Piotr Zwiernik

Lecture notes in Mathematics for Economics and Statistics

January 19, 2023

Preface

These lecture notes are intended to give a concise introduction to modern real analysis with a view towards applications in economics, finance, and statistics. The main aim of these notes is to provide students with tools that are essential to grasp basics of optimisation, fixed point theory, and vector calculus. Another important reason behind these lectures is to introduce student into formal mathematical thinking.

The most popular real analysis textbooks are typically designed for first-year undergraduates in mathematics. This affects the exposition. For example, matrix algebra is used scarcely and some topological concepts are developed in full generality. Our lecture notes are intended for different audience. The student is assumed to be familiar with standard algebra and calculus concepts taught during the first years in Economics, Statistics and related degrees. The topology in $\mathbb{R}^n$ is more streamlined and differentiation is build around the idea of close connections between calculus and linear algebra.

The whole material is divided into two. The core part of the lecture will take about 16 lectures. The second part can be used to complement the core material with topics that are relevant for students.

Preparing these lecture notes we benefited from several excellent textbooks:

1. John H. Hubbard, Barbara Burke Hubbard, *Vector Calculus, Linear Algebra, and Differential Forms.*
2. Charles C. Pugh, *Real Mathematical Analysis.*
3. Walter Rudin, *Principles of Mathematical Analysis.*

These notes also benefited from comments of Orestis Vravosinos, Miguel Espinosa, Malachy Gavan, and Christian Brownlees.

Contents

Part I
The core part of the lecture

Chapter 1
Preliminaries (2 lectures)

1.1 Foundations of mathematics

Mathematics is expressed in the language of set theory. A **set** is simply a collection of elements. For example the set of **natural numbers** $\mathbb{N}$ consists of all numbers $1, 2, 3, \ldots$. For a set A and element x we write $x \in A$ if x belongs to A and $x \notin A$ otherwise. For example $0 \notin \mathbb{N}$ but $10 \in \mathbb{N}$. Sets can contain other sets. For example, the set of **integers** $\mathbb{Z}$ that consists of all numbers of the form

$$\ldots, -3, -2, -1, 0, 1, 2, 3, \ldots$$

contains all natural numbers. If A is contained in B, we write $A \subset B$ and say that A is a subset of B. Formally, we write $A \subset B$ if $x \in A$ implies that $x \in B$. We have $\mathbb{N} \subset \mathbb{Z}$.

In mathematics sets have typically some further structure. For example, two numbers can be added, subtracted, multiplied, or divided. The set of natural numbers is closed under addition, meaning that the sum of two natural numbers is a natural number. It is not however closed under subtraction since $2 - 3 \notin \mathbb{N}$. The set of integers is closed under addition and subtraction but it is not closed under division because $\frac{3}{2} \notin \mathbb{Z}$. A minimal system of numbers closed under basic arithmetic operations (addition, subtraction, multiplication, division) is the set of **rational numbers**

$$\mathbb{Q} \;=\; \left\{ \tfrac{m}{n} : \;\; m, n \in \mathbb{Z}, n \neq 0 \right\} \;=\; \left\{ \tfrac{m}{n} : \;\; m \in \mathbb{Z}, n \in \mathbb{N} \right\}.$$

Although $\mathbb{Q}$ is closed under arithmetic operations, as we will see shortly, $\mathbb{Q}$ is not good enough for calculus, which will lead us to the system of **real numbers** $\mathbb{R}$.

Quickly recall basic concepts of the set theory. For two sets A, B their **union** is

$$A \cup B \;=\; \{x : x \in A \text{ or } x \in B\},$$

their **intersection** is

$$A \cap B \;=\; \{x : x \in A \text{ and } x \in B\},$$

their **difference** is

$$A \setminus B \;=\; \{x : x \in A \text{ and } x \notin B\}.$$

Exercise 1.1. Show that $A \cap B = A \setminus (A \setminus B)$.

For any collection of sets E_α, $\alpha \in A$ we define

$$\bigcup_{\alpha \in A} E_\alpha \;=\; \{x : x \in E_\alpha \text{ for some } \alpha \in A\}$$

and

$$\bigcap_{\alpha \in A} E_\alpha \;=\; \{x : x \in E_\alpha \text{ for all } \alpha \in A\}.$$

For example, if $A = [0,1]$ and $E_\alpha = [\alpha, 1+\alpha]$ then $\bigcup_{\alpha\in A} E_\alpha = [0,2]$ and $\bigcap_{\alpha\in A} E_\alpha = \{1\}$. The **empty set** is the set that has no elements and it is denoted by $\emptyset$. A **singleton** is a set that contains only one element.

The complement of a set A is $A^c = \{x : x \notin A\}$ and we have the following useful result.

Theorem 1.1 (De Morgan's Law). *Let $\{E_\alpha\}$ for $\alpha \in A$ is any collection of sets, then*

$$\Big(\bigcup_{\alpha \in A} E_\alpha\Big)^c \;=\; \bigcap_{\alpha \in A} E_\alpha^c.$$

Proof. We will show that x lies in the set on the left if and only if it lies in the set on the right. Indeed, $x \in \big(\bigcup_{\alpha\in A} E_\alpha\big)^c$ if and only if $x \notin \bigcup_{\alpha\in A} E_\alpha$, or in other words, for every $\alpha \in A$, $x \notin E_\alpha$. This is if and only if $x \in E_\alpha^c$ for all $\alpha \in A$. □

De Morgan's Law is the first mathematical statement in these notes that required a proof. We discuss the concept of a proof more closely starting with introducing some mathematical logic.

Statements in mathematics can be either true (1) or false (0). Thus, $2+2 = 4$ is an example of a true statement, whereas $2+2 = 5$ is false. For a statement P by $\neg P$ denote its **negation**, that is, the statement that is true if and only if P is false. For example, the negation of $x + 2 = 4$ is $x + 2 \neq 4$ and the negation of $m > n$ is $m \leq n$. **Conjunction** of two statements P and Q, denoted by $P \wedge Q$ is a statement that is true only if both P and Q are true. **Disjunction**, denoted by $P \vee Q$, is a statement that is false only if both P and Q are false. The **conditional** $P \Rightarrow Q$, is a statement that is false only if P is true and Q is false. The conditional reads "P implies Q" and it says:

if P is true then Q is true but Q can be true for other reasons. This is all summarized in Table 1.1.

Table 1.1 The logical values for $P \wedge Q$, $P \vee Q$ and $P \Rightarrow Q$ depending on the values of P and Q.

P	Q	$P \wedge Q$	$P \vee Q$	$P \Rightarrow Q$
0	0	0	0	1
0	1	0	1	1
1	0	0	1	0
1	1	1	1	1

Exercise 1.2. We write $P \equiv Q$ if both P and Q have the same logical value. Verify, using tables like Table 1.1, that irrespective of the logical value of P and Q:

(i) $\neg(P \vee Q) \quad \equiv \quad \neg P \wedge \neg Q$,
(ii) $\neg(P \wedge Q) \quad \equiv \quad \neg P \vee \neg Q$,
(iii) $P \Rightarrow Q \quad \equiv \quad \neg Q \Rightarrow \neg P \quad \equiv \quad \neg P \vee Q$
(iv) $\neg(P \Rightarrow Q) \quad \equiv \quad P \wedge \neg Q$

Statements in mathematics often involve quantifiers. We write $\forall$ to denote "for all" and $\exists$ to denote "there exists". For example

$$\forall n \in \mathbb{Z} \quad \exists q \in \mathbb{Q} \quad \text{such that } q > n$$

is an example of such a statement. Order of the quantifiers is important. For example the statement

$$\forall n \in \mathbb{N} \ \ \forall m \in \mathbb{N} \ \ \exists p \text{ prime} \quad \text{such that } nm < p$$

is true. Indeed, recall that $p \in \mathbb{N}$ is a prime number if the only natural numbers dividing p and 1 and p itself (e.g. 3, 7, 11, 89). Now the statement essentially follows from the fact that there exist arbitrary large prime numbers. Six different proofs of this fact can be found in Section 1.1 of Martin Aigner, Günter M. Ziegler, "Proofs from THE BOOK"[1]. On the other hand, the statement

$$\forall n \in \mathbb{N} \ \ \exists p \text{ prime} \quad \text{such that } \forall m \in \mathbb{N} \ \ nm < p$$

is false.

Statements involving quantifiers can be negated easily. We simply swap $\forall$ with $\exists$ and negate the assertion. For example, the negation of

$$\forall n \in \mathbb{N} \ \ \exists p \in \mathbb{Q} \text{ such that } n < p$$

[1] https://www.emis.de/classics/Erdos/textpdf/aigzieg/aigzieg.pdf

is the statement

$$\exists n \in \mathbb{N} \text{ such that } \forall p \in \mathbb{Q}\ n \geq p.$$

Make sure you understand why by starting with a single quantifier statement.

Nearly all mathematical assertions can be expressed as the conditional $P \Rightarrow Q$. In order to prove such a conditional statement, it is sometimes easier to prove $\neg Q \Rightarrow \neg P$, which is equivalent by Exercise 1.2. Such statement is called **contrapositive**.

Exercise 1.3. Using the contrapositive statement prove that if n^2 is an even number then n must be even.

To prove a statement P it is often easier to show that $\neg P$ implies a statement which is false (and so $\neg P$ must be false). This proof technique is called **proof by contradiction**. We illustrate it proving two basic results.

Theorem 1.2. *There are no solutions to $x^2 - y^2 = 1$ such that $x, y \in \mathbb{N}$.*

The negation of this statement is that there exists a pair of natural numbers x, y such that $x^2 - y^2 = 1$ and our proof begins by assuming that this is true.

Proof. Suppose that $x, y \in \mathbb{N}$ and $x^2 - y^2 = 1$. Then $(x - y)(x + y) = 1$. Since $x > y$ we get that $x - y, x + y \in \mathbb{N}$. The only possibility now is that $x - y = x + y = 1$. This system of linear equations has only one solution $x = 1$, $y = 0$. This leads to a contradiction because 0 is not a natural number. □

Our second example is the following result.

Theorem 1.3. $\sqrt{2} \notin \mathbb{Q}$.

Proof. Suppose that $\sqrt{2} \in \mathbb{Q}$ or, in other words, $\sqrt{2} = \frac{m}{n}$ with $m \in \mathbb{Z}$ and $n \in \mathbb{N}$ where m and n have no common divisors. Since $2n^2 = m^2$ it must be that m is even, say $m = 2k$ for some $k \in \mathbb{Z}$. But then $2n^2 = 4k^2$ and so $n^2 = 2k^2$, which by Exercise 1.3 implies that n is even. However, if both m and n are even, they have a common factor, which leads to a contradiction. □

As you can see the structure of a proof by contradiction is rather simple:

> **Theorem.** P.
> *Proof.* Suppose not P. Then... Then... Then... Contradiction. □

Now try it by yourself generalising the proof of Theorem 1.3.

Exercise 1.4. Generalize Theorem 1.3 and show that $\sqrt{p} \notin \mathbb{Q}$ for every prime number $p \geq 2$.

Another common proof technique is the proof **by induction**. Suppose that we have a sequence $P(n)$ of statements indexed by natural numbers n and we want to show that $P(n)$ is true for all $n \in \mathbb{N}$. We then first prove it for $n = 1$. Next, we show that the fact that it holds for some n implies that it holds for $n + 1$. For example, to prove that

$$1 + 2 + \cdots + n \;=\; \tfrac{n(n+1)}{2} \tag{1.1}$$

holds for all $n \in \mathbb{N}$ we first make sure (1.1) holds for $n = 1$ (easy). Now suppose (1.1) holds for some n. We want to show that it holds for $n+1$, that is,

$$1 + 2 + \cdots + n + (n + 1) \;=\; \tfrac{(n+1)(n+2)}{2}.$$

But this follows immediately because

$$1 + 2 + \cdots + n + (n + 1) \;=\; \tfrac{n(n+1)}{2} + (n + 1) \;=\; \tfrac{(n+1)(n+2)}{2}.$$

Exercise 1.5. Show that $n! > 2^n$ for all $n \geq 4$.

1.2 Real numbers

Natural numbers $\mathbb{N}$ are used to count objects. In order to allow basic arithmetic operations, we need to extend natural numbers to the rational numbers $\mathbb{Q}$. The set of rational numbers is however not sufficiently rich; for example, we have seen that $\sqrt{2} \notin \mathbb{Q}$. In order to do calculus, we will work over the set of real numbers $\mathbb{R}$. This section provides a basic understanding of why the reals are special.

Definition 1.1. Let $E \subset \mathbb{R}$. If there exists $\beta \in \mathbb{R}$ such that $x \leq \beta$ for all $x \in E$ then E is said to be **bounded above** and β is called an **upper bound** of E.

Definition 1.2. Let $E \subset \mathbb{R}$ be bounded above. Suppose there exists $\alpha \in \mathbb{R}$ such that:

(i) α is an upper bound of E
(ii) if $\gamma < \alpha$ then γ is not an upper bound of E.

Then α is called the **supremum** of E. We write $\alpha = \sup E$.

Exercise 1.6. In the same way we can define bounded below, lower bound, and infimum $\inf E$. Write carefully those definitions.

Example 1.1. If $E = \{0, 1\} \subset \mathbb{R}$ then $\sup E = 1$ and $\inf E = 0$. If $E = \{x \in \mathbb{Q} : x \leq 0\}$ then E has no lower bound and $\sup E = 0$. Indeed, 0 is clearly an upper bound and there is no smaller upper bound because $0 \in E$.

Slightly more sophisticated examples are given in the end of this section.

Exercise 1.7. Show that if the supremum of E exists then it is necessarily unique.

The reason why we added "Suppose there exists" as part of the definition of the supremum is that it is not a priori clear that such a number exists for every E. For an illustration of what can be the issue define the maximum of E as $\max E = \{\alpha \in E : \ \alpha \geq x \text{ for all } x \in E\}$. Then for $E = \{\frac{n-1}{n} : \ n \in \mathbb{N}\}$ the maximum does not exists, which leads to the following corrected definition.

Definition 1.3. Let $E \subset \mathbb{R}$. Suppose that there exists $\alpha \in E$ such that $\alpha \geq x$ for all $x \in E$. Then α is called the maximum of E.

Exercise 1.8. Show that if $E \subset \mathbb{R}$ is a finite set then $\max E$ exists.

Exercise 1.9. Let $E \subset \mathbb{R}$. Show that if $\max E$ exists then $\max E = \sup E$.

One of the most important properties of the real numbers is that every set bounded above admits the supremum, which means that "Suppose there exists" can be removed from Definition 1.2. Although it is often proven from the first principles, to simplify the discussion, we state it as an axiom.

The completeness axiom: Every nonempty subset E of $\mathbb{R}$ that is bounded from above has a supremum in $\mathbb{R}$.

We list some consequences of this axiom that will be frequently used in this course.

Theorem 1.4. $\forall x \in \mathbb{R}\ \exists n \in \mathbb{N}$ *such that* $n > x$.

Proof. We proceed by contradiction. Suppose there exists $x \in \mathbb{R}$ such that $x \geq n$ for all $n \in \mathbb{N}$. Then x is the upper bound of $\mathbb{N}$. Then, there exists $\alpha = \sup \mathbb{N}$. Since $\alpha - 1 < \alpha$, $\alpha - 1$ is not an upper bound of $\mathbb{N}$ and so there exists $n \in \mathbb{N}$ such that $n > \alpha - 1$, but then $n + 1 \in \mathbb{N}$ and $n + 1 > \alpha$, which contradicts the fact that α is an upper bound of $\mathbb{N}$. □

Theorem 1.5. $\forall x > 0\ \exists n \in \mathbb{N}$ *such that* $0 < \frac{1}{n} < x$.

Proof. By the previous theorem there is $n \in \mathbb{N}$ such that $n > \frac{1}{x}$. But this means that $\frac{1}{n} < x$. □

This result shows in particular that the interval $(0, x)$ always contains a rational number. With a little bit of extra work we get the following.

Theorem 1.6. *For any two real numbers $x < y$, there exists a rational number q such that $x < q < y$.*

Proof. If x, y have different signs then the theorem is obviously true with $q = 0$. If $x, y \leq 0$ then we can apply the theorem to $-x, -y$ so, without loss of generality, we can assume $0 \leq x < y$. The case $x = 0$ was covered in the previous theorem so assume $0 < x < y$. By Theorem 1.5, there exists $n_1, n_2 \in \mathbb{N}$ such that $0 < \frac{1}{n_1} < y - x$ and $\frac{1}{n_2} < x$. Let $n = \max\{n_1, n_2\}$. We will show that for some $k \in \mathbb{N}$ we have $x < \frac{k}{n} < y$. The set $E = \{k \in \mathbb{N} : k \leq nx\}$ is non-empty ($1 \in E$) and bounded above by nx. By Theorem 1.4 there exists $m \in \mathbb{N}$ such that $m > nx$ so E must be finite (at most $m - 1$ elements). Let $l = \sup E = \max E$ (c.f. Exercises 1.8 and 1.9) then $l \leq nx$ and $l + 1 > nx$. Together with the fact that $n(y - x) > 1$ this implies that

$$nx \ < \ l + 1 \ \leq \ nx + 1 \ < \ nx + n(y - x) \ = \ ny.$$

We conclude that $x < \frac{l+1}{n} < y$. □

We close this section with a couple of examples.

Example 1.2. Let $E = \{x \in \mathbb{Q} : \ x < 0\}$. This set has no lower bound but 0 is an obvious upper bound. To show that $\sup E$ is actually equal to 0 we need to show that no smaller upper bound is possible. Suppose $\alpha < 0$ and α is an upper bound of E. This is impossible because, by Theorem 1.6, there exists a rational number q such that $\alpha < q < 0$. We conclude that 0 is indeed the supremum.

Example 1.3. If $E = \{\frac{1}{n} : n \in \mathbb{N}\}$. Then $\sup E = 1$ and $\inf E = 0$. Indeed, 1 is an upper bound and so it has to be the supremum because $1 \in E$. It is also clear that 0 is a lower bound. The fact that there is no larger lower bound follows by Theorem 1.5.

Denote by $\overline{\mathbb{R}}$ the extended real line. As a set, $\overline{\mathbb{R}}$ is simply equal to $\mathbb{R} \cup \{-\infty, +\infty\}$, where $-\infty, +\infty$ are formal symbols. For every $x \in \mathbb{R}$ we set $-\infty < x < +\infty$, which extends the ordering in $\mathbb{R}$ to $\overline{\mathbb{R}}$. We also extend the basic arithmetic on $\mathbb{R}$ to $\overline{\mathbb{R}}$ in an obvious way. We require that both the addition and the multiplication are commutative on $\overline{\mathbb{R}}$. Moreover,

$$x + (+\infty) = +\infty, \quad x + (-\infty) = -\infty,$$

$$(+\infty) + (+\infty) = +\infty, \quad (-\infty) + (-\infty) = -\infty,$$

To extend multiplication, apart from commutativity we require that

$$x(\pm\infty) = \begin{cases} \pm\infty & \text{if } x > 0, \\ 0 & \text{if } x = 0, \\ \mp\infty & \text{if } x < 0. \end{cases}$$

$$(+\infty)(+\infty) \ = \ (-\infty)(-\infty) \ = \ +\infty, \quad (+\infty)(-\infty) \ = \ -\infty.$$

For any $x, y \in \overline{\mathbb{R}}$ we define $x - y$ as $x + (-y)$ whenever it is defined. Note that for example $(+\infty) - (+\infty)$ is *not* defined. Finally, we define division on $\overline{\mathbb{R}}$

$$x/y = 0 \quad \text{if } x \in \mathbb{R} \text{ and } y \in \{-\infty, +\infty\}.$$
$$(\pm\infty)/y = (1/y)(\pm\infty) \qquad \text{if } y \in \mathbb{R} \setminus \{0\}.$$

The only surprise in the list is that 0 times $\pm\infty$ is 0.

1.3 The Euclidean space

Let $k \in \mathbb{N}$ then by $\mathbb{R}^k$ denote the set of k-tuples of real numbers $\mathbf{x} = (x_1, \ldots, x_k)$, where x_i are the **coordinates** of $\mathbf{x}$. Any two k-tuples can be added together $\mathbf{x} + \mathbf{y} = (x_1 + y_1, \ldots, x_k + y_k)$ and multiplied by a scalar $\lambda \cdot \mathbf{x} = (\lambda x_1, \ldots, \lambda x_k)$. We say that $\mathbb{R}^k$ forms a vector space (more on that in Section 5.1) and call $\mathbf{x}$ a **vector**.

For any two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^k$ we define their **(standard) scalar product** $\langle \mathbf{x}, \mathbf{y} \rangle$ as

$$\langle \mathbf{x}, \mathbf{y} \rangle \;=\; x_1 y_1 + \cdots + x_k y_k \;\in\; \mathbb{R}.$$

Directly from the definition it follows that the scalar product is symmetric, that is, $\langle \mathbf{x}, \mathbf{y} \rangle = \langle \mathbf{y}, \mathbf{x} \rangle$ for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^k$.

Exercise 1.10. Show that the scalar product is **bilinear**, that is, for all $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbb{R}^k$ and $\alpha, \beta \in \mathbb{R}$

$$\langle \alpha\mathbf{x} + \beta\mathbf{y}, \mathbf{z} \rangle \;=\; \alpha\langle \mathbf{x}, \mathbf{z} \rangle + \beta\langle \mathbf{y}, \mathbf{z} \rangle$$

and

$$\langle \mathbf{x}, \alpha\mathbf{y} + \beta\mathbf{z} \rangle \;=\; \alpha\langle \mathbf{x}, \mathbf{y} \rangle + \beta\langle \mathbf{x}, \mathbf{z} \rangle.$$

The scalar product induces the **Euclidean norm** of $\mathbf{x}$

$$\|\mathbf{x}\| \;:=\; \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle} \;=\; \sqrt{x_1^2 + \ldots + x_k^2}.$$

The norm $\|\mathbf{x}\|$ is interpreted as the length of the vector $\mathbf{x}$. This is confirmed by the basic properties it satisfies.

Exercise 1.11. Let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^k$, $\alpha \in \mathbb{R}$. Show that

(i) $\|\mathbf{x}\| \geq 0$ and $\|\mathbf{x}\| = 0$ if and only if $\mathbf{x} = \mathbf{0}$.
(ii) $\|\alpha\mathbf{x}\| = |\alpha|\|\mathbf{x}\|$.

The following fundamental inequality will be used many times in this course.

Theorem 1.7 (Cauchy-Schwarz inequality). *For every $\mathbf{x}, \mathbf{y} \in \mathbb{R}^k$ we have that*

$$|\langle \mathbf{x}, \mathbf{y} \rangle| \ \leq \ \|\mathbf{x}\| \, \|\mathbf{y}\|.$$

Moreover, the inequality is strict unless $\mathbf{x} = \lambda \mathbf{y}$ *for some* $\lambda \in \mathbb{R}$.

Proof. The inequality trivially holds (as equality) when $\mathbf{y} = \mathbf{0}_k$ so suppose $\mathbf{y} \neq \mathbf{0}_k$. For every $\lambda \in \mathbb{R}$

$$\|\mathbf{x} - \lambda \mathbf{y}\|^2 \ = \ \sum_{i=1}^{k} (x_i - \lambda y_i)^2 \ \geq \ 0. \tag{1.2}$$

Using bilinearity and symmetry of the scalar product (c.f. Exercise 1.10), we get

$$\lambda^2 \|\mathbf{y}\|^2 - 2\lambda \langle \mathbf{x}, \mathbf{y} \rangle + \|\mathbf{x}\|^2 \ \geq \ 0. \tag{1.3}$$

Think about the expression on the left as a function of λ; this is a quadratic function with a strictly positive coefficient of λ^2. Therefore it achieves a unique minimum at

$$\lambda^* \ = \ \tfrac{\langle \mathbf{x}, \mathbf{y} \rangle}{\|\mathbf{y}\|^2}.$$

Since the inequality in (1.3) holds for every λ, it also holds for λ^*, which gives

$$\tfrac{\langle \mathbf{x}, \mathbf{y} \rangle^2}{\|\mathbf{y}\|^2} - 2\tfrac{\langle \mathbf{x}, \mathbf{y} \rangle^2}{\|\mathbf{y}\|^2} + \|\mathbf{x}\|^2 \ \geq \ 0.$$

Rearranging gives that $\langle \mathbf{x}, \mathbf{y} \rangle^2 \ \leq \ \|\mathbf{x}\|^2 \|\mathbf{y}\|^2$, which implies the Cauchy-Schwarz inequality. Note that the inequality (1.2) is actually always strict unless $\mathbf{x} - \lambda \mathbf{y} = \mathbf{0}_k$, which completes the proof. □

Another elementary inequality that will be frequently used is the triangle inequality.

Theorem 1.8 (The Euclidean triangle inequality). *For any* $\mathbf{x}, \mathbf{y} \in \mathbb{R}^k$

$$\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|.$$

Proof. Using bilinearity and symmetry of the scalar product we get

$$\|\mathbf{x} + \mathbf{y}\|^2 = \langle \mathbf{x} + \mathbf{y}, \mathbf{x} + \mathbf{y} \rangle = \langle \mathbf{x}, \mathbf{x} \rangle + \langle \mathbf{y}, \mathbf{y} \rangle + 2\langle \mathbf{x}, \mathbf{y} \rangle$$

and the Cauchy-Schwarz inequality further gives

$$\langle \mathbf{x}, \mathbf{x} \rangle + \langle \mathbf{y}, \mathbf{y} \rangle + 2\langle \mathbf{x}, \mathbf{y} \rangle \leq \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 + 2\|\mathbf{x}\| \, \|\mathbf{y}\| = (\|\mathbf{x}\| + \|\mathbf{y}\|)^2.$$

Taking square roots on both sides establishes the triangle inequality. □

Using the norm, we can compute the **Euclidean distance** between any two elements $\mathbf{x}, \mathbf{y}$ of $\mathbb{R}^k$ as the norm $\|\mathbf{x} - \mathbf{y}\|$. The set $\mathbb{R}^k$ equipped with this distance function is called the **Euclidean space**. Denote $d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|$. Directly from the properties of the norm it follows that the Euclidean distance satisfies

1. positive definiteness: $d(x, y) \geq 0$ and $d(x, y) = 0$ if and only if $x = y$.
2. symmetry: $d(x, y) = d(y, x)$.
3. triangle inequality: $d(x, z) \leq d(x, y) + d(y, z)$.

In the next section we generalize the concept of the Euclidean space, where $\mathbb{R}^k$ can be possibly replaced with another set X with a distance function $d : X \times X \to \mathbb{R}$ satisfying the three basic axioms above.

Chapter 2
Topology in metric spaces (2 lectures)

Metaremark: It is fine to focus in this course only on the Euclidean space. In this case, reading through the following section, simply ignore the definition of a metric space: Further on, whenever we mention a metric space X with a distance d, think of $\mathbb{R}^k$ with the Euclidean distance. Almost the whole course can be followed assuming $X = \mathbb{R}^k$.

2.1 Cardinality of sets

If X, Y are two sets then a function $f : X \to Y$ is called a bijection if it is one-to-one and onto, that is,

(i) if $f(x) = f(x')$ then $x = x'$,
(ii) for every $y \in Y$ there exists $x \in X$ such that $f(x) = y$.

In that case we say that X and Y are **bijective**, or that they have the same **cardinality**. A set X is finite if it is bijective with $\{1, \dots, d\}$ for some $d \in \mathbb{N}$. If X is bijective with $\mathbb{N}$ we say that X is **countable**. If X is either finite or countable we say that X is **at most countable**. If X is not at most countable we say it is uncountable.

Directly by definition, X is countable if we can enumerate its elements as $X = \{x_1, x_2, x_3, \dots\}$. Then $f(n) = x_n$ is the defining bijection between $\mathbb{N}$ and X.

Exercise 2.1. Show that $\mathbb{Z}$ is countable.

Exercise 2.2. Show that if X is countable then X^2 is countable.

Theorem 2.1. *Every infinite subset of a countable set is countable.*

Proof. If X is countable then the elements of X can be arranged in a sequence $\{x_n\}_{n\in\mathbb{N}}$. If $E \subset X$ is infinite then let n_1 be the smallest index such that

$x_{n_1} \in E$. Since E is infinite, there exists the smallest $n_2 > n_1$ such that $x_{n_2} \in E$. We proceed recursively obtaining the sequence $\{x_{n_k}\}_{k\in\mathbb{N}}$. The function $f(k) = x_{n_k}$ establishes bijection between $\mathbb{N}$ and E. □

Exercise 2.3. Show that $\mathbb{Q}$ is countable.

Exercise 2.4. Prove that if A is a countable set and B its finite subset then $A \setminus B$ is countable.

Not all sets are countable. For example $\mathbb{R}$ is not. To see this we first show the following.

Theorem 2.2. *Let A be the set of all binary sequences. Then A is uncountable.*

Proof. Suppose A is (infinitely) countable and let $f : \mathbb{N} \to A$ be the corresponding bijection. Let $(b_n)_{n\in\mathbb{N}}$ be a binary sequence such that, for every $n \in \mathbb{N}$, $1 - b_n$ is equal to the n-th element of the sequence $f(n)$. By construction, $(b_n)_{n\in\mathbb{N}}$ is not equal to of the sequences $f(m)$, $m \in \mathbb{N}$, and so it does not lie in the image of f (f cannot be a bijection). □

Theorem 2.3. *The set $[0, 1]$ is uncountable.*

To prove this result, use the binary representation of the real numbers. We omit the details.

Exercise 2.5. Let A be an uncountable set and let B be a countable set. Show that $A \setminus B$ is uncountable.

2.2 Sequences in metric spaces

A **metric space** X is a set, the elements of which are referred as **points** of X, together with a function $d : X \times X \to \mathbb{R}$ satisfying:

1. positive definiteness: $d(p, q) \geq 0$ and $d(p, q) = 0$ if and only if $p = q$.
2. symmetry: $d(p, q) = d(q, p)$ for all $p, q \in X$.
3. triangle inequality: $d(p, q) \leq d(p, r) + d(r, q)$ for all $p, q, r \in X$.

The function d is called the **distance function** (or a **metric**); $d(p, q)$ is the distance between p and q.

The canonical example is the Euclidean space where $X = \mathbb{R}^k$ and $d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x}-\mathbf{y}\|$. The fact that this is indeed a distance function was shown in the end of Section 1.3. Another example is the discrete metric space where X is an arbitrary nonempty set and $d(p, q) = 1$ for all $p \neq q$. The following exercise gives further examples important in applied mathematics.

Exercise 2.6. Let $X = \mathbb{R}^k$ and define

$$d_{\text{sum}}(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{k} |x_i - y_i|, \qquad d_{\max}(\mathbf{x}, \mathbf{y}) = \max_{i=1,\dots,k} |x_i - y_i|.$$

Show that both d_{sum} and $d_{\max}$ define distance functions on $\mathbb{R}^k$.

For any $p \in X$ and $r > 0$ the **neighborhood** of p of **radius** r is

$$N_r(p) \ := \ \{q \in X : \ d(p, q) < r\}. \tag{2.1}$$

In the Euclidean space $N_r(p)$ is simply an open ball with center p and radius r. But for different metric spaces $N_r(p)$ may look nothing like a ball.

Exercise 2.7. Draw the neighborhood $N_1(\mathbf{0}_2)$ in the real plane $\mathbb{R}^2$ with metrics $d_{\max}$ and d_{sum}.

A **sequence of points** in a metric space X is a list $p_1, p_2, \dots$, where the points p_n lie in X. We write (p_n) or $(p_n)_{n\in\mathbb{N}}$. More formally, a sequence of points in X is *any* map $f : \mathbb{N} \to X$. The n-th element of the sequence is $f(n) = p_n$.

Definition 2.1. The sequence (p_n) **converges** to $p \in X$ if

$$\forall \epsilon > 0 \quad \exists N \in \mathbb{N} \text{ such that } \quad \forall n \geq N \quad d(p_n, p) < \epsilon.$$

We write $p_n \to p$ or $\lim_{n\to\infty} p_n = p$; p is called the **limit** of (p_n). If (p_n) does not converge, we say it **diverges**.

In other words: for every $\epsilon > 0$ (think ϵ very small) there exists a moment in the sequence $N \in \mathbb{N}$ such that from this moment on *all* elements in the sequence lie in the ϵ neighborhood of p.

Exercise 2.8. Show that the sequence $x_n = \frac{1}{n}$ in $\mathbb{R}$ converges to 0 as $n \to \infty$. Show that the sequence $x_n = (-1)^n$ diverges.

Exercise 2.9. Suppose that $p_n = p$ for all $n \in \mathbb{N}$, that is, (p_n) is a constant sequence. Show that $p_n \to p$.

Proposition 2.1. *$p_n \to p$ if and only if $d(p_n, p) \to 0$.*

Proof. It follows directly by writing down the definition of $p_n \to p$ and $d(p_n, p) \to 0$. □

Proposition 2.1 allows to reduce convergence analysis to real sequences. The following simple results will be very useful in this context.

Lemma 2.1. *For every real number λ if $x_n \to x$ then $\lambda x_n \to \lambda x$.*

Proof. If $\lambda = 0$ then λx_n is a constant sequence and converges to zero. If $\lambda \neq 0$ fix $\epsilon > 0$ and $\exists N \in \mathbb{N}$ such that $|x_n - x| < \epsilon/|\lambda|$ for all $n \geq N$. For every such n also $|\lambda x_n - \lambda x| = |\lambda||x_n - x| < \epsilon$ which establishes convergence. □

Lemma 2.2. *If (x_n), (y_n) are two real sequences such that $x_n \to x$, $y_n \to y$ then also $x_n + y_n \to x + y$.*

Proof. Fix $\epsilon > 0$ and let $N \in \mathbb{N}$ be such that $|x_n - x| < \epsilon/2$, $|y_n - y| < \epsilon/2$ for all $n \geq N$. For every such n we then have

$$|(x_n + y_n) - (x + y)| \ \leq \ |x_n - x| + |y_n - y| \ < \ \epsilon.$$

□

Exercise 2.10. Let (x_n) and (y_n) be two real sequences Suppose $0 \leq x_n \leq y_n$ for all $n \in \mathbb{N}$. Show that if $y_n \to 0$ then $x_n \to 0$.

Proposition 2.2. *If (p_n) converges to p then p is unique.*

Proof. Suppose that p, p' be two distinct points such that $p_n \to p$ and $p_n \to p'$. By Proposition 2.1, $d(p_n, p) \to 0$ and $d(p_n, p') \to 0$. By the triangle inequality

$$d(p, p') \overset{\triangle}{\leq} d(p, p_n) + d(p', p_n).$$

By Lemma 2.2, the right hand side of this inequality converges to 0. This is only possible if $d(p, p') = 0$ for otherwise taking $\epsilon = d(p, p')$ would give a contradiction with this convergence. We conclude that $p = p'$. □

Remark 2.1. In the above proof we introduced a convention that an inequality that follows from the triangle inequality is written as $\overset{\triangle}{\leq}$.

A set $E \subset X$ is **bounded** if there exists $p \in X$ and $r > 0$ such that $E \subset N_r(p)$. A sequence (p_n) in X is bounded if its range is a bounded set.

Proposition 2.3. *Every convergent sequence is bounded.*

Proof. Suppose $p_n \to p$ then there exists $N \in \mathbb{N}$ such that $d(p, p_n) < 1$ for all $n \geq N$. Put

$$r = \max\{1, d(p, p_1), \ldots, d(p, p_N)\}$$

then $d(p, p_n) \leq r$ for all $n \in \mathbb{N}$ and so the range of (p_n) is contained in $N_r(p)$. □

We conclude this section by discussing subsequences.

Definition 2.2. Let (p_n) be a sequence in a metric space X. Let (n_k) for $k \in \mathbb{N}$ be a sequence of natural numbers such that $n_1 < n_2 < \ldots$ Then the sequence (p_{n_k}), that is, $p_{n_1}, p_{n_2}, \ldots$ is called a **subsequence** of (p_n). If (p_{n_k}) converges as $k \to \infty$, its limit is called a **subsequential limit** of (p_n).

For example if $x_n = (-1)^n$ then (x_n) does not converge but it has two subsequential limits -1 and 1.

Proposition 2.4. *We have $p_n \to p$ if and only if $p_{n_k} \to p$ for* every *subsequence (p_{n_k}).*

Proof. Since (p_n) is its own subsequence, the left direction is immediate. For the forward direction note that for any $\epsilon > 0$ there exists $N \in \mathbb{N}$ such that $d(p, p_n) < \epsilon$ for all $n \geq N$. Fix a subsequence (p_{n_k}) and let $K \in \mathbb{N}$ be any natural number such that $n_K \geq N$ (it must exist because (n_k) is a strictly increasing sequence of natural numbers). Then for all $k \geq K$ we have $d(p, p_{n_k}) < \epsilon$. Since ϵ was arbitrary, we conclude that $p_{n_k} \to p$ as $k \to \infty$. □

The concept of subsequences is useful to define limit inferior and limit superior.

Definition 2.3. Let $E \subseteq \mathbb{R}$ be the set of subsequential limits of a real sequence (x_n). Define

$$\liminf_{n\to\infty} x_n := \inf E, \qquad \limsup_{n\to\infty} x_n := \sup E.$$

Exercise 2.11. Show that $x_n \to x$ if and only if $\liminf_{n\to\infty} x_n = \limsup_{n\to\infty} x_n$.

2.3 Point sequences in Cartesian products

Let X_i for $i = 1, \ldots, k$ be metric spaces with distance functions d_i respectively. There are three natural ways to define a distance function on the Cartesian product $X = X_1 \times \cdots \times X_k$. For $\mathbf{p} = (p_1, \ldots, p_k)$ and $\mathbf{q} = (q_1, \ldots, q_k)$ in X define

$$\begin{aligned} d_E(\mathbf{p}, \mathbf{q}) &= \sqrt{d_1(p_1, q_1)^2 + \cdots + d_k(p_k, q_k)^2} \\ d_{\max}(\mathbf{p}, \mathbf{q}) &= \max_{i=1,\ldots,k} d_i(p_i, q_i) \\ d_{\text{sum}}(\mathbf{p}, \mathbf{q}) &= d_1(p_1, q_1) + \cdots + d_k(p_k, q_k) \end{aligned}$$

Exercise 2.12. Show that these three formulas define valid metrics on X.

In some aspects the three metrics are similar. We have the following result.

Proposition 2.5. $d_{\max} \leq d_E \leq d_{\text{sum}} \leq k d_{\max}$.

Proof. Dropping the smaller terms inside the square root shows that $d_{\max} \leq d_E$; comparing the square of d_E and the square of d_{sum} shows that the latter has the terms of the former and the cross terms besides, so $d_E \leq d_{\text{sum}}$; and clearly d_{sum} is no larger than k times its greatest term, so $d_{\text{sum}} \leq k d_{\max}$. □

This sequence of inequalities has very important consequences.

Theorem 2.4. *Let $X = X_1 \times \cdots \times X_k$ be a product of k metric spaces (X_i, d_i). The following are equivalent for a sequence $\mathbf{p}_n = (p_{1n}, \ldots, p_{kn})$ in $X = X_1 \times \cdots \times X_k$:*

(i) $(\mathbf{p}_n)$ converges with respect to the metric $d_{\max}$.
(ii) $(\mathbf{p}_n)$ converges with respect to the metric d_E.
(iii) $(\mathbf{p}_n)$ converges with respect to the metric d_{sum}.
(iv) For every $i = 1, \ldots, k$ the sequence (p_{in}) converges in X_i.

Proof. By Proposition 2.1, $\mathbf{p}_n \to \mathbf{p}$ in X with a metric d if and only if $d(\mathbf{p}_n, \mathbf{p}) \to 0$. By Proposition 2.5,

$$0 \;\le\; d_{\max}(\mathbf{p}_n, \mathbf{p}) \;\le\; d_E(\mathbf{p}_n, \mathbf{p}) \;\le\; d_{\text{sum}}(\mathbf{p}_n, \mathbf{p}) \;\le\; k d_{\max}(\mathbf{p}_n, \mathbf{p}).$$

Thus, converging any of the three metrics to zero implies convergence to zero of the other two, which shows equivalence of (i),(ii), and (iii).

To show that all *four* items are equivalent, it is enough to show $(i) \Leftrightarrow (iv)$. Item (i) means that $\max_{i=1,\ldots,k} d_i(p_{in}, p_i) \to 0$. Equivalently, $d_i(p_{in}, p_i) \to 0$ for all i, which is another way to say that $p_{in} \to p$ for all i. □

Remark 2.2. If convergence in one metric is equivalent to convergence in another metric we say that these metrics are **equivalent**. Theorem 2.4 implies that d_E, $d_{\max}$, and d_{sum} are equivalent on X.

Corollary 2.1 (Convergence in $\mathbb{R}^k$). *A sequence of vectors $(\mathbf{x}_n)$ in the Euclidean space converges if and only if each component sequence (x_{in}) converges, $1 \le i \le k$. The limit of the vector sequence is the vector*

$$\mathbf{x} \;=\; \lim_{n\to\infty} \mathbf{x}_n \;=\; \left(\lim_{n\to\infty} x_{1n}, \ldots, \lim_{n\to\infty} x_{kn} \right).$$

2.4 Closed sets and open sets

Recall that $N_r(p)$ denotes the neighborhood of p of radius r, see (2.1). We say that $p \in E$ is an **interior point** of $E \subset X$ if $N_r(p) \subset E$ for some $r > 0$.

Definition 2.4. E is an **open set** (in X) if each point of E is an interior point of E.

Example 2.1. Three examples of open sets: (i) the open interval $(0, 1)$ is an open set, (ii) $\mathbb{R}^k$, (iii) the set of $n \times n$ matrices that are invertible.

Exercise 2.13. Show that for every $r > 0$ and every $p \in X$, the neighborhood $N_r(p)$ is an open set.

A point $p \in X$ is a **limit of** E[1] if there exists a sequence (p_n) in E such that $p_n \to p$. Every $p \in E$ is a limit of E because it is a limit of the constant sequence (p_n) where $p_n = p$ for all $n \in \mathbb{N}$.

Definition 2.5. E is a **closed set** (in X) if it contains all its limits.

Example 2.2. Three examples of closed sets: (i) closed interval $[0, 1]$, (ii) $\mathbb{R}^k$, (iii) the set of orthogonal $n \times n$ matrices.

Remark 2.3. The underlying metric space is important in the above definitions. For example, $(0, 1)$ is open in $X = \mathbb{R}$ but it is closed (and open) in $X = (0, 1)$. Make sure you understand why.

The following simple lemma will be used in many proofs.

Lemma 2.3. *If (p_n) is a sequence in X such that $d(p, p_n) < 1/n$ for all $n \geq 1$, then $p_n \to p$.*

Proof. Follows immediately from Proposition 2.1. □

A point $p \in E$ is an **isolated point** of E if there exists a neighborhood $N_r(p)$ such that $N_r(p) \cap E = \{p\}$. For example, the point $(0, 0)$ is an isolated point of $E = \{(x, y) \in \mathbb{R}^2 : \ x^3 - x^2 - y^2 \geq 0\}$, see Figure 2.1.

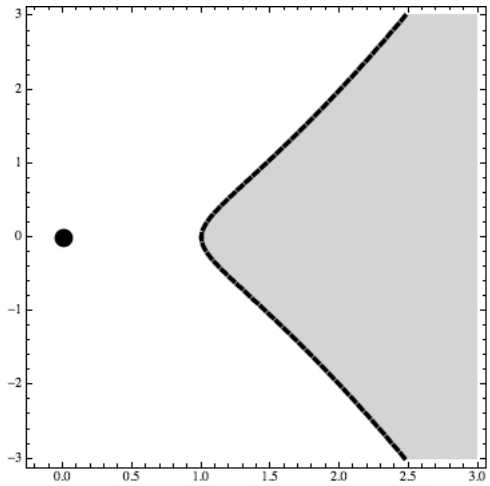

Fig. 2.1 The set of points satisfying $x^3 - x^2 - y^2 \geq 0$.

The following lemma gives an alternative characterization of the isolated point.

[1] Note that this notion is not the same as the limit point in Rudin's book. In this we follow the book of Pugh.

Lemma 2.4. *A point $p \in E$ is an isolated point of E if and only if for every sequence (p_n) in E converging to p there must exist $N \in \mathbb{N}$ such that $p_n = p$ for all $n \geq N$.*

Proof. If p is an isolated point of E then $N_\epsilon(p) \cap E = \{p\}$ for some $\epsilon > 0$. For the definition of convergence to be satisfied, there must exist $N \in \mathbb{N}$ such that $p_n = p$ for all $n \geq N$.

For the opposite direction suppose that p is not an isolated point. Then for each $n \in N$ the neighborhood $N_{1/n}(p)$ contains a point $p_n \in E$ such that $p_n \neq p$. By Lemma 2.3, $p_n \to p$. For this sequence there does not exist $N \in \mathbb{N}$ such that $p_n = p$ for $n \geq N$. □

Exercise 2.14. If E is a finite set of points then all points of E are isolated (why?). Conclude that every finite set of points is closed.

There are examples of infinite sets whose all points are isolated. For example, the set of natural numbers $\mathbb{N}$ is closed in $X = \mathbb{R}$.

Theorem 2.5. *$E \subset X$ is open if and only if the complement of E is closed.*

Proof. "⇒" Suppose $p_n \to p$ and $p_n \in E^c$, we need to show that $p \in E^c$. Well, if $p \notin E^c$ then $p \in E$, and since E is open, there exists $r > 0$ such that $d(p, q) < r$ implies $q \in E$. Since $p_n \to p$ we have $d(p_n, p) < r$ for large n, which implies $p_n \in E$, contrary to the sequence being in E^c. We conclude that p must lie in E^c, which proves that E^c is closed.

"⇐" Now assume that E^c is closed and take any $p \in E$. We want to show that p is an interior point of E. If there is no $r > 0$ such that $d(p, q) < r$ implies that $q \in E$ then we can take $r = 1/n$ for $n \geq 1$ to construct a sequence $p_n \in E^c$ such that $d(p_n, p) < 1/n$. By Lemma 2.3, this sequence in E^c converges to $p \in E$. This contradicts closedness of E^c. □

The **topology** $\mathcal{T}$ of a metric space X is the collection of all its open sets.

Theorem 2.6. *$\mathcal{T}$ has three properties:*

(a) Every union of open sets is an open set.
(b) The intersection of finitely many open sets is an open set.
(c) $\emptyset$ and X are open sets.

Proof. (a) Let $G = \bigcup_{\alpha \in A} G_\alpha$, where A is arbitrary and all G_α are open. If $p \in G$ then $p \in G_\alpha$ for some α. Since G_α is open, p is an interior point of G_α, that is, there exists $r > 0$ such that $N_r(p) \subset G_\alpha$. But then $N_r(p) \subset G$ and so p is an interior point of G. Since p was arbitrary, G is open.

(b) Let $G = \bigcap_{i=1}^n G_i$. If $p \in G$, then $p \in G_i$ for all $i = 1, \ldots, n$ and there exist neighborhoods $N_{r_i}(p) \subset G_i$ for some $r_i > 0$. Take $r = \min\{r_1, \ldots, r_n\}$ then $r > 0$ and $N_r(p) \subset G_i$ for all i and so $N_r(p) \subset G$, which proves that G is open.

(c) In both cases clearly all points are interior. □

An infinite intersection of open sets does not need to be open. For example, let $G_i = (-\frac{1}{n}, \frac{1}{n})$ then $G = \bigcap_{i=1}^{\infty} G_i = \{0\}$, which is not open (in $\mathbb{R}$).

Exercise 2.15. Show that every intersection of closed sets is closed.

Exercise 2.16. Show that every open subset of $\mathbb{R}$ is an at most countable union of open intervals.

Let E be a subset of a metric space X. Its closure, interior, and boundary are defined as follows:

1. **closure** $\overline{E}$ is the set of limits of E.
2. **interior** E° is the set of all interior points of E.
3. **boundary** $\partial E = \overline{E} \cap \overline{E^c}$.

Theorem 2.7. *The closure of $E \subset X$ is a closed set and $E = \overline{E}$ if and only if E is closed.*

Proof. Suppose that $p_n \to p$ and each p_n lies in $\overline{E}$. We claim that $p \in \overline{E}$. Since p_n is a limit of E there is a sequence $(p_{n,k})_{k\in\mathbb{N}}$ in E converging to p_n as $k \to \infty$. Thus, for every $n \in \mathbb{N}$ there exists $q_n = p_{n,k(n)} \in E$ such that $d(p_n, q_n) < \frac{1}{n}$. Then

$$d(p, q_n) \;\le\; d(p, p_n) + d(p_n, q_n) \;\to\; 0 \qquad \text{as } n \to \infty$$

and so $q_n \to p$, which implies that $p \in \overline{E}$. For the second part: "$\Rightarrow$" is part of the first statement. "$\Leftarrow$" follows from the definition of a closed set. □

Exercise 2.17. Show that the boundary of E is closed.

Exercise 2.18. Show that the interior E° of E is the largest (with respect to inclusion) open set contained in E. More precisely, E° is is an open set contained in E with the property that $U \subset E$ and U open implies that $U \subset E^\circ$. Conclude that E^0 is the union of all open sets contained in E.

Exercise 2.19. Similarly like in the exercise above show that $\overline{E}$ is the smallest closed set containing E. In other words, $\overline{E}$ is the intersection of all closed sets containing E.

We finish this section with an important result on subsets of the real line.

Theorem 2.8. *If $E \subset \mathbb{R}$ is bounded above, then $\sup E \in \overline{E}$. If E is bounded below, then $\inf E \in \overline{E}$. Hence $\sup E$ and $\inf E$ lie in E if E is closed.*

Proof. We proof only the sup-part of the statement and the arguments for infimum are similar. Let $\alpha = \sup E$. If $\alpha \in E$ then $\alpha \in \overline{E}$. Assume $\alpha \notin E$. Since α is the least upper bound, for every $n \in \mathbb{N}$ there exists $x_n \in E$ such that $\alpha - \frac{1}{n} < x_n < \alpha$. By Lemma 2.3, the sequence (x_n) converges to α and so α is a limit of E (lies in $\overline{E}$). The second part follows directly from Theorem 2.7. □

Chapter 3
Continuity (2 lectures)

3.1 Functional limits and continuity

So far we have studied metric spaces. Now we add to this picture also mappings between metric spaces.

Definition 3.1 (Functional limit). Let X, Y be metric spaces. Let $E \subset X$ and $f : E \to Y$ and suppose that p is a limit of E but not an isolated point of E. Then we say that $f(x) \to q$ as $x \to p$, or $\lim_{x\to p} f(x) = q$ if for every sequence (p_n) in E such that $p_n \neq p$ and $p_n \to p$ we have $f(p_n) \to q$.

To understand why we rule out the possibility for p to be an isolated point recall from Lemma 2.4 that if p is an isolated point then there is *no* sequence in E such that $p_n \neq p$ and $p_n \to p$. In particular, the condition in the definition would hold for every real q!

Remark 3.1. If $\lim_{x\to p} f(x)$ exists then it is unique because $\lim_{n\to\infty} f(p_n)$ is unique.

Example 3.1. Let f and g be real-valued functions given by

$$f(x) = \begin{cases} 1 & x = 0 \\ 0 & \text{otherwise} \end{cases} \qquad \text{and} \qquad g(x) = \begin{cases} 1 & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

then $\lim_{x\to 0} f(x) = 0$ however $\lim_{x\to 0} g(x)$ does not exist.

Example 3.2. Let $f : \mathbb{R} \to \mathbb{R}$ be given by $f(x) = \frac{|x|}{x}$. Then $\lim_{x\to 0} f(x)$ does not exist. For example, taking sequences $x_n = \frac{1}{n}$ and $x'_n = -\frac{1}{n}$ we get $f(x_n) = 1$ and $f(x'_n) = -1$ for all n and so $f(x_n) \to 1$, $f(x'_n) \to -1$.

Exercise 3.1. Let $f, g : \mathbb{R} \to \mathbb{R}$. Conclude from Lemma 2.2 that if $\lim_{x\to p} f(x) = q$ and $\lim_{x\to p} g(x) = q'$ then $\lim_{x\to p}(f + g)(x) = q + q'$.

Proposition 2.1 has the following translation to functional limits.

Remark 3.2. We have $f(x) \to q$ if and only if $d_Y(f(x), q) \to 0$.

In the following example we discuss a more abstract situation to emphasise some subtle points in the definition of the functional limit.

Example 3.3. If a is a real number, then $\lim_{x\to a} \frac{a^2-x^2}{a-x} = 2a$; in particular the limit exists. In this problem we investigate to what extent this still holds for matrices. Let $A \in \mathbb{R}^{k\times k}$ be a square matrix. What does it mean to say that

$$\lim_{X\to A} (A - X)^{-1}(A^2 - X^2) \qquad \text{exists?}$$

We follow the definition of the functional limit. Let $F(X) = (A-X)^{-1}(A^2 - X^2)$. The domain of $F(X)$, denoted by E, is the set of all $X \in \mathbb{R}^{k\times k}$ such that $A - X$ is invertible. Although $A \notin E$, A is a limit point of E. With this notation $\lim_{X\to A} F(X)$ exists if there exists a matrix B such that for every sequence X_n of matrices in E that converges to A, with $X_n - A$ invertible, we have $F(X_n) \to B$ (a suitable distance for matrices to define convergence is discussed in Section 5.4). Now consider some special cases. If A is the identity matrix $\mathbb{I}_k$, we can use the same trick as in the univariate case and rewrite $(\mathbb{I}_k^2 - X^2) = (\mathbb{I}_k - X)(\mathbb{I}_k + X)$, which gives

$$(\mathbb{I}_k - X)^{-1}(\mathbb{I}_k^2 - X^2) = \mathbb{I}_k + X.$$

Therefore, if $X_n \to \mathbb{I}_k$ as $n \to \infty$, then $(\mathbb{I}_k - X)^{-1}(\mathbb{I}_k^2 - X^2) \to B = 2\mathbb{I}_k$. Now suppose that $k = 2$ and

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

For this A we can easily find two different sequences X_n in E for which the limits of $(A - X_n)^{-1}(A^2 - X_n^2)$ are different. For example

$$\begin{bmatrix} \frac{1}{n} & 1 \\ 1 & \frac{1}{n} \end{bmatrix} \qquad \text{and} \qquad \begin{bmatrix} \frac{1}{n} & 1 \\ 1 & -\frac{1}{n} \end{bmatrix}.$$

We now introduce one of the most fundamental concepts of real analysis.

Definition 3.2. A function $f : E \to Y$ is **continuous at a point** $p \in E$ if for every sequence (p_n) such that $p_n \to p$ also $f(p_n) \to f(p)$. We say that f is **continuous** if it is continuous at every point $p \in E$.

Proposition 3.1. *For any fixed* $\mathbf{y} \in \mathbb{R}^k$ *the function* $f(\mathbf{x}) = \|\mathbf{x} - \mathbf{y}\|$ *is continuous. Taking* $\mathbf{y} = \mathbf{0}$ *gives that the norm is a continuous function.*

The proof is left as an easy exercise.

Lemma 2.4 shows that if p is an isolated point of $E \subset X$ then for every $p_n \to p$ we also have that $f(p_n) = f(p)$ from certain point on. In particular, f is continuous at p. If p is a limit of E but not an isolated point of E then we have the following

Proposition 3.2. *If p is not an isolated point of E and $f : E \to Y$ then f is continuous at p if and only if* $\lim_{x\to p} f(x) = f(p)$.

Exercise 3.2. Prove Proposition 3.2.

Exercise 3.3. Let $f : X \to Y$. Show that if there exists $M \in \mathbb{R}$ such that $d_Y(f(p_1), f(p_2)) \leq M d_X(p_1, p_2)$ for any two $p_1, p_2 \in X$ then f is continuous.

The following result follows easily from the definition.

Theorem 3.1. *A composition of continuous functions is continuous.*

Proof. Let $p_n \to p$. Since f is continuous, $f(p_n) \to f(p)$. Since g is continuous, we get that $g(f(p_n)) \to g(f(p))$ and therefore $g \circ f$ is continuous. □

We are often studying mappings $\mathbf{f} : E \to \mathbb{R}^m$, where $E \subset X$. Every such function is given by specifying its components $f_i : E \to \mathbb{R}$ for $i = 1, \ldots, m$. We write $\mathbf{f} = (f_1, \ldots, f_m)$. The following result shows that studying limits of $\mathbf{f}$ can be reduced to studying limits of its components.

Theorem 3.2. *Let $\mathbf{f} : E \to \mathbb{R}^m$, where $E \subset X$. Suppose p is a limit of E but not an isolated point of E. Then*

$$\lim_{x\to p} \mathbf{f}(x) = \mathbf{q} \qquad \Longleftrightarrow \qquad \lim_{x\to p} f_i(x) = q_i \text{ for all } i = 1, \ldots, m.$$

Proof. We have $\lim_{x\to p} \mathbf{f}(x) = \mathbf{q}$ if and only if for every (p_n) such that $p_n \to p$ also $\mathbf{f}(p_n) \to \mathbf{q}$. By Corollary 2.1 the sequence $\mathbf{f}(p_n)$ converges to $\mathbf{q}$ if and only if each $f_i(p_n)$ converges and

$$\lim_{n\to\infty} \mathbf{f}(p_n) \;=\; (\lim_{n\to\infty} f_1(p_n), \ldots, \lim_{n\to\infty} f_m(p_n)) \;=\; (q_1, \ldots, q_n).$$

The last equality gives that $\lim_{x\to p} f_i(x) = q_i$ for all $i = 1, \ldots, m$. □

This implies that continuity can also be checked component-wise.

Theorem 3.3. *Suppose that $\mathbf{f} : E \to \mathbb{R}^m$ where $E \subset X$. Then $\mathbf{f}$ is continuous if and only if each component $f_i : E \to \mathbb{R}$ is continuous.*

Proof. If p is an isolated point of E then all functions are continuous at this point. Otherwise $\mathbf{f}$ is continuous if and only if $\lim_{x\to p} \mathbf{f}(x) = \mathbf{f}(p)$ (Proposition 3.2). By Theorem 3.2 this is equivalent to $\lim_{x\to p} f_i(x) = f_i(p)$, which is equivalent to all f_i being continuous. □

3.2 Continuity of arithmetic in $\mathbb{R}$

We will now discuss continuity of basic arithmetic operations in $\mathbb{R}$. Addition is a mapping $\text{Sum} : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ that assigns to (x, y) the real number $x + y$. Subtraction and multiplication are also such mappings. Division is a mapping $\mathbb{R} \times (\mathbb{R} \setminus \{0\}) \to \mathbb{R}$ that assigns to (x, y) the number x/y.

Theorem 3.4. *The arithmetic operations of $\mathbb{R}$ are continuous.*

Proof. Let $(x_n, y_n) \to (x, y)$. By Corollary 2.1, equivalently $x_n \to x$ and $y_n \to y$. The fact that $x_n + y_n \to x + y$ follows by Lemma 2.2. The proof for subtraction is exactly the same. To show that multiplication is continuous note that

$$|x_n y_n - xy| \;\le\; |x_n - x|\,|y_n| + |x|\,|y_n - y| \;\le B(|x_n - x| + |y_n - y|),$$

where B is some positive number that bounds both $|x|$ and $|y_n|$ for all $n \in \mathbb{N}$ (B exists by Proposition 2.3). Using Lemma 2.1 and Lemma 2.2, the right hand side converges to zero. We conclude that multiplication is also continuous.

For division, since $y_n \to y$ with $y, y_n \neq 0$, there exists $N \in \mathbb{N}$ such that $|y_n - y| < |y|/2$ for all $n \ge N$. Since $||y_n| - |y|| \le |y_n - y|$, we get $|y_n| > |y|/2$, and so $\frac{1}{|y_n y|} < \frac{2}{|y|^2}$ for all $n \ge N$. As a result, we can bound above the following expression for all $n \ge N$

$$\left|\tfrac{x_n}{y_n} - \tfrac{x}{y}\right| \;=\; \left|\tfrac{x_n y - x y_n}{y_n y}\right| \;\overset{\triangle}{\le}\; \tfrac{|x_n - x|\,|y| + |x|\,|y_n - y|}{|y_n y|} \;\le\; B(|x_n - x| + |y_n - y|),$$

where B is some positive constant. This again implies continuity. □

Definition 3.3. If $f, g : E \to \mathbb{R}$, $E \subset X$, then $f+g$, $f-g$, $f \cdot g$ are real-valued functions on E with values $(f+g)(x) = f(x)+g(x)$, $(f-g)(x) = f(x)-g(x)$, and $(f \cdot g)(x) = f(x)g(x)$. For all points such that $g(x) \neq 0$ we can also define f/g through $(f/g)(x) = f(x)/g(x)$.

Theorem 3.5. *The sums, differences, products, and quotients of real-valued continuous functions are continuous (as long as the denominators are not equal to zero).*

Proof. Take, for example, the sum $f + g$ where $f, g : E \to \mathbb{R}$ are continuous. It is the composition of functions

$$\begin{gathered} E \;\xrightarrow{(f,g)}\; \mathbb{R} \times \mathbb{R} \;\xrightarrow{\text{Sum}}\; \mathbb{R} \\ x \mapsto (f(x), g(x)) \mapsto f(x) + g(x), \end{gathered}$$

where the first arrow is continuous by Theorem 3.3 and the second by Theorem 3.4. Thus, the composition is continuous by Theorem 3.1. The proof for $f - g$, fg and f/g is the same. □

Theorem 3.5 has a counterpart in $\mathbb{R}^k$.

Theorem 3.6. *If $\mathbf{f}, \mathbf{g} : E \to \mathbb{R}^k$ are continuous functions then $\mathbf{f} + \mathbf{g}$ and $\langle \mathbf{f}(x), \mathbf{g}(x) \rangle$ are continuous.*

Proof. If $\mathbf{f}$, $\mathbf{g}$ are continuous then all components f_i, g_i are continuous. By the previous theorem, $f_i + g_i$ are continuous for $i = 1, \ldots, m$ and so $\mathbf{f} + \mathbf{g}$ is continuous. For the scalar product: since all $f_i g_i$ are continuous, so is their sum. □

Exercise 3.4. Show how Proposition 3.1 now follows directly from Theorem 3.6.

Exercise 3.5. Show that all polynomial functions are continuous.

The proof of Theorem 3.4 can be also recycled to prove the following result.

Proposition 3.3. *Let $f, g : E \to \mathbb{R}$, $E \subset \mathbb{R}$ open, and $p \in E$. Suppose $\lim_{x\to p} f(x) = a$, $\lim_{x\to p} g(x) = b$ then*

$$\lim_{x\to p}(f+g)(x) = a + b, \quad \text{and} \quad \lim_{x\to p} f(x)g(x) = ab.$$

Moreover, if $b \neq 0$ then also $\lim_{x\to p} \frac{f(x)}{g(x)} = \frac{a}{b}$.

Proof. We prove that $\lim_{x\to p} f(x)+g(x) = a+b$ all other cases being similar. Since $\lim_{x\to p} f(x) = a$ and $\lim_{x\to p} g(x) = b$ for every sequence $x_n \to p$ such that $x_n \neq p$ we have $f(x_n) \to a$ and $g(x_n) \to b$. By the proof of Theorem 3.4 the sum $f(x_n) + g(x_n) = (f + g)(x_n)$ converges to $a + b$. □

Example 3.4. In Example 2.1 we learned that the set of invertible $n \times n$ matrices is open in $\mathbb{R}^{n\times n}$. Proving this without discussing continuity may be tedious. Continuity gives us tools to provide a simple argument. A matrix A is invertible if and only if $\det(A) \neq 0$. It is then enough to show that the set $\{A : \det(A) = 0\}$ is closed. Since $\det(A)$ is a polynomial in the entries of A, it is a continuous function on $\mathbb{R}^{n\times n}$. If $A_n \to A$ with $\det(A_n) = 0$ then continuity implies that also $\det(A) = 0$, which shows that $\{A : \det(A) = 0\}$ is closed.

Exercise 3.6. Suppose that $f : X \to \mathbb{R}$ is a continuous function. Show that the set of zeros of f is a closed set. How about any level set of f?

3.3 Alternative characterizations of continuity*

The standard way of defining continuous functions is, so called, (ϵ, δ)-condition. For completeness of the discussion we formulate it as a theorem. We then discuss other popular characterisations of continuity that have some advantages.

Theorem 3.7. *A function $f : X \to Y$ is continuous at p if and only if*

$$\forall \epsilon > 0 \;\; \exists \delta > 0 \text{ s.t. } \quad \Big(d(p,q) < \delta \;\; \Rightarrow \;\; d(f(p), f(q)) < \epsilon \Big).$$

Proof. Suppose that f satisfies the (ϵ, δ)-condition condition and $p_n \to p$. Then $f(p_n)$ is a sequence in Y. The condition implies that for every $\epsilon > 0$, there exists $\delta > 0$ such that $d(x,p) < \delta$ implies $d(f(x), f(p)) < \epsilon$. Convergence implies that there exists $N \in \mathbb{N}$ such that $d(p_n, p) < \delta$ for all $n \geq N$. Then also $d(f(p_n), f(p)) < \epsilon$ and so $f(p_n) \to f(p)$, which shows continuity of f.

We prove the converse in contrapositive form: if f does not satisfy the (ϵ, δ)-condition then there exists $\epsilon > 0$ such that for all $\delta > 0$ we have a point q such that $d(p,q) < \delta$ and $d(f(p), f(q)) \geq \epsilon$. Taking $\delta = 1/n$ for $n \in \mathbb{N}$ we construct a sequence q_n converging to p (by Lemma 2.3) such that $f(q_n)$ does not converge to $f(p)$. This shows that f is not continuous. □

Another important reformulation of continuity builds on the concept of preimage. Let $f : X \to Y$ be given. The **preimage** of a set $V \subset Y$ is

$$f^{\text{pre}}(V) \;:=\; \{p \in X : \; f(p) \in V\}. \tag{3.1}$$

For example, if $f : \mathbb{R}^2 \to \mathbb{R}$ is defined by $f(x,y) = x^2 + y^2 + 2$ then the preimage of the interval $[3,6]$ is the annulus in the plane with inner radius 1 and outer radius 2.

Theorem 3.8. *A function $f : X \to Y$ is continuous if and only if the preimage $f^{\text{pre}}(V)$ of any open set $V \subset Y$ is open in X.*

Proof. Suppose that f is continuous on X and let $V \subset Y$ be open. We need to show that every point $p \in f^{\text{pre}}(V)$ is an interior point of $f^{\text{pre}}(V)$. Since V is open, $f(p)$ is an interior point of V, that is, there exists $\epsilon > 0$ such that $d_Y(f(p), y) < \epsilon$ implies that $y \in V$. By Theorem 3.7, there exists $\delta > 0$ such that $d_X(p,x) < \delta$ implies that $d_Y(f(p), f(x)) < \epsilon$, or in other words, $f(x) \in V$. Thus, $x \in f^{\text{pre}}(V)$ as soon as $d_X(p,x) < \delta$. This proves that x is an interior point.

Conversely, suppose that $f^{\text{pre}}(V)$ is open in X for every open set V in Y. Fix $p \in X$ and $\epsilon > 0$. The set $V = \{y : d_Y(f(p), y) < \epsilon\}$ is a neighborhood in Y and so open. Since $f^{\text{pre}}(V)$ is open, there exists $\delta > 0$ such that $x \in f^{\text{pre}}(V)$ as soon as $d_X(p,x) < \delta$. But if $x \in f^{\text{pre}}(V)$. then $f(x) \in V$ and so $d_Y(f(x), f(p)) < \epsilon$. This is precisely the definition of continuity of f at p as characterized by Theorem 3.7. □

Exercise 3.7. Prove that a function $f : X \to Y$ is continuous at $p \in X$ if and only if for any open $V \subset Y$ such that $p \in f^{\text{pre}}(V)$ the point p is an interior point of $f^{\text{pre}}(V)$.

Exercise 3.8. Prove that a function $f : X \to Y$ is continuous if and only if the preimage $f^{\text{pre}}(C)$ of any closed set $C \subset Y$ is closed in X.

3.4 Semicontinuity*

Discussing correspondences in Chapter 13 we will find conceptually useful to derive the following alternative way of thinking about continuity of real-valued functions.

Definition 3.4. A function $f : E \to \mathbb{R}$ is is **lower semicontinuous at** p if p is an interior point of the set $\{x : f(x) > f(p) - \epsilon\}$. Similarly, f is **upper semicontinuous at** $p \in E$ if for each $\epsilon > 0$, p is an interior point of $\{x : f(x) < f(p) + \epsilon\}$.

Recall the definition of limit inferior and limit superior given in Definition 2.3. This definition can be extended to functional limits of real valued functions. Namely, consider all sequences x_n, $x_n \to p$, $x_n \neq p$ and all converging subsequences $f(x_{n_k})$. Then $\liminf_{x\to p} f(x) \in \overline{\mathbb{R}}$ is the infimum of the set of all such subsequential limits.

The following result offers a useful reformulation of Definition 3.4.

Theorem 3.9. *The function $f : E \to \overline{\mathbb{R}}$ is lower semicontinuous at $p \in E$ if and only if $\liminf_{x\to p} f(x) \geq f(p)$. Similarly, $f : E \to \overline{\mathbb{R}}$ is upper semicontinuous at $p \in E$ if and only if $\limsup_{x\to p} f(x) \leq f(p)$.*

Remark 3.3. The condition $\liminf_{x\to p} f(x) \geq f(p)$ can be translated as follows

$$x_n \to p \qquad \Longrightarrow \qquad \liminf_{n\to\infty} f(x_n) \geq f(p).$$

Proof of Theorem 3.9. We will provide the proof for lower semicontinuity. For the right implication, let $x_n \in E$ be any sequence such that $x_n \to p$ and $f(x_n)$ converges ($\lim_{n\to\infty} f(x_n) = q$). By the definition of lower semicontinuity, for every $\epsilon > 0$ there exists $N \in \mathbb{N}$ such that for every $n \geq N$ $f(x_n) > f(p) - \epsilon$. Since this holds for an arbitrary ϵ, we conclude that $q \geq f(p)$. This proves that $\liminf_{x\to p} f(x) \geq f(p)$. For the other implication, suppose that for some $\epsilon > 0$ p is not an interior point of $x : f(x) > f(p) - \epsilon\}$. Then, there exists a sequence $x_n \to p$ such that $f(x_n) \leq f(p) - \epsilon$. In particular, for any subsequence (x_{n_k}) such that $f(x_{n_k})$ converges, the limit is less than or equal to $f(p) - \epsilon$. But this implies that $\liminf_{x\to p} f(x) \leq f(p) - \epsilon < f(p)$, which is a contradiction. □

As a corrollary, we get the following characterization of continuous functions.

Theorem 3.10. *A function $f : E \to \mathbb{R}$ is continuous at p if and only if it is both upper and lower semicontinuous at p.*

We say that $f : E \to \mathbb{R}$ is upper semicontinuous if it is upper semicontinuous at every $p \in E$.

Exercise 3.9. Show that $f : E \to \mathbb{R}$ is upper semicontinuous if and only if $\{x : f(x) < a\}$ is open for every $a \in \mathbb{R}$.

Chapter 4
Compactness and completeness (2 lectures)

4.1 Compact sets

The following is probably the most important concept of real analysis.

Definition 4.1. A subset E of a metric space X is (sequentially) **compact** if every sequence $(p_n)_{n\in\mathbb{N}}$ in E has a subsequence $(p_{n_k})_{k\in\mathbb{N}}$ that converges to a limit in E.

Every finite set is compact because a sequence (p_n) contained in a finite set repeats a term infinitely often, and the corresponding constant subsequence converges.

Theorem 4.1. *Every compact set is closed and bounded.*

Proof. Suppose that E is a compact subset of the metric space X and that p is a limit of E. There is a sequence (p_n) in E converging to p. By compactness, some subsequence (p_{n_k}) converges to some $q \in E$. But every subsequence of a convergent sequence converges to the same limit and so $p = q$ and $p \in E$. Thus E is closed.

To see that E is bounded, choose and fix any point $p \in X$. Either E is bounded, or else, for each $n \in \mathbb{N}$ there is a point $p_n \in E$ such that $d(p, p_n) \geq n$. Compactness implies that some subsequence (p_{n_k}) converges. Convergent sequences are bounded by Proposition 2.3, which contradicts the fact that $d(p, p_{n_k}) \to \infty$ as $k \to \infty$. Therefore p_n cannot exist. □

Exercise 4.1. Show that the intersection of arbitrarily many compact sets and the union of finitely many compact sets is compact.

Theorem 4.2. *The closed interval $[a, b] \subset \mathbb{R}$ is compact.*

Proof. Let (x_n) be a sequence in $[a, b]$ and set

$$C \;=\; \{x \in [a, b] :\; x_n < x \text{ only finitely often}\}.$$

Equivalently, for all but finitely many n, $x_n \geq x$. Since $a \in C$ we know that $C \neq \emptyset$. Clearly b is an upper bound on C. By the least upper bound property of $\mathbb{R}$ there exists $c = \sup C$ with $c \in [a, b]$. We will show that a subsequence of (x_n) converges to c. Let $n_1 = 1$. For every $k \geq 2$, the interval $(c - \frac{1}{k}, c + \frac{1}{k})$ contains infinitely many elements of (x_n); for otherwise $c + \frac{1}{k} \in C$, which contradicts c being an upper bound for C. This means that we can always pick $n_k > n_{k-1}$ such that $|x_{n_k} - c| < \frac{1}{k}$. This subsequence converges to c by Lemma 2.3.

□

To pass from $\mathbb{R}$ to $\mathbb{R}^k$ we think about compactness for Cartesian products equipped in one of the induced metrics introduced in Section 2.3.

Theorem 4.3. *The Cartesian product of two compact sets is compact.*

Proof. Let $(a_n, b_n) \in A \times B$ be given where $A \subset X$ and $B \subset Y$ are compact. There exists a subsequence (a_{n_k}) that converges to $a \in A$ as $k \to \infty$. The subsequence (b_{n_k}) has a sub-subsequence $(b_{n_{k(l)}})$ that converges to some point $b \in B$. The sub-subsequence $(a_{n_k(l)})$ continues to converge to the point a. Thus, by Theorem 2.4,

$$(a_{n_{k(l)}}, b_{n_{k(l)}}) \longrightarrow (a, b)$$

as $l \to \infty$. This implies that $A \times B$ is compact. □

Corollary 4.1. *The Cartesian product of k compact sets is compact.*

Proof. Write $A_1 \times A_2 \times \cdots \times A_m = A_1 \times (A_2 \times \cdots \times A_m)$ and perform induction on m. □

Corollary 4.2. *Every k-cell $[a_1, b_1] \times \cdots \times [a_k, b_k]$ in $\mathbb{R}^k$ is compact.*

As a corollary we get the following important result.

Theorem 4.4 (Bolzano-Weierstrass). *Every bounded sequence in $\mathbb{R}^k$ has a convergent subsequence.*

Here is a simple fact about compact sets.

Theorem 4.5. *Every closed subset of a compact set is compact.*

Proof. If E is a closed subset of the compact set K and if (p_n) is a sequence of points in E then clearly (p_n) is also a sequence in K, so by compactness of K there is a subsequence p_{n_k} converging to a limit $p \in K$. Since E is closed, p lies in E, which proves that E is compact. □

Now we come to the first partial converse of Theorem 4.1, which works in the special case of $\mathbb{R}^k$.

Theorem 4.6 (Heine-Borel). *A set $K \subset \mathbb{R}^k$ is compact if and only if it is closed and bounded.*

Proof. The forward direction follows from Theorem 4.1. For the other direction: Boundedness implies that K is contained in some k-cell, which is compact by Corollary 4.2. Since K is closed, Theorem 4.5 implies that K is compact. □

The Heine-Borel Theorem states that closed and bounded subsets of Euclidean space are compact, but it is important to remember that a closed and bounded subset of a general metric space may fail to be compact. For example, the set $\mathbb{N}$ of natural numbers equipped with the discrete metric is closed in itself, and it is bounded. But it is not compact. For example, consider the sequence $1, 2, 3, \ldots$

If $E_1 \supset E_2 \supset \cdots \supset E_n \supset E_{n+1} \supset \ldots$ then (E_n) is a **nested sequence** of sets. Its intersection is

$$\bigcap_{n=1}^{\infty} E_n \;=\; \{p :\; p \in E_n \text{ for each } n\}.$$

For example, we could take E_n to be the disc $\{\mathbf{z} \in \mathbb{R}^2 : \|\mathbf{z}\| \leq 1/n\}$. The intersection of all sets E_n is then the singleton $\{\mathbf{0}\}$.

Theorem 4.7. *The intersection of a nested sequence of compact nonempty sets is compact and nonempty.*

Proof. Let (E_n) be such a sequence. By Theorem 4.1, each E_n is closed. The intersection of closed sets is closed. Thus $\bigcap E_n$ is a closed subset of the compact set E_1 and so it is also compact by Theorem 4.5. It remains to show that this intersection is nonempty.

Since E_n are nonempty, we can from each E_n choose a point p_n. The sequence (p_n) lies in E_1. Compactness of E_1 implies that (p_n) has a convergent subsequence (p_{n_k}) converging to $p \in E_1$. Now fix $n \geq 2$. The sub-subsequence of (p_{n_k}) with all $n_k \geq n$ lies in E_n and so, since E_n is closed, $p \in E_n$. Since n was arbitrary, $p \in \bigcap_{n \geq 1} E_n$ and so $\bigcap E_n$ is nonempty. □

In general, without compactness a similar result cannot hold. For example, the following two sets are empty: $\bigcap_{n=1}^{\infty}(0, \frac{1}{n})$, $\bigcap_{n=1}^{\infty}[n, \infty)$.

The **diameter** of a nonempty set $E \subset X$ is the supremum of the distances $d(x, y)$ between points in E

$$\operatorname{diam}(E) = \sup_{p,q \in E} d(p, q).$$

For example, both $(0, 1)$ and $[0, 1]$ have diameter 1.

Corollary 4.3. *If in addition to being nested, nonempty, and compact, the sets E_n have diameter that tends to 0 as $n \to \infty$ then $E = \bigcap E_n$ is a single point.*

Proof. For each $n \in \mathbb{N}$, E is a subset of E_n, which implies that E has diameter zero. Since distinct points lie at positive distance from each other, E consists of at most one point, while by Theorem 4.7 it consists of at least one point. □

4.2 Open coverings*

In our discussion of compact sets we followed Charles Pugh. The standard approach is more topological and may be occasionally useful in some aspects of analysis like measure theory; c.f. Chapter 10.

A collection $\mathcal{U}$ of subsets of X **covers** $E \subset X$ if E is contained in the union of sets in $\mathcal{U}$. We say that $\mathcal{U}$ is an **open covering** of E if $\mathcal{U}$ covers E and all sets in $\mathcal{U}$ are open. If $\mathcal{V}$ also covers E and each set in $\mathcal{V}$ also lies in $\mathcal{U}$ we say that $\mathcal{V}$ is a **subcovering** of E.

If every open covering of E has a finite subcovering, we say that E is **covering compact**. The point is here that such a finite subcovering must exist for *every* open covering of E. Just the fact that a finite open covering exist is obvious as a single (open) set X covers every set in X.

Theorem 4.8. *If X is a metric space and $E \subset X$ is covering compact then E is sequentially compact.*

Proof. Suppose E is covering compact but not sequentially compact. Then there exists a sequence (p_n) in E with no converging subsequence. It follows that each point $q \in E$ has a neighborhood U_q with only finitely many elements of (p_n) in it. The collection $\{U_q : q \in E\}$ is an open covering of E. By covering compactness, there is a finite subcovering $U_{q_1}, \ldots, U_{q_m}$ of E. Since each U_{q_i} has finitely many elements of (p_n) in it, it follows that (p_n) has finitely many elements, a contradiction. □

The opposite implication of Theorem 4.8 also holds. The proof relies on the concept of a Lebesgue number. A **Lebesgue number** of a covering $\mathcal{U}$ of E is a positive real number λ such that for each $p \in E$ there is some $U \in \mathcal{U}$ with $N_\lambda(p) \subset U$. Having one number that works for all $p \in E$ is a very strong requirement and, in general, if E is not compact cannot be achieved.

Lemma 4.1. *Every open covering of a sequentially compact set has a Lebesgue number $\lambda > 0$.*

Proof. Suppose not, that is, $\mathcal{U}$ is an open covering of a sequentially compact set E, and yet for every $\lambda > 0$ there is $p \in E$ such that no $U \in \mathcal{U}$ contains $N_\lambda(p)$. Take $\lambda = 1/n$ for $n \in \mathbb{N}$ and let $p_n \in E$ be a point such that no $U \in \mathcal{U}$ contains $N_{1/n}(p_n)$. By sequential compactness there is a subsequence (p_{n_k}) that converges to some $q \in E$. Since $\mathcal{U}$ is an open covering, there exists $U \in \mathcal{U}$ and $r > 0$ such that $N_r(q) \subset U$. If k is large then $d(p_{n_k}, q) < r/2$ and $1/n_k < r/2$, which implies by the triangle inequality that

$$N_{1/n_k}(p_{n_k}) \subset N_r(p) \subset U,$$

a contradiction. □

Theorem 4.9. *If X is a metric space and $E \subset X$ is sequentially compact then E is covering compact.*

Proof. Let $\mathcal{U}$ be an open covering of a sequentially compact E. By Lemma 4.1, $\mathcal{U}$ has a Lebesgue number $\lambda > 0$. Choose any $p_1 \in E$ and some $U_1 \in \mathcal{U}$ such that $N_\lambda(p_1) \subset U_1$. If $E \subset U_1$ then $\{U_1\}$ is a finite subcovering and the theorem is proved. Otherwise, there exists $p_2 \in E \setminus U_1$. Let $U_2 \in \mathcal{U}$ be such that $N_\lambda(p_2) \subset U_2$. Now either $E \subset U_1 \cup U_2$ in which case we are done, or otherwise, we continue producing a sequence (p_n) in E and a sequence (U_n) in $\mathcal{U}$ such that

$$N_\lambda(p_n) \subset U_n \text{ and } p_{n+1} \in E \setminus (U_1 \cup \cdots \cup U_n).$$

We will now show that such sequences must lead to contradiction. By sequential compactness there is a subsequence (p_{n_k}) that converges to some $p \in E$. For a large k, $d(p_{n_k}, p) < \lambda$ and so $p \in N_\lambda(p_{n_k}) \subset U_{n_k}$. All p_{n_l} with $l > k$ lie outside U_{n_k}, which contradicts their convergence to p. □

4.3 Continuity and compactness

Next we discuss how compact sets behave under continuous transformations.

Theorem 4.10. *If $f : X \to Y$ is continuous and K is a compact subset of X then $f(K)$ is compact. That is, the continuous image of a compact set is compact.*

Proof. Suppose that (q_n) is a sequence in $f(K)$. For each n choose a point $p_n \in K$ such that $f(p_n) = q_n$. By compactness of K there exists a subsequence (p_{n_k}) that converges to some point $p \in K$. By continuity of f it follows that

$$q_{n_k} \;=\; f(p_{n_k}) \;\to\; f(p) \in f(K)$$

as $k \to \infty$. Thus, every sequence (q_n) in $f(K)$ has a subsequence converging to a limit point in $f(K)$, which shows that $f(K)$ is compact. □

The following important result is an immediate corollary.

Theorem 4.11. *Let $f : X \to \mathbb{R}$ be a continuous function. If X is compact then there exist points $p, q \in X$ such that $f(p) = \inf_{x \in X} f(x)$ and $f(q) = \sup_{x \in X} f(x)$. In other words, f attains its optima.*

Proof. By Theorem 4.10 $f(X)$ is compact and so bounded and closed. By Theorem 2.8 it contains its supremum and infimum. □

A function $f : X \to Y$ is a **bijection** if it is one-to-one and onto, that is, (i) if $f(x) = f(x')$ then $x = x'$, and (ii) $\forall y \in Y\ \exists x \in X$ such that $f(x) = y$. If $f : X \to Y$ is a bijection then we define its inverse $f^{-1} : Y \to X$ as the function satisfying $f^{-1}(f(x)) = x$. For example if $f : \mathbb{R} \to \mathbb{R}$ is defined by $f(x) = 2x$ then $f^{-1}(x) = x/2$.

Definition 4.2. If $f : X \to Y$ is a bijection and f is continuous and the inverse function $f^{-1} : Y \to X$ is also continuous then f is a **homeomorphism**. We say that X and Y are **homeomorphic**.

Theorem 4.12. *If X is compact and $f : X \to Y$ is a continuous bijection then f is a homeomorphism.*

Proof. We need to show that $f^{-1} : Y \to X$ is continuous. Suppose that $y_n \to y$ in Y and let $x_n = f^{-1}(y_n)$, $x = f^{-1}(y)$. We must now show that $x_n \to x$. Suppose (x_n) does not converge to x. Then

$$\exists \epsilon > 0\ \ \forall N \in \mathbb{N}\ \ \exists n \geq N \quad d(x_n, x) \geq \epsilon.$$

In particular, there is a subsequence (x_{n_k}) contained in X that lies outside of the neighborhood $N_\epsilon(x)$. This and the fact that X is compact implies that (x_{n_k}) has a convergent subsequence $x_{n_{k_l}} \to x' \neq x$. By continuity of f, $f(x_{n_{k_l}}) \to f(x')$. However, $f(x_{n_{k_l}}) = y_{n_{k_l}} \to y = f(x)$ and so $f(x) = f(x')$ and $x = x'$ because f is a bijection. This gives a contradiction. □

Compactness of X is essential in Theorem 4.12. Consider the following example.

Example 4.1. Let $C = \{(x, y) \in \mathbb{R}^2 :\ x^2 + y^2 = 1\}$. The function $f(t) = (\cos(t), \sin(t))$ is a continuous bijection from $[0, 2\pi)$ to C. However, the inverse is not continuous. Consider a sequence (p_n) in C converging to $p = (1, 0)$ from below. Then $f^{-1}(p_n) \to 2\pi$ and $f^{-1}(p) = 0$.

4.4 Cauchy sequences and completeness

A sequence (p_n) in a metric space X is a **Cauchy sequence** if

$$\forall \epsilon > 0\ \ \exists N \in \mathbb{N} \text{ such that } \forall m, n \geq N \quad d(p_m, p_n) < \epsilon.$$

If p_n converges then it is Cauchy, which can be showed directly by the triangle inequality. The opposite may not be true. For example, the sequence r_n of rational numbers

$$1.4, 1.41, 1.414, 1.4142, 1.41421, 1.414213, \ldots$$

given by finer and finer decimal expansion of $\sqrt{2}$ is Cauchy. Given $\epsilon > 0$ choose $N > -\log_{10} \epsilon$. If $m, n \geq N$ then $|r_m - r_n| \leq 10^{-N} < \epsilon$. Nevertheless, (r_n) does not converge in $\mathbb{Q}$.

Often in real analysis we want to show that a sequence converges but without computing explicitly the limit. Showing that a sequence is Cauchy may be easier, and so it is important to develop conditions under which a Cauchy sequence converges.

Definition 4.3. A metric space is **complete** if every Cauchy sequence in X converges.

Start with the following exercise.

Exercise 4.2. Show that every Cauchy sequence is bounded.

Theorem 4.13. $\mathbb{R}^k$ *is complete.*

Proof. If (p_n) is Cauchy then it is bounded. By Theorem 4.4 (p_n) has a subsequence (p_{n_k}) that converges some point p. We will show that $p_n \to p$.

Since (p_n) is Cauchy there exists an N such that if $n, m \geq N$ then $d(p_n, p_m) < \epsilon/2$. Since $p_{n_k} \to p$ as $k \to \infty$, there exists $n^* = n_k$ for some k such that $d(p, p_{n^*}) < \epsilon/2$. There are infinitely many such n^* and so we can safely assume that $n^* \geq N$. Now for every $n \geq N$ we have

$$d(p, p_n) \overset{\triangle}{\leq} d(p, p_{n^*}) + d(p_{n^*}, p_n) \leq \tfrac{\epsilon}{2} + \tfrac{\epsilon}{2} = \epsilon,$$

which completes the verification that (p_n) converges.

□

Exercise 4.3. Let $a \in (0, 1)$ and consider the real sequence $s_n = 1 + a + \cdots + a^n = \sum_{k=0}^{n} a^k$. Show that s_n is Cauchy. Argue that $s_n \to \frac{1}{1-a}$.

Exercise 4.4. Let $C[0, 1]$ denote the space of all bounded real-valued functions on $[0, 1] \subset \mathbb{R}$. Prove that this space is a metric space with $d(f, g) = \sup_{x \in [0,1]} |f(x) - g(x)|$. Then prove that this space is complete.

Chapter 5
Basic linear algebra (2 lectures)

5.1 Vector space and its dimension

A subset $V \subset \mathbb{R}^k$ is a **vector space** if (i) $\mathbf{0}_k$ lies in V (ii) for any two $\mathbf{x}, \mathbf{y} \in V$ also $\mathbf{x} + \mathbf{y} \in V$, (iii) if $\mathbf{x} \in V$ and $\lambda \in \mathbb{R}$ then $\lambda \cdot \mathbf{x}$ lies in V. The real space $\mathbb{R}^k$ is a trivial example of a vector space. Another example is given by the plane $x_1 + x_2 + x_3 = 0$ in $\mathbb{R}^3$. However, the *affine* plane $x_1 + x_2 + x_3 = 1$ is not a vector space because $\mathbf{0}_3$ does not satisfy the defining equation and so property (i) fails to hold.

This extends extends to other familiar situations like vector spaces in the space $\mathbb{R}^{m\times n}$ of $m \times n$ matrices. The zero element is simply the matrix of zeros and addition and scalar multiplication are defined in the usual way.

The vector spaces structure can be defined on more abstract spaces as long as addition and scalar multiplication are properly defined together with the zero element $\mathbf{0}$. An example of an abstract vector space is the set of all bounded real-valued functions over $[0, 1]$. Here $f + g$ is as in defined by $(f + g)(x) = f(x) + g(x)$; $\lambda \cdot f$ is defined by $(\lambda \cdot f)(x) = \lambda f(x)$ and $\mathbf{0}$ is the function, which is constant equal to zero.

Given vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^k$ and scalars $\lambda_1, \ldots, \lambda_n \in \mathbb{R}$, the vector

$$\lambda_1\mathbf{x}_1 + \lambda_2\mathbf{x}_2 + \cdots + \lambda_n\mathbf{x}_n \quad \in \quad \mathbb{R}^k$$

is called a **linear combination** of $\mathbf{x}_1, \ldots, \mathbf{x}_n$ and the scalars $\lambda_1, \lambda_2, \ldots, \lambda_n$ are called the **coefficients** of this linear combination.

Definition 5.1. Let $E \subset \mathbb{R}^k$ be a nonempty subset, then the **span** of E, denoted by $\text{span}(E)$, is the set of all (finite) linear combinations of elements of E. More precisely, $\mathbf{x} \in \text{span}(E)$ if there is a natural number n, points $\mathbf{x}_1, \ldots, \mathbf{x}_n \in E$, and coefficients $\lambda_1, \ldots, \lambda_n$ such that

$$\mathbf{x} \quad = \quad \lambda_1\mathbf{x}_1 + \lambda_2\mathbf{x}_2 + \cdots + \lambda_n\mathbf{x}_n.$$

Exercise 5.1. Show that the span of E is always a vector space.

Definition 5.2. A set of vectors $E = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ in vector space V is **linearly independent** if

$$\lambda_1\mathbf{x}_1 + \lambda_2\mathbf{x}_2 + \cdots + \lambda_n\mathbf{x}_n = \mathbf{0}_k \qquad \Longrightarrow \qquad \lambda_1 = \lambda_2 = \cdots = \lambda_n = 0.$$

Note that $E = \{\mathbf{x}_1\}$ is linearly independent if and only if $\mathbf{x}_1 \neq \mathbf{0}_k$. For two elements, $E = \{\mathbf{x}_1, \mathbf{x}_2\}$ is linearly independent if and only if there is no $\lambda \in \mathbb{R}$ such that $\mathbf{x}_2 = \lambda\mathbf{x}_1$.

Example 5.1. To check if $(1,1,0)$, $(1,0,1)$, $(0,1,1)$ are linearly independent we check that

$$\lambda_1(1,1,0) + \lambda_2(1,0,1) + \lambda_3(0,1,1) \;=\; (\lambda_1 + \lambda_2, \lambda_1 + \lambda_3, \lambda_2 + \lambda_3),$$

which is equal to the zero vector only if

$$\lambda_1 + \lambda_2 \;=\; \lambda_1 + \lambda_3 \;=\; \lambda_2 + \lambda_3 \;=\; 0$$

and we easily check that this is possible only if $\lambda_1 = \lambda_2 = \lambda_3 = 0$. Therefore these three vectors are linearly independent.

We leave as an exercise to prove the following two facts.

Lemma 5.1. *A set of vectors $E = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ in some vector space $V \subset \mathbb{R}^k$ is linearly independent if and only if no element in E can be written as a linear combination of the others.*

Lemma 5.2. *Any nonempty subset of a linearly independent set is linearly independent.*

Definition 5.3. If a vector space V contains d linearly independent vectors but no $d+1$ linearly independent vectors, then the **dimension** of V is d; we write $\dim(V) = d$.

The fact that $\dim(V)$ is well-defined follows from Lemma 5.2. Indeed, suppose that there exists two different numbers d, d' satisfying Definition 5.3 and suppose $d' > d$. Let E be a subset of V with d' elements that is linearly independent. By Lemma 5.2 each subset of E is linearly independent. It follows that V has a subset of size $d+1$, which is linearly independent contradicting the fact that d satisfies Definition 5.3.

Definition 5.4. A set E **spans** V if $V = \operatorname{span}(E)$. A **basis** of V is any linearly independent set that spans V.

The fact that every vector space with finite dimension has a basis will be proved later in Theorem 5.3. We first discuss importance of this concept.

Theorem 5.1. *The set $E = \{\mathbf{x}_1, \ldots, \mathbf{x}_d\}$ is a basis of V if and only if every $\mathbf{x} \in V$ can be written* uniquely *as a linear combination of elements of E.*

Proof. Every $\mathbf{x} \in V$ can be written as a linear combination of elements of E if and only if E spans V. This can be done uniquely if and only if E is linearly independent. Indeed, if there is an element that has two representations with coefficients $\lambda_1, \ldots, \lambda_d$ and $\lambda'_1, \ldots, \lambda'_d$ (not all equal) then

$$(\lambda_1 - \lambda'_1)\mathbf{x}_1 + \cdots + (\lambda_d - \lambda'_d)\mathbf{x}_d \;=\; \mathbf{0}_k$$

which is possible if and only if E is linearly dependent. □

Theorem 5.1 says that that a fixed basis of V induces a bijection $f : \mathbb{R}^d \to V$, $(\lambda_1, \ldots, \lambda_d) \mapsto \lambda_1\mathbf{x}_1 + \cdots + \lambda_d\mathbf{x}_d$; we call f a parameterization of V.

Example 5.2. Let $\mathbf{e}_i \in \mathbb{R}^k$ be the vector with i-th coordinate equal to 1 and all other coordinates equal to zero. Then $\mathbf{e}_1, \ldots, \mathbf{e}_k$ forms a basis of $\mathbb{R}^k$ called the standard basis of $\mathbb{R}^k$. If $\mathbf{x} = (x_1, \ldots, x_k)$ then $\mathbf{x} = x_1\mathbf{e}_1 + \cdots + x_k\mathbf{e}_k$.

Theorem 5.2. *Let V be spanned by a set $E = \{\mathbf{x}_1, \ldots, \mathbf{x}_r\}$, then* $\dim(V) \le r$.

Proof. Suppose E spans V and $\dim(V) > r$. If $\dim(V) > r$, then V contains a linearly independent set $F = \{\mathbf{y}_1, \ldots, \mathbf{y}_{r+1}\}$. Since E spans V, $\mathbf{y}_1 = \sum_{k=1}^{r} b_k\mathