = \{x \in \mathbb{R}^n : Ax \le \mathbf{0}\}$.

☒ Since P is nonempty and a convex cone, it follows that $\mathbf{0} \in P$. Consequently, $A\mathbf{0} = \mathbf{0} \le b$. So if $Ax \le \mathbf{0}$, then $Ax \le b$, showing that $P \supseteq \{x \in \mathbb{R}^n : Ax \le \mathbf{0}\}$.

☒ To establish the inclusion $P \subseteq \{x \in \mathbb{R}^n : Ax \le \mathbf{0}\}$, we need to show that $Ax \le \mathbf{0}$ 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 $\lambda \ge 0$, it follows that $a_i(\lambda x) = \lambda(a_i x)$ is not bounded from above, contradicting that $a_i x \le b_i$ for all $x \in P$.

14.4 A polytope is the convex hull of a finite set of vectors $\{v_1, \dots, v_m\}$. Let $V \in \mathbb{R}^{n\times m}$ have these vectors as its columns. Then the polytope can be written as

$$\{x \in \mathbb{R}^n : \text{there is a } \lambda \in \mathbb{R}^m \text{ with } x = V\lambda, \lambda \ge \mathbf{0}, \sum_i \lambda_i = 1\}$$

which is the projection of the polyhedron

$$\{(x,\lambda) \in \mathbb{R}^{n+m} : x - V\lambda \le 0, -x + V\lambda \le 0, -\lambda \le 0, \sum_i \lambda_i \le 1, -\sum_i \lambda_1 \le -1\}$$

by projecting away the coordinates of λ one at a time. By Fourier-Motzkin elimination, this is a polyhedron.

15.1 (a) Rewrite the inequalities in terms of upper and lower bounds on x_1:

$$\begin{aligned} x_1 &\le 4 - x_2 - x_3 \\ x_1 &\le 2x_2 + x_3 \\ -1 + x_2 + x_3 &\le x_1 \\ 7 - 3x_2 - 4x_3 &\le x_1 \end{aligned}$$

This means that x_1 must satisfy

$$\max\{-1 + x_2 + x_3, 7 - 3x_2 - 4x_3\} \le x_1 \le \min\{4 - x_2 - x_3, 2x_2 + x_3\}. \tag{187}$$

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:

$$\begin{aligned} -1 + x_2 + x_3 &\le 4 - x_2 - x_3 \\ -1 + x_2 + x_3 &\le 2x_2 + x_3 \\ 7 - 3x_2 - 4x_3 &\le 4 - x_2 - x_3 \\ 7 - 3x_2 - 4x_3 &\le 2x_2 + x_3 \end{aligned}$$

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

$$\max\{-1, \frac{3}{2} - \frac{3}{2}x_3, \frac{7}{5} - x_3\} \le x_2 \le \frac{5}{2} - x_3. \tag{188}$$

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 (188) and the feasible x_1 follow from (187).

(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 $y^\top b < 0$.

15.2 (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 15.3, 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 15.3:

$$\begin{aligned} \{x \in \mathbb{R}^n : Ax = b\} \neq \emptyset &\Leftrightarrow \left\{x \in \mathbb{R}^n : \begin{bmatrix} A \\ -A \end{bmatrix} x \leq \begin{bmatrix} b \\ -b \end{bmatrix}\right\} \neq \emptyset \\ &\Leftrightarrow \left\{(y_1, y_2) \in \mathbb{R}^{m+m} : (y_1, y_2)^\top \begin{bmatrix} A \\ -A \end{bmatrix} = \mathbf{0}^\top, (y_1, y_2)^\top \begin{bmatrix} b \\ -b \end{bmatrix} < 0, (y_1, y_2) \geq \mathbf{0}\right\} = \emptyset \\ &\Leftrightarrow \left\{(y_1, y_2) \in \mathbb{R}^{m+m} : (y_1 - y_2)^\top A = \mathbf{0}^\top, (y_1 - y_2)^\top b < 0, (y_1, y_2) \geq \mathbf{0}\right\} = \emptyset \\ &\Leftrightarrow \left\{y \in \mathbb{R}^m : y^\top A = \mathbf{0}^\top, y^\top b < 0\right\} = \emptyset. \end{aligned}$$

(c) Using Theorem 15.3:

$$\begin{aligned} \{x \in \mathbb{R}^n : Ax \leq b, x \geq \mathbf{0}\} \neq \emptyset &\Leftrightarrow \left\{x \in \mathbb{R}^n : \begin{bmatrix} A \\ -I \end{bmatrix} x \leq \begin{bmatrix} b \\ \mathbf{0} \end{bmatrix}\right\} \neq \emptyset \\ &\Leftrightarrow \left\{(y_1, y_2) \in \mathbb{R}^{m+m} : (y_1, y_2)^\top \begin{bmatrix} A \\ -I \end{bmatrix} = \mathbf{0}^\top, (y_1, y_2)^\top \begin{bmatrix} b \\ \mathbf{0} \end{bmatrix} < 0, (y_1, y_2) \geq \mathbf{0}\right\} = \emptyset \\ &\Leftrightarrow \left\{(y_1, y_2) \in \mathbb{R}^{m+m} : y_1^\top A - y_2^\top = \mathbf{0}^\top, y_1^\top b < 0, (y_1, y_2) \geq \mathbf{0}\right\} = \emptyset \\ &\Leftrightarrow \left\{y \in \mathbb{R}^m : y^\top A \geq \mathbf{0}^\top, y^\top b < 0, y \geq \mathbf{0}\right\} = \emptyset. \end{aligned}$$

(d) Let $\mathbf{1} = (1, \dots, 1) \in \mathbb{R}^n$ be the vector with all coordinates equal to 1. Using Theorem 15.1:

$$\begin{aligned}
\{x \in \mathbb{R}^n : Ax = \mathbf{0}, \mathbf{0} \neq x \geq \mathbf{0}\} \neq \emptyset &\Leftrightarrow \{x \in \mathbb{R}^n : Ax = \mathbf{0}, \sum_i x_i = 1, x \geq \mathbf{0}\} \neq \emptyset \\
&\Leftrightarrow \left\{x \in \mathbb{R}^n : \begin{bmatrix} A \\ \mathbf{1}^\top \end{bmatrix} x = \begin{bmatrix} \mathbf{0} \\ 1 \end{bmatrix}, x \geq \mathbf{0}\right\} \neq \emptyset \\
&\Leftrightarrow \left\{(y, y_{m+1}) \in \mathbb{R}^{m+1} : (y, y_{m+1})^\top \begin{bmatrix} A \\ \mathbf{1}^\top \end{bmatrix} \geq \mathbf{0}^\top, (y, y_{m+1})^\top \begin{bmatrix} \mathbf{0} \\ 1 \end{bmatrix} < 0\right\} = \emptyset \\
&\Leftrightarrow \left\{(y, y_{m+1}) \in \mathbb{R}^{m+1} : y^\top A + y_{m+1}\mathbf{1}^\top \geq \mathbf{0}^\top, y_{m+1} < 0\right\} = \emptyset \\
&\Leftrightarrow \left\{y \in \mathbb{R}^m : y^\top A > \mathbf{0}^\top\right\} = \emptyset.
\end{aligned}$$

(e) Rescaling vector x if necessary and using Theorem 15.3:

$$\begin{aligned}
\{x \in \mathbb{R}^n : Ax = \mathbf{0}, x > \mathbf{0}\} \neq \emptyset &\Leftrightarrow \{x \in \mathbb{R}^n : Ax = \mathbf{0}, x \geq \mathbf{1}\} \neq \emptyset \\
&\Leftrightarrow \left\{x \in \mathbb{R}^n : \begin{bmatrix} A \\ -A \\ -I \end{bmatrix} x \leq \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ -\mathbf{1} \end{bmatrix}\right\} \neq \emptyset \\
&\Leftrightarrow \left\{(y_1, y_2, y_3) \in \mathbb{R}^{3m} : (y_1, y_2, y_3)^\top \begin{bmatrix} A \\ -A \\ -I \end{bmatrix} = \mathbf{0}^\top, (y_1, y_2, y_3)^\top \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ -\mathbf{1} \end{bmatrix} < 0, (y_1, y_2, y_3) \geq \mathbf{0}\right\} = \emptyset \\
&\Leftrightarrow \left\{(y_1, y_2, y_3) \in \mathbb{R}^{3m} : (y_1 - y_2)^\top A - y_3^\top = \mathbf{0}^\top, -y_3^\top \mathbf{1} < 0, (y_1, y_2, y_3) \geq \mathbf{0}\right\} = \emptyset \\
&\Leftrightarrow \left\{(y_1, y_2, y_3) \in \mathbb{R}^{3m} : (y_1 - y_2)^\top A = y_3^\top, y_1, y_2 \geq \mathbf{0}, \mathbf{0} \neq y_3 \geq \mathbf{0}\right\} = \emptyset \\
&\Leftrightarrow \{y \in \mathbb{R}^m : \mathbf{0} \neq y^\top A \geq \mathbf{0}^\top\} = \emptyset.
\end{aligned}$$

(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 15.3:

$$\begin{aligned}
\{x \in \mathbb{R}^n : Ax \leq \mathbf{0}, x > \mathbf{0}\} \neq \emptyset &\Leftrightarrow \{x \in \mathbb{R}^n : Ax \leq \mathbf{0}, x \geq \mathbf{1}\} \neq \emptyset \\
&\Leftrightarrow \left\{x \in \mathbb{R}^n : \begin{bmatrix} A \\ -I \end{bmatrix} x \leq \begin{bmatrix} \mathbf{0} \\ -\mathbf{1} \end{bmatrix}\right\} \neq \emptyset \\
&\Leftrightarrow \left\{(y_1, y_2) \in \mathbb{R}^m \times \mathbb{R}^n : \begin{bmatrix} y_1^\top & y_2^\top \end{bmatrix} \begin{bmatrix} A \\ -I \end{bmatrix} = \mathbf{0}^\top, \begin{bmatrix} y_1^\top & y_2^\top \end{bmatrix} \begin{bmatrix} \mathbf{0} \\ -\mathbf{1} \end{bmatrix} < 0, (y_1, y_2) \geq \mathbf{0}\right\} = \emptyset \\
&\Leftrightarrow \left\{(y_1, y_2) \in \mathbb{R}^m \times \mathbb{R}^n : y_1^\top A = y_2^\top, y_2^\top \mathbf{1} > 0, y_1, y_2 \geq \mathbf{0}\right\} = \emptyset \\
&\Leftrightarrow \left\{y \in \mathbb{R}^m : \mathbf{0}^\top \neq y^\top A \geq \mathbf{0}, y \geq \mathbf{0}\right\} = \emptyset.
\end{aligned}$$

15.3 PROOF USING THEOREM 15.1: 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 15.1, 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 15.3: 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 15.3, 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 $y_1^\top A + y_2^\top B + y_3^\top C = \mathbf{0}^\top, \mathbf{0} \neq y_1 \geq \mathbf{0}, y_2 \geq \mathbf{0}$.

15.4 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 15.3 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 $\underbrace{(z_1+z_2)^\top}_{=\mathbf{0}^\top} Az_4 = 0$, whereas on the other

$$(z_1+z_2)^\top Az_4 = (z_1+z_2)^\top(z_1+z_3) = \underbrace{z_1^\top z_1}_{>0} + \underbrace{z_1^\top z_3}_{\geq 0} + \underbrace{z_2^\top z_1}_{\geq 0} + \underbrace{z_2^\top z_3}_{\geq 0} > 0.$$

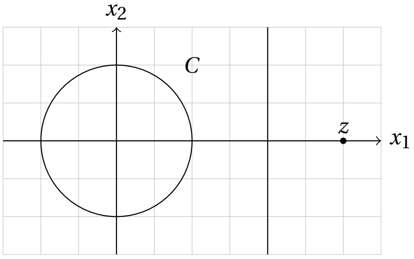
16.1 By the parallelogram law:

$$\begin{aligned}\|x-z\|^2 &= \|\tfrac{1}{2}(x_1+x_2)-z\|^2 = \|\tfrac{1}{2}(x_1-z)+\tfrac{1}{2}(x_2-z)\|^2\\ &= 2\|\tfrac{1}{2}(x_1-z)\|^2 + 2\|\tfrac{1}{2}(x_2-z)\|^2 - \|\tfrac{1}{2}(x_1-z)-\tfrac{1}{2}(x_2-z)\|^2 = \tfrac{1}{2}\|x_1-z\|^2+\tfrac{1}{2}\|x_2-z\|^2-\tfrac{1}{4}\|x_1-x_2\|^2.\end{aligned}$$

If $x_1 \neq x_2$, this is smaller than the squared distance of x_1 or x_2 to z.

16.2 (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

$$\begin{array}{rcrl} c^\top(1,0) &=& c_1 & \leq 0,\\ c^\top(-1,0) &=& -c_1 & \leq 0,\\ c^\top(0,1) &=& c_2 & \leq 0,\\ c^\top(0,-1) &=& -c_2 & \leq 0,\end{array}$$

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 $y \in C_2$.

$$\left\{\begin{array}{ll} x=(0,0)\in C_1, & y=(1,0)\in C_2\\ x=(0,0)\in C_1, & y=(-1,0)\in C_2\\ x=(0,1)\in C_1, & y=(0,0)\in C_2\\ x=(0,-1)\in C_1, & y=(0,0)\in C_2\end{array}\right. \text{ gives } \left\{\begin{array}{lrcrl} c^\top x = & 0 & \leq & c_1 & = c^\top y,\\ c^\top x = & 0 & \leq & -c_1 & = c^\top y,\\ c^\top x = & c_2 & \leq & 0 & = c^\top y,\\ c^\top x = & -c_2 & \leq & 0 & = c^\top y,\end{array}\right.$$

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 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 15.1), 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

$$\begin{cases} 2y_1 + 2y_2 \geq 0 \\ -y_1 - 2y_2 \geq 0 \\ -y_1 + y_2 < 0 \end{cases}$$

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 cone$\{v_1, v_2\}$: since $y^\top v_1 \geq 0$ and $y^\top v_2 \geq 0$, it follows that

$$y^\top z = 3 \cdot (-1) - 2 \cdot 1 = -5 < 0 \leq y^\top v \qquad \text{for all } v \in \text{cone}\{v_1, v_2\}.$$

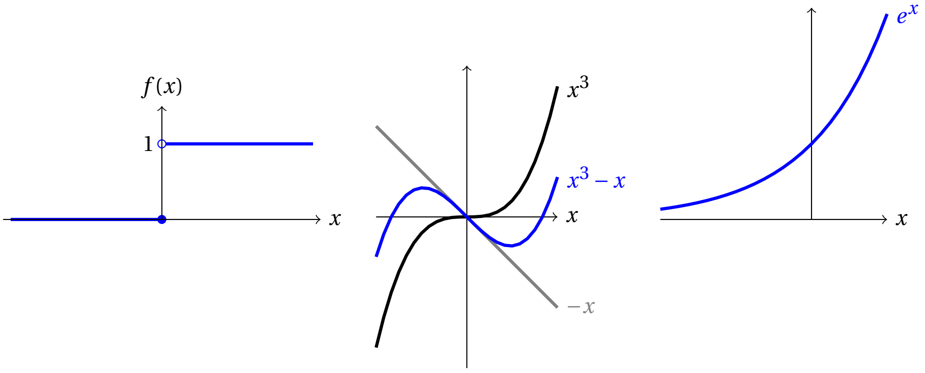
For $-5 < \delta < 0$, the previous line shows that hyperplane $H = \{x \in \mathbb{R}^2 : y^\top x = \delta\} = \{x \in \mathbb{R}^2 : 3x_1 - 2x_2 = \delta\}$ *strictly* separates z from cone$\{v_1, v_2\}$.

17.2 (a) Function f with $f(x) = 0$ if $x \leq 0$ and $f(x) = 1$ otherwise is weakly increasing, hence quasiconcave (Thm. 17.8), but is not continuous at $x = 0$.

(b) The function $x \mapsto -x$ is decreasing, the function $x \mapsto x^3$ is increasing: they are quasiconcave (Thm. 17.8). Their sum, the function $x \mapsto x^3 - x = x(x+1)(x-1)$ first goes up, then down, then up again: by Thm. 17.8, it is not quasiconcave.(Thm. 17.2)

(c) The exponential function $x \mapsto e^x$ is increasing, hence quasiconcave (Thm. 17.8), and has a nonnegative second derivative, hence convex (Thm. 17.4); it is not concave because the area below its graph is not a convex set.

The graphs for the three different subquestions are drawn below.



17.3 (a) No. Function $f : \mathbb{R} \to \mathbb{R}$ with $f(x) = x$ is linear, hence concave, and $g : \mathbb{R} \to \mathbb{R}$ with $g(x) = x^3$ is strictly increasing. Their composition $g \circ f$ has $(g \circ f)(x) = g(f(x)) = x^3$, which is not concave: the line piece between the points $(0, 0)$ and $(1, 1)$ on its graph lies above its graph.

(b) Yes. Let $x, y \in C$ and $\lambda \in [0, 1]$. We need to prove that

$$g\big(f(\lambda x + (1-\lambda)y)\big) \geq \min\big\{g(f(x)), g(f(y))\big\}.$$

Without loss of generality, assume that $f(x) \geq f(y)$. Since g is strictly increasing, also

$$g(f(x)) \geq g(f(y)), \qquad \text{i.e.,} \qquad \min\big\{g(f(x)), g(f(y))\big\} = g(f(y)).$$

Since f is quasiconcave,

$$f(\lambda x + (1-\lambda)y) \geq \min\{f(x), f(y)\} = f(y).$$

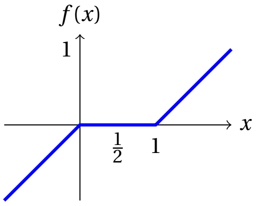
Since g is strictly increasing,

$$g\big(f(\lambda x + (1-\lambda)y)\big) \geq g(f(y)) = \min\big\{g(f(x)), g(f(y))\big\}.$$

17.4 Below is the graph of the function $f : \mathbb{R} \to \mathbb{R}$ with

$$f(x) = \begin{cases} x & \text{if } x \leq 0, \\ 0 & \text{if } 0 < x < 1, \\ x-1 & \text{if } x \geq 1. \end{cases}$$

Point $x = \frac{1}{2}$ is a local maximum, since $f\left(\frac{1}{2}\right) \geq f(x)$ for all x in its neighborhood $(0,1)$, but not a global maximum, since points $x > 1$ achieve a higher function value.



19.1 (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: $X_{LD} = \{(0,0),(2,8)\}$.

19.2 Argue yourself that in all three cases the Extreme Value Theorem (Thm. 13.3) 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 19.4: the constraints are affine functions. So a maximum must satisfy the KKT conditions. Assigning multipliers $\mu_1, \dots, \mu_4$ 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:

$$\begin{array}{ll} \text{maximize} & f(x) = -4x_1 + 3x_2 \\ \text{with} & h_1(x) = x_1 + x_2 - 4 \le 0 \\ & h_2(x) = -x_2 - 7 \le 0 \\ & h_3(x) = (x_1-3)^2 - x_2 - 1 \le 0 \end{array}$$

We can apply case 2 of Theorem 19.4: 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:

$$-6x_1^2 + 4x_1 - 1 + 3\mu_2 x_1^2 - \mu_3(6x_1^2 - 12x_1 + 12) = 0 \tag{189}$$

$$1 + \mu_1 - \mu_2 + \mu_3 = 0 \tag{190}$$

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: (189) gives $\mu_3 = -1/12$, contradicting $\mu_3 \geq 0$.

If $x_1 = 2$, the KKT conditions cannot be satisfied: (189) becomes $\mu_2 - \mu_3 = 17/12$, whereas (190) 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 (190) that $\mu_2 = 1$. Substituting these into (189), 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 19.1, 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 $(1,1)$.

19.3 (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:

$$\mu_1 = -2x_2, \qquad \mu_2 = 2x_1, \qquad \mu_3 = 8x_3 - 2x_1. \tag{191}$$

☒ 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 (191) 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 (191) 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 (191) 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.

19.4 (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 $x_1 = 3/2$.

19.5 (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 16.1.

(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 19.7. 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:

$$-2x_1 + \lambda x_2 = 0 \tag{192}$$

$$-2x_2 + \lambda x_1 = 0 \tag{193}$$

$$-2x_3 - \lambda \quad = 0 \tag{194}$$

Consider two cases:

CASE 1: $x_1 = 0$. Then (193) gives $x_2 = 0$, feasibility gives $x_3 = 5$, (194) 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 (193) gives $x_2 \neq 0$. Rewriting (192) and (193) 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 $\lambda = -2$, (194) gives $x_3 = 1$, (193) 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 $\lambda = 2$, (194) gives $x_3 = -1$, (193) 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 (192) and (193) 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 $x = (2,-2,1)$.

19.6 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 19.8. 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

$$x_2 - 2\lambda = 0,$$
$$x_1 - 2\lambda = 0,$$

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.

19.7 (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:

$$\begin{aligned} 2(1-\lambda)x_1 + \quad\quad 6x_2 &= 0 \\ 6x_1 + 8(1-\lambda)x_2 &= 0 \end{aligned}$$

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.

19.8 (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 $\nabla f(x) - \lambda_1 \nabla g_1(x) - \lambda_2 \nabla g_2(x) = \mathbf{0}$:

$$\begin{aligned} 2x_1 - 2\lambda_1 x_1 - \ \ \lambda_2 &= 0 \\ 4x_2 - 8\lambda_1 x_2 \quad\quad &= 0 \\ 4x_3 \quad\quad\quad - 2\lambda_2 &= 0 \end{aligned}$$

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.

19.9 (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

$$\begin{aligned} 3\mu_1 + \mu_2 - \mu_3 \quad\quad &= 2(1-x_1) \\ -2\mu_1 + \mu_2 \quad\quad - \mu_4 &= 1 \end{aligned}$$

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

$$\begin{aligned} \mu_1 + 2\mu_2 x_1 &= 3x_1^2 \\ 2\mu_2 x_2 &= 1 \\ 2\mu_1 x_3 \qquad &= 0 \end{aligned}$$

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

$$\begin{aligned} 1 - 2\mu_1 x_1 - 2\mu_2 x_1 &= 0 \\ 2 - 2\mu_1 x_2 \qquad\quad &= 0 \\ -2\mu_1 x_3 - 2\mu_2 x_3 &= 0 \end{aligned}$$

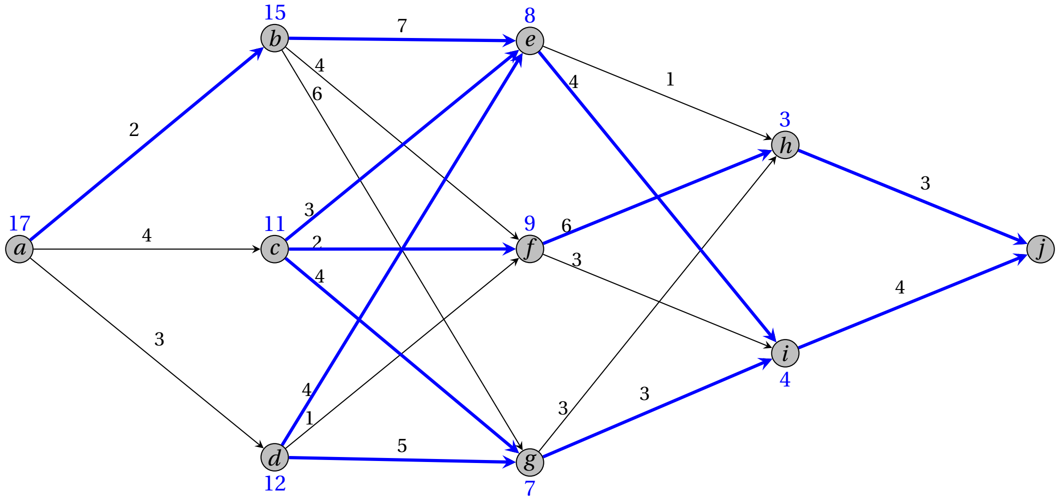
which together with the nonnegativity constraints on μ_1 and μ_2 gives that $\mu_1 > 0, x_2 > 0, x_1 > 0, x_3 = 0$.

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.

20.1 Arguing as in the shortest path problem, but maximizing instead of minimizing, we obtain the following figure. Optimal controls are indicated by the blue lines; the value $J_t(x)$ of the value functions is indicated by a blue number above the nodes. The longest path is $a \to b \to e \to i \to j$ with length 17.



20.2 Legend: the table provides the values of $J_i(x)$. Numbers in bold face indicate that the optimal control at that stage is to pack the item. Boxed numbers indicate that *both* packing and not packing the item are optimal controls at that stage.

(a)

state x	1	2	3	4	5
7	16	**16**	**12**	**9**	**4**
6	16	**16**	**12**	**9**	**4**
5	12	$\boxed{12}$	**12**	**9**	**4**
4	11	**11**	9	**9**	**4**
3	9	9	9	**9**	**4**
2	5	5	5	**5**	**4**
1	4	4	4	4	**4**
0	0	0	0	0	0

stage i

It follows by considering $J_1(7) = 16$ that the optimal value is 16, obtained by packing items 2, 4, and 5. Two things to notice:

1. If remaining capacity is 5 at stage 2:

   ☒ packing item 2 gives value $v(2) + J_3(5 - w(2)) = 7 + J_3(2) = 7 + 5 = 12$;
   ☒ not packing item 2 gives value $J_3(5) = 12$.

   So $J_2(5) = 12$ and both actions (pack/don't pack item 2) are optimal. We indicate both actions being optimal by putting the number in a box.

2. The optimum is obtained by packing items with weight $w(2) + w(4) + w(5) = 3 + 2 + 1 = 6$, less than the total capacity: it is not always optimal to stuff the knapsack to full capacity.

(b)

	8	15	15	**15**	**14**	**9**	**3**
	7	13	$\boxed{13}$	**13**	**11**	**9**	**3**
	6	12	12	**12**	9	**9**	**3**
	5	12	12	**12**	**8**	**6**	**3**
state x	4	12	12	**12**	**8**	3	**3**
	3	9	9	**9**	**8**	3	**3**
	2	7	7	**7**	**5**	3	**3**
	1	4	4	**4**	3	3	**3**
	0	0	0	0	0	0	0
		1	2	3	4	5	6

stage i

It follows by considering $J_1(8) = 15$ that the optimal value is 15, obtained by packing items 3, 4, 5.

(c)

	6	**17**	15	15	**15**	$\boxed{7}$	**7**
	5	**16**	**13**	**12**	**8**	$\boxed{7}$	**7**
	4	12	12	**12**	**8**	7	**7**
state x	3	**11**	7	7	7	7	**7**
	2	7	7	7	7	7	**7**
	1	**4**	0	0	0	0	0
	0	0	0	0	0	0	0
		1	2	3	4	5	6

stage i

If follows from considering $J_1(6) = \mathbf{17}$ that the optimal value is 17, obtained by packing items 1, 2, and 6.

20.3 (a) For $i = 1$ and capacity $x = 0, \ldots, W$, it is optimal to pack item i if and only if capacity allows it. This gives the desired expression for $J_1(x)$. For other $i = 2, \ldots, n$ and $x = 0, \ldots, W$:

☒ If $w(i) > x$, it is not feasible to pack item i, so the optimal value $J_i(x)$ is obtained by packing items $\{1, \ldots, i-1\}$ optimally: $J_i(x) = J_{i-1}(x)$.

☒ If $w(i) \leq x$, there are two options:

1. Don't pack item i. Then the best you can achieve is to pack the remaining items $1, \ldots, i-1$ optimally, resulting in value $J_{i-1}(x)$.
2. Do pack item i. This gives value $v(i)$ and remaining capacity $x - w(i)$ for items $1, \ldots, i-1$. The latter problem has optimal value $J_{i-1}(x - w(i))$, so together this decision to pack item i generates maximal payoff $v(i) + J_{i-1}(x - w(i))$.

Evidently, the optimal thing to do of these two options is the one that generates the highest payoff. So taking the maximum of these two numbers gives $J_i(x) = \max\{J_{i-1}(x), v(i) + J_{i-1}(x - w(i))\}$.

(b) Legend: the table provides the values of $J_i(x)$. Numbers in bold face indicate that the optimal control at that stage is to pack the item. Boxed numbers indicate that *both* packing and not packing the item are optimal controls at that stage. Again, the optimal value is 133, obtained from packing items 1, 2, 4, and 7.

state x	1	2	3	4	5	6	7
10	**60**	**120**	120	**130**	130	130	**133**
9	**60**	**120**	120	**130**	130	130	130
8	**60**	**120**	120	120	120	120	120
7	**60**	$\boxed{60}$	**100**	100	100	100	100
6	**60**	$\boxed{60}$	60	**70**	70	70	**73**
5	**60**	$\boxed{60}$	60	**70**	70	70	**73**
4	**60**	60	60	**70**	70	70	70
3	**60**	60	60	60	60	60	60
2	0	0	0	**10**	10	10	**13**
1	0	0	0	**10**	10	10	10
0	0	0	0	0	0	0	0

stage i

20.4 (a) Since $J_n(x)$ is the smallest capacity needed to obtain aggregate value x from packing a subset of items $\{1,\dots,n\}$ and our knapsack has capacity W, we're interested in finding the largest aggregate value x whose corresponding capacity $J_n(x)$ is still below W: find $\max\{x : J_n(x) \le W\}$.

(b) At stage $i = 1$, we consider subsets of $\{1\}$. There are only two, so the only feasible aggregate values are 0 (by not packing anything) or $v(1)$ by packing item 1. These correspond to weight 0 and $w(1)$, respectively, which gives the desired expression for J_1. For other $i = 2,\dots,n$ and $x = 0,\dots,V$:

☒ If $v(i) > x$, aggregate value x cannot be obtained from packing item i, so the optimal value $J_i(x)$ is obtained using items $1,\dots,i-1$ only: $J_i(x) = J_{i-1}(x)$.

☒ If $v(i) \le x$, there are two options:

1. Don't pack item i. Then aggregate value x must be obtained from items $1,\dots,i-1$ only, which requires minimal weight $J_{i-1}(x)$.
2. Do pack item i. This takes weight $w(i)$ and the remainder of the aggregate value $x - v(i)$ must be obtained with items from $1,\dots,i-1$ which requires minimal weight $J_{i-1}(x - v(i))$. In total, the minimal weight for aggregate value x if you do pack item i is $w(i) + J_{i-1}(x - v(i))$.

The optimal thing to do of these two options is the one that takes minimal weight. So taking the minimum of these two numbers gives $J_i(x) = \min\{J_{i-1}(x), w(i) + J_{i-1}(x - v(i))\}$.

(c) In that example, there are $n = 7$ items with aggregate value $v(1) + \cdots + v(7) = 198$. A table with 7 stages and $i = 0,\dots,198$ states has $7 \cdot 199 = 1393$ entries. You probably won't enjoy computing all of them.

20.5 ☒ A maximum exists: using the system equation, the problem can be rewritten to the maximization of a continuous goal function over controls in the compact set $[0,1]^{T+1}$.

☒ I prove by (backward) induction that J_t is of the form $J_t(x) = \alpha_t x^2$ for some $\alpha_t > 1$.

$$J_T(x) = \sup_{u\in[0,1]} f(T,x,u) = \sup_{u\in[0,1]} x^2(2-u) = 2x^2 \tag{195}$$

with set of optimal controls $[0,1]$ if $x = 0$ and $\{0\}$ if $x \ne 0$. So $J_T(x) = \alpha_T x^2$ with $\alpha_T = 2$. For the induction step, let $t \in \{0,\dots,T-1\}$ and let $J_{t+1}(x) = \alpha_{t+1}x^2$ for some $\alpha_{t+1} > 1$.

$$\begin{aligned} J_t(x) &= \sup_{u\in[0,1]} f(t,x,u) + J_{t+1}(g(t,x,u)) \\ &= \sup_{u\in[0,1]} x^2(2-u) + \alpha_{t+1}(xu)^2 \\ &= \sup_{u\in[0,1]} x^2\left[\alpha_{t+1}u^2 - u + 2\right] \end{aligned}$$

If $x = 0$, all controls $u \in [0,1]$ are optimal. If $x \ne 0$, $\alpha_{t+1} > 1$ makes $u \mapsto \alpha_{t+1}u^2 - u + 2$ (strictly) convex in u: it is maximized in one of its end points. Since $u = 0$ gives 2 and $u = 1$ gives $\alpha_{t+1} + 1 > 2$, the optimal control is $u = 1$ and

$$J_t(x) = \alpha_t x^2 \text{ with } \alpha_t = \alpha_{t+1} + 1. \tag{196}$$

☒ (195) and (196) give $J_t(x) = (T - t + 2)x^2$ for all $t \in \{0, \dots, T\}$. So the maximal value of the goal function is $J_0(x_0) = (T+2)x_0^2$. The set of optimal controls is

$$\begin{cases} [0,1] \text{ at all } t & \text{if } x_0 = 0, \\ \{0\} \text{ at } T \text{ and } \{1\} \text{ at all other times} & \text{if } x_0 \neq 0. \end{cases}$$

20.6 ☒ A maximum exists: using the system equation, the problem can be rewritten to the maximization of a continuous goal function over controls in the compact set $[0,1]^{T+1}$.

☒ I prove by (backward) induction that J_t is of the form $J_t(x) = \alpha_t x + \beta_t$ for some $\alpha_t, \beta_t \in \mathbb{R}$.

$$J_T(x) = \sup_{u \in [0,1]} f(T, x, u) = \sup_{u \in [0,1]} x - u = x$$

with set of optimal controls $\{0\}$. So $J_T(x) = \alpha_T x + \beta_T$ with $\alpha_T = 1, \beta_T = 0$. For the induction step, let $t \in \{0, \dots, T-1\}$ and let $J_{t+1}(x) = \alpha_{t+1} x + \beta_{t+1}$ for some $\alpha_{t+1}, \beta_{t+1} \in \mathbb{R}$.

$$\begin{aligned} J_t(x) &= \sup_{u \in [0,1]} f(t, x, u) + J_{t+1}(g(t, x, u)) \\ &= \sup_{u \in [0,1]} x - u + \alpha_{t+1}(x + u) + \beta_{t+1} \\ &= \sup_{u \in [0,1]} (\alpha_{t+1} + 1)x + (\alpha_{t+1} - 1)u + \beta_{t+1}. \end{aligned}$$

We maximize over $u \in [0,1]$. So if the coefficient $\alpha_{t+1} - 1$ of u is positive, the optimal control is to choose $u = 1$. If it is negative, the optimal control is to choose $u = 0$. If it is zero, the u-term drops out and every control in $[0,1]$ is optimal. That is summarized in the expression

$$J_t(x) = \begin{cases} (\alpha_{t+1} + 1)x + \alpha_{t+1} + \beta_{t+1} - 1, & \text{optimal control } \pi(t,x) = 1 & \text{if } \alpha_{t+1} > 1, \\ (\alpha_{t+1} + 1)x + \beta_{t+1}, & \text{optimal control } \pi(t,x) \in [0,1] & \text{if } \alpha_{t+1} = 1, \\ (\alpha_{t+1} + 1)x + \beta_{t+1}, & \text{optimal control } \pi(t,x) = 0 & \text{if } \alpha_{t+1} < 1. \end{cases}$$

This is of the required affine form $J_t(x) = \alpha_t x + \beta_t$, but the exact shape of α_t and β_t depends on the next period. Let's work backward. We already saw that $J_T(x) = x = \alpha_T x + \beta_T$ with $\alpha_T = 1, \beta_T = 0$ and optimal control $u(T) = 0$. To find J_{T-1}, note that $\alpha_T = 1$, so we are in the middle case of the expression above:

$$J_{T-1}(x) = (\alpha_T + 1)x + \beta_T = 2x \qquad \text{with optimal control } u(T-1) \in [0,1] \text{ arbitrarily.}$$

To find J_{T-2}, note that $\alpha_{T-1} = 2 > 1$, so we are in the top case of the expression above:

$$J_{T-2}(x) = (\alpha_{T-1} + 1)x + \alpha_{T-1} + \beta_{T-1} - 1 = 3x + 2 + 0 - 1 = 3x + 1 \qquad \text{with optimal control } u(T-2) = 1.$$

And so on:

$$\begin{aligned} J_{T-3}(x) &= 4x + 1 + 2 && \text{(with optimal control } u(T-3) = 1) \\ J_{T-4}(x) &= 5x + 1 + 2 + 3 && \text{(with optimal control } u(T-4) = 1) \\ J_{T-5}(x) &= 6x + 1 + 2 + 3 + 4 && \text{(with optimal control } u(T-5) = 1) \end{aligned}$$

and since $1 + 2 + \dots + (k-1) = \frac{1}{2}k(k-1)$, this can be written as

$$J_{T-k}(x) = (k+1)x + \frac{1}{2}k(k-1).$$

☒ So the optimal value of the goal function is $J_0(x_0) = (T+1)x_0 + \frac{1}{2}T(T-1)$ with set of optimal controls $\{0\}$ at T, $[0,1]$ at $T-1$, and $\{1\}$ at all other times.

20.7 ☒ In the final period T, if the state is x, then

$$J_T(x) = \sup_{u \in (0,1]} f(T, x, u) = \sup_{u \in (0,1]} \ln u + x = \ln 1 + x = x$$

with optimal control/policy $\pi(T, x) = 1$.

☒ The next step contains the crucial insight for the general solution. Using that

$$f(T-1,x,u)=\ln u+x, \quad g(T-1,x,u)=x-u, \quad J_T(x)=x$$

all are affine (constant plus linear) functions of x, the DPA applied to time $T-1$ and state x implies that the same holds for J_{T-1}:

$$\begin{aligned} J_{T-1}(x) &= \sup_{u\in(0,1]} f(T-1,x,u)+J_T(g(T-1,x,u)) \\ &= \sup_{u\in(0,1]} \ln u+x+J_T(x-u) \\ &= \sup_{u\in(0,1]} \ln u+x+(x-u) \\ &= \sup_{u\in(0,1]} \ln u-u+2x \end{aligned}$$

The derivative of $\ln u-u+2x$ w.r.t. u is $\frac{1}{u}-1>0$ for all $u\in(0,1)$, making it increasing in u. So the optimal control/policy is $\pi(T-1,x)=1$ and

$$J_{T-1}(x)=\ln 1-1+2x=2x-1,$$

once again an affine function of x.

☒ So let us conjecture that $J_t:\mathbb{R}\to\mathbb{R}$ is of the form $J_t(x)=\alpha_t x+\beta_t$ for some $\alpha_t,\beta_t\in\mathbb{R}$.

☒ Above, we saw that $J_T(x)=x$, so the claim is true at time T with $\alpha_T=1, \beta_T=0$. We will now show that if J_{t+1} is of the desired form, then so is J_t. By induction, we can then conclude that J_t is of the desired form for all $t\in\{0,\dots,T\}$.

☒ So let $t\in\{0,\dots,T-1\}$ and suppose the claim is true for $t+1$: $J_{t+1}(x)=\alpha_{t+1}x+\beta_{t+1}$ for some $\alpha_{t+1},\beta_{t+1}\in\mathbb{R}$. We use the DPA to prove that the claim is true for t:

$$\begin{aligned} J_t(x) &= \sup_{u\in(0,1]} f(t,x,u)+J_{t+1}(g(t,x,u)) \\ &= \sup_{u\in(0,1]} \ln u+x+J_{t+1}(x-u) \\ &= \sup_{u\in(0,1]} \ln u+x+\alpha_{t+1}(x-u)+\beta_{t+1} \\ &= \sup_{u\in(0,1]} \ln u-\alpha_{t+1}u+\beta_{t+1}+(\alpha_{t+1}+1)x \end{aligned}$$

The goal function in the previous line is a concave function of u. The partial derivative w.r.t. u is $\frac{1}{u}-\alpha_{t+1}$. So we need to distinguish two cases:

1. If $\alpha_{t+1}<1$, the partial derivative is positive for all $u\in(0,1]$, so the largest feasible control $\pi(t,x)=1$ is optimal. Substitution in the expression for J_t gives:

$$J_t(x)=\ln 1-\alpha_{t+1}\cdot 1+\beta_{t+1}+(\alpha_{t+1}+1)x=(\alpha_{t+1}+1)x-\alpha_{t+1}+\beta_{t+1}.$$

2. If $\alpha_{t+1}\geq 1$, the optimum is achieved by setting the partial derivative $\frac{1}{u}-\alpha_{t+1}$ equal to zero, finding optimal control $\pi(t,x)=\frac{1}{\alpha_{t+1}}$. Substitution in the expression for J_t gives:

$$J_t(x)=\ln\frac{1}{\alpha_{t+1}}-\alpha_{t+1}\cdot\frac{1}{\alpha_{t+1}}+\beta_{t+1}+(\alpha_{t+1}+1)x=(\alpha_{t+1}+1)x-\ln\alpha_{t+1}-1+\beta_{t+1}.$$

Summarizing these two cases, we see that J_t is of the desired affine form:

$$J_t(x)=\begin{cases} (\alpha_{t+1}+1)x-\alpha_{t+1}+\beta_{t+1} & \text{with optimal control } \pi(t,x)=1 \quad \text{if } \alpha_{t+1}<1, \\ (\alpha_{t+1}+1)x-\ln\alpha_{t+1}-1+\beta_{t+1} & \text{with optimal control } \pi(t,x)=\frac{1}{\alpha_{t+1}} \quad \text{if } \alpha_{t+1}\geq 1. \end{cases}$$

☒ From this expression, we see that $\alpha_t = \alpha_{t+1} + 1$. Together with $\alpha_T = 1$, this gives $\alpha_t = (T + 1 - t)$. Since these numbers are greater than or equal to one, the difference equation for β_t becomes

$$\beta_t = -\ln \alpha_{t+1} - 1 + \beta_{t+1} = -\ln(T - t) - 1 + \beta_{t+1}$$

with $\beta_T = 0$. Consequently,

$$\begin{aligned}
\beta_T &= 0 \\
\beta_{T-1} &= -\ln 1 - 1 + 0 = -1 \\
\beta_{T-2} &= -\ln 2 - 1 + (-1) = -2 - \ln 2 \\
\beta_{T-3} &= -\ln 3 - 1 + (-2 - \ln 2) = -3 - \ln 2 - \ln 3 = -3 - \ln 1 \cdot 2 \cdot 3 \\
&\vdots \\
\beta_t &= -(T - t) - \ln 1 \cdot 2 \cdot \dots \cdot (T - t)
\end{aligned}$$

☒ In particular, the optimal value of the goal function is

$$J_0(x_0) = \alpha_0 x_0 + \beta_0 = (T + 1)x_0 - T - \ln 1 \cdot 2 \cdots T$$

with optimal control 1 at T and $\frac{1}{T-t}$ at all other times $t \in \{0, \dots, T - 1\}$.

ALTERNATIVE METHOD: As a bit of a curiosity, here is another method to solve the problem, using substitution to get rid off the state variables and using standard static optimization. The system equation $x(t + 1) = x(t) - u(t)$ with initial state $x(0) = x_0$ gives

$$\begin{aligned}
x(0) &= x_0 \\
x(1) &= x_0 - u(0) \\
x(2) &= x_0 - u(0) - u(1) \\
&\vdots \\
x(T) &= x_0 - u(0) - u(1) - \dots - u(T - 1)
\end{aligned}$$

so

$$\sum_{t=0}^{T} x(t) = (T + 1)x_0 - Tu(0) - (T - 1)u(1) - \dots - u(T - 1) = (T + 1)x_0 - \sum_{t=0}^{T} (T - t)u(t).$$

Substitution of this expression into the goal function makes the problem equivalent to

$$\text{maximize } \sum_{t=0}^{T} (\ln u(t) - (T - t)u(t)) + (T + 1)x_0 \text{ with } u(0), u(1), \dots, u(T) \in (0, 1].$$

The goal function is concave and its partial derivative with respect to $u(t)$ is

$$\frac{1}{u(t)} - (T - t).$$

At $t = T$, this derivative is positive, making $u^*(T) = 1$ the optimal control. At $t \in \{0, \dots, T - 1\}$, the first-order condition $\frac{1}{u(t)} - (T - t) = 0$ gives optimal control $u^*(t) = \frac{1}{T-t}$. The maximal value of the goal function is

$$\sum_{t=0}^{T-1} \left(\ln \frac{1}{T - t} - 1 \right) + (\ln 1 - (T - T) \cdot 1) + (T + 1)x_0 = (T + 1)x_0 - T - \ln(1 \cdot 2 \cdots T).$$

REMARK: in this exercise, substitution is by far the simplest method. You cannot always do this, however, and the dynamic programming algorithm and the maximum principle — to be discussed in a later section — are more reliable approaches to general problems.

21.1 Using Maximum Principle I (Theorem 21.1), the Hamiltonian is

$$H(t,x,u,p) = f(t,x,u) + pg(t,x,u) = \ln u + x + p(x-u),$$

which is concave in (x,u), so by Theorem 21.2, the maximum principle gives us the desired solution. According to Maximum Principle I, in the optimum $(u^*(0),\dots,u^*(T),x^*(0),\dots,x^*(T))$, there are $p_1,\dots,p_{T+1} \in \mathbb{R}$ with

$$\begin{aligned} p_{T+1} &= 0 \\ p_t &= \frac{\partial}{\partial x} H(t,x^*(t),u^*(t),p_{t+1}) = 1 + p_{t+1} \qquad (t = 1,\dots,T) \end{aligned}$$

so that

$$p_t = (T+1-t) \qquad (t = 1,\dots,T+1)$$

and for all $t = 0,\dots,T$ and all $u(t) \in (0,1]$:

$$\begin{aligned} \frac{\partial}{\partial u} H(t,x^*(t),u^*(t),p_{t+1})(u(t)-u^*(t)) &= \left(\frac{1}{u^*(t)} - p_{t+1}\right)(u(t)-u^*(t)) \\ &= \left(\frac{1}{u^*(t)} - (T-t)\right)(u(t)-u^*(t)) \qquad (197) \\ &\le 0. \end{aligned}$$

Let us solve these inequalities for each time:

☒ At time $t = T$, the inequality reads

$$\frac{1}{u^*(T)}(u(T)-u^*(T)) \le 0$$

for all $u(T) \in (0,1]$. Since $u^*(T)$ is positive, it must be that $u(T) - u^*(T) \le 0$ for all $u(T) \in (0,1]$, i.e., control $u^*(T)$ must be the largest possible: $u^*(T) = 1$.

☒ At time $t = T-1$, the inequality reads

$$\left(\frac{1}{u^*(T-1)} - 1\right)(u(T-1)-u^*(T-1)) \le 0 \qquad (198)$$

for all $u(T-1) \in (0,1]$. This implies that $u^*(T-1) = 1$, making the first term in the product equal to zero. No other control can be optimal: if $u^*(T-1) \in (0,1)$, the first term in the product is positive, so there are feasible controls $u(T-1) > u^*(T-1)$ for which the product in (198) is larger than zero, a contradiction!

☒ At any other time $t \in \{0,\dots,T-2\}$, the inequality reads

$$\left(\frac{1}{u^*(t)} - (T-t)\right)(u(t)-u^*(t)) \le 0$$

for all $u(t) \in (0,1]$. The optimal control $u^*(t)$ cannot be equal to one, because then both the first and the second term in the product become negative at all controls $u(t) \in (0,1)$. So the optimal control $u^*(t)$ must lie in the interior $(0,1)$ of the control region. Remark 21.2 then implies that the partial derivative $\frac{1}{u^*(t)} - (T-t)$ must be zero: the optimal control is $u^*(t) = \frac{1}{T-t}$.

Of course, this gives us the same optimal controls

$$u^*(T) = 1 \text{ and } u^*(t) = \frac{1}{T-t} \text{ for all other } t \in \{0,\dots,T-1\}$$

as in Exercise 20.7 and consequently the same maximal payoff $(T+1)x_0 - T - \ln 1 \cdot 2 \cdots T$.

21.2 Using Maximum Principle I (Theorem 21.1), the Hamiltonian is

$$H(t,x,u,p) = f(t,x,u) + pg(t,x,u) = x - u + p(x+u),$$

which is concave in (x,u), so by Theorem 21.2, the maximum principle gives us the desired solution. According to maximum principle I, in the optimum $(u^*(0),\dots,u^*(T),x^*(0),\dots,x^*(T))$, there are $p_1,\dots,p_{T+1}\in\mathbb{R}$ with

$$\begin{aligned} p_{T+1} &= 0 \\ p_t &= \frac{\partial}{\partial x}H(t,x^*(t),u^*(t),p_{t+1}) = 1+p_{t+1} && (t=1,\dots,T) \end{aligned}$$

so that

$$p_t = (T+1-t) \qquad (t=1,\dots,T+1)$$

and for all $t=0,\dots,T$ and all $u(t)\in[0,1]$:

$$\begin{aligned} \frac{\partial}{\partial u}H(t,x^*(t),u^*(t),p_{t+1})(u(t)-u^*(t)) &= (-1+p_{t+1})(u(t)-u^*(t)) \\ &= (T-t-1)(u(t)-u^*(t)) \\ &\le 0. \end{aligned}$$

Let us solve these inequalities for each time:

⊠ At time $t=T$, the inequality reads

$$-(u(T)-u^*(T)) = u^*(T)-u(T)\le 0$$

for all $u(T)\in[0,1]$. So $u^*(T)=0$.

⊠ At time $t=T-1$, the term $T-t-1$ equals zero, so the inequality is valid for all $u^*(T-1)\in[0,1]$.

⊠ At any other time $t\in\{0,\dots,T-2\}$, the term $T-t-1$ is positive, so the inequality requires $u(t)\le u^*(t)$ for all $u(t)\in[0,1]$, so $u^*(t)=1$.

Of course, this gives us the same solution as in Exercise 20.6.

22.1 In both answers, we use the sandwich property.

(a) Repeated application of (a) gives $|x_{N+k}|\le\alpha^k|x_N|$ for all $k\in\mathbb{N}$. Sequence $(y_t)_{t\in\mathbb{Z}_+}$ with terms

$$|x_0|,|x_1|,\dots,|x_N|,\alpha|x_N|,\alpha^2|x_N|,\alpha^3|x_N|,\dots$$

satisfies $|x_t|\le|y_t|$ for all t and is summable: for each $k\in\mathbb{N}$, the partial sum of the first $N+k$ terms is

$$\sum_{t=0}^{N}|x_t|+\sum_{t=1}^{k}\alpha^t|x_N| = \sum_{t=0}^{N}|x_t|+\frac{\alpha-\alpha^{k+1}}{1-\alpha}|x_N|,$$

with limit $\sum_{t=0}^{N}|x_t|+\frac{\alpha}{1-\alpha}|x_N|$. By Remark 22.1, sequence $x_0,x_1,x_2,\dots$ is summable.

(b) Function $x\mapsto\frac{1}{x^\beta}$ on $(0,\infty)$ is decreasing, so for all $t>1$ and all x between $t-1$ and t: $\frac{1}{t^\beta}\le\frac{1}{x^\beta}$. Integrating over the interval between $t-1$ and t, we see that $\frac{1}{t^\beta}\le\int_{t-1}^{t}\frac{1}{x^\beta}\,\mathrm{d}x$. Sequence $(y_t)_{t\in\mathbb{Z}_+}$ with terms

$$|x_1|,|x_2|,\dots,|x_{N-1}|,\frac{\alpha}{N^\beta},\frac{\alpha}{(N+1)^\beta},\frac{\alpha}{(N+2)^\beta},\dots$$

satisfies $|x_t|\le|y_t|$ for all t and is summable: the partial sum of its first terms is

$$\begin{aligned} \sum_{t=0}^{N-1}|x_t|+\sum_{t=N}^{N+k}\frac{\alpha}{t^\beta} &\le \sum_{t=0}^{N-1}|x_t|+\sum_{t=N}^{N+k}\int_{t-1}^{t}\frac{\alpha}{x^\beta}\,\mathrm{d}x \\ &= \sum_{t=0}^{N-1}|x_t|+\int_{N-1}^{N+k}\frac{\alpha}{x^\beta}\,\mathrm{d}x \\ &= \sum_{t=0}^{N-1}|x_t|+\left[\frac{\alpha}{1-\beta}x^{1-\beta}\right]_{x=N-1}^{x=N+k} \end{aligned}$$

with limit $\sum_{t=0}^{N-1}|x_t|-\frac{\alpha}{1-\beta}(N-1)^{1-\beta}$. By Remark 22.1, sequence $x_0,x_1,x_2,\dots$ is summable.

22.2 This exercise is a slight generalization of Exercise 4.2 in N. L. Stokey and R. E. Lucas (1989) *Recursive methods in economic dynamics*, Harvard University Press.

(a1) The first inequality is evident for $t = 0$. For larger t, it follows from repeated application of (iii). For the second inequality, note that for each coordinate $i \in \{1,\dots,k\}$ of $x(t)$:

$$x(t)_i \le \sqrt{x(t)_i^2} \le \|x(t)\| \le \theta^t \|x_0\|,$$

which is equivalent with the desired vector inequality.

(a2) Let $t \in \mathbb{Z}_+$. Then

$$\begin{aligned} f(x(t), x(t+1)) &\le f(x(t), \mathbf{0}) && \text{(as } x(t+1) \ge \mathbf{0} \text{ and } f \text{ is decr. in its final } k \text{ coord.)} \\ &\le f(\theta^t \|x_0\|\mathbf{1}, \mathbf{0}) && \text{(as } x(t) \le \theta^t\|x_0\|\mathbf{1} \text{ and } f \text{ is incr. in its first } k \text{ coord.)} \\ &\le f(\|x_0\|\mathbf{1}, \mathbf{0}) && (\theta^t\|x_0\|\mathbf{1} \le \|x_0\|\mathbf{1} \text{ as } \theta \in (0,1] \text{ and } f \text{ is incr. in its first } k \text{ coord.)} \end{aligned}$$

(b) Since $\theta > 1$ by (iv), we have that $0 < \frac{1}{\theta^t} \le 1$, so $\|x_0\|\mathbf{1} = \frac{\theta^t}{\theta^t}\|x_0\|\mathbf{1} + \left(1 - \frac{1}{\theta^t}\right)\mathbf{0}$ expresses $\|x_0\|\mathbf{1}$ as a convex combination of the zero vector and $\theta^t\|x_0\|\mathbf{1}$. Consequently,

$$\begin{aligned} f(\|x_0\|\mathbf{1}, \mathbf{0}) &\ge \frac{1}{\theta^t} f(\theta^t\|x_0\|\mathbf{1}, \mathbf{0}) + \left(1 - \frac{1}{\theta^t}\right) f(\mathbf{0}, \mathbf{0}) && \text{(by (vi))} \\ &= \frac{1}{\theta^t} f(\theta^t\|x_0\|\mathbf{1}, \mathbf{0}) && \text{(by (v))} \end{aligned}$$

Multiplying both sides by θ^t proves $f(\theta^t\|x_0\|\mathbf{1}, \mathbf{0}) \le \theta^t f(\|x_0\|\mathbf{1}, \mathbf{0})$. Repeating the first steps of (a2), we find

$$f(x(t), x(t+1)) \le f\left(\theta^t\|x_0\|\mathbf{1}, \mathbf{0}\right) \le \theta^t f\left(\|x_0\|\mathbf{1}, \mathbf{0}\right).$$

Now multiply by β^t to prove (160).

(c) Let $(x(t))_{t \in \mathbb{Z}_+}$ be feasible and let $t \in \mathbb{Z}_+$. Then

$$\begin{aligned} \beta^t f(x(t), x(t+1)) &\le \beta^t f(x(t), \mathbf{0}) && \text{(as } x(t+1) \ge \mathbf{0} \text{ and } f \text{ is decr. in its final } k \text{ coord.)} \\ &\le \beta^t \theta f(x(t-1), \mathbf{0}) && \text{(by (iv))} \\ &\le \cdots \\ &\le \beta^t \theta^t f(x(0), \mathbf{0}) && \text{(by (iv))} \\ &= \left(\beta\theta\right)^t f(x_0, \mathbf{0}) && \text{(by feasibility)} \end{aligned}$$

23.1 (a) Since $f(x,u) = 0$ and $g(x,u) = x/\beta$, the Bellman equation for a function $V : X \to \mathbb{R}$ says that for all $x \in X = \mathbb{R}$:

$$V(x) = \sup_{u \in U(x)} f(x,u) + \beta V(g(x,u)) = \sup_{u \in \mathbb{R}} 0 + \beta V(x/\beta).$$

Since the right side is independent of u, this simply reduces to

$$V(x) = \beta V(x/\beta). \tag{199}$$

(b) Fix some constant $c \in \mathbb{R}$ and define the constant function $V_0 : X \to \mathbb{R}$ with $V_0(x) = c$ for all states x. To compute V_1, replace V on the right side of the Bellman equation with V_0 and solve the corresponding optimization problem. That is, for each $x \in [0,1]$:

$$V_1(x) = \sup_{u \in \mathbb{R}} 0 + \beta \underbrace{V_0(x/\beta)}_{=c}.$$

The right side is independent of u, so this reduces to

$$V_1(x) = \beta c.$$

This says that if you start with a constant function c, you get a new constant function βc. This implies that $V_k(x) = \beta^k c$. And since $\beta \in (0,1)$, $\beta^k c$ tends to 0 as k goes to infinity: the sequence of functions $V_0, V_1, V_2, \dots$ indeed converges to the optimal value function that is constant at zero.

(c) Substituting such a V on both sides of (199), we must verify that for all states $x \in \mathbb{R}$:

$$V(x) = ax + b|x| = \beta V(x/\beta) = \beta\big(a(x/\beta) + b|x/\beta|\big).$$

Since $1/\beta > 0$, this follows easily: the right side can be rewritten as

$$\beta\big(a(x/\beta) + b|x/\beta|\big) = \beta a \frac{x}{\beta} + \beta b \left|\frac{x}{\beta}\right| = ax + \beta b \frac{1}{\beta}|x| = ax + b|x|.$$

23.2 (a) States and controls lie between 0 and 1 and $f(x,u) = \sqrt{xu}$ is (weakly) increasing in both arguments, so f achieves values between $f(0,0) = 0$ and $f(1,1) = 1$.

(b) Consider a function $V : X \to \mathbb{R}$. Since $f(x,u) = \sqrt{xu}$ and $g(x,u) = 1 - u$, the Bellman equation says that for all $x \in [0,1]$:

$$V(x) = \sup_{u \in U(x)} f(x,u) + \beta V(g(x,u)) = \sup_{u \in [0,1]} \sqrt{xu} + \beta V(1-u).$$

(c) To compute V_1, replace V on the right side of the Bellman equation with V_0 and solve the corresponding optimization problem. That is, for each $x \in [0,1]$:

$$V_1(x) = \sup_{u \in [0,1]} \sqrt{xu} + \beta \underbrace{V_0(1-u)}_{=0} = \sup_{u \in [0,1]} \sqrt{xu}.$$

Since $\sqrt{xu}$ is (weakly) increasing in u it follows that the optimal control u is any number in $[0,1]$ if $x = 0$ and the optimal control u is 1 if $x \neq 0$. Substituting this in the right side, we obtain $V_1(x) = \sqrt{x}$.

To compute V_2, replace V on the right side of the Bellman equation with V_1 and solve the corresponding optimization problem. That is, for each $x \in [0,1]$:

$$V_2(x) = \sup_{u \in [0,1]} \sqrt{xu} + \beta V_1(1-u) = \sup_{u \in [0,1]} \sqrt{xu} + \beta\sqrt{1-u}.$$

The optimum on the right might be achieved in one of the endpoints. Control $u = 0$ gives value $\sqrt{xu} + \beta\sqrt{1-u} = \beta$, whereas $u = 1$ gives value $\sqrt{x}$. Or the optimum might be in an interior point, where the first order condition must hold; the derivative of the goal function must be zero:

$$\frac{\sqrt{x}}{2\sqrt{u}} - \frac{\beta}{2\sqrt{1-u}} = 0 \iff \frac{\sqrt{x}}{\sqrt{u}} = \frac{\beta}{\sqrt{1-u}}.$$

Squaring both sides and solving for u gives $u = \frac{x}{x+\beta^2}$. Substituting this into the goal function gives

$$\sqrt{xu} + \beta\sqrt{1-u} = \sqrt{x + \beta^2}.$$

Comparing this with the values $\sqrt{x}$ and β in the endpoints, we see that this is the maximum we're searching for: $V_2(x) = \sqrt{x+\beta^2}$.

(d) Substituting this V on both sides of the Bellman equation, we need to verify that for all $x \in [0,1]$:

$$V(x) = \frac{\sqrt{x + \beta^2(1-x)}}{1-\beta^2} = \sup_{u \in [0,1]} \sqrt{xu} + \beta V(1-u) = \sup_{u \in [0,1]} \sqrt{xu} + \beta\frac{\sqrt{(1-u) + \beta^2 u}}{1-\beta^2}. \tag{200}$$

So we solve the problem on its right side and show that its solution coincides with the expression on the left. The goal function $u \mapsto \sqrt{xu} + \beta\frac{\sqrt{(1-u)+\beta^2 u}}{1-\beta^2}$ is a concave function[5] of $u \in [0,1]$, so points satisfying the first-order condition give a maximum. That first-order condition is

$$\frac{\sqrt{x}}{2\sqrt{u}} - \frac{\beta(1-\beta^2)}{2\sqrt{(1-u)+\beta^2 u}} = 0 \iff \frac{\sqrt{x}}{\sqrt{u}} = \frac{\beta(1-\beta^2)}{\sqrt{(1-u)+\beta^2 u}}.$$

[5] I won't show that here, but it is enough to compute the second derivative and show that it is less than or equal to zero.

Squaring both sides and solving for u shows that the optimal control is

$$u^*(x) = \frac{x}{x+\beta^2(1-x)}, \tag{201}$$

which indeed lies in the set $[0,1]$ of feasible controls. Substituting this into the goal function and simplifying gives

$$\begin{aligned}
\sqrt{xu} + \beta\frac{\sqrt{(1-u)+\beta^2 u}}{1-\beta^2} &= \sqrt{\frac{x^2}{x+\beta^2(1-x)}} + \frac{\beta}{1-\beta^2}\sqrt{\frac{\beta^2(1-x)+\beta^2 x}{x+\beta^2(1-x)}} \\
&= \frac{1}{\sqrt{x+\beta^2(1-x)}}\left(x+\frac{\beta^2}{1-\beta^2}\right) \\
&= \frac{1}{\sqrt{x+\beta^2(1-x)}}\frac{x+\beta^2(1-x)}{1-\beta^2} \\
&= \frac{\sqrt{x+\beta^2(1-x)}}{1-\beta^2},
\end{aligned}$$

which is, as we needed to show, equal to the left side of the Bellman equation (200).

(e) Since the instantaneous payoff function f is bounded, the optimal value function is bounded. The function V is bounded (it is increasing in $x \in [0,1]$, so it achieves values between $V(0)$ and $V(1)$) and satisfies the Bellman equation, making it the optimal value function by Theorem 23.2.

The optimal policy function is the function that assigns to each state x the optimal control that solves the optimization problem in (200); we found it in expression (201): it is the function $u^* : [0,1] \to [0,1]$ with $u^*(x) = \frac{x}{x+\beta^2(1-x)}$.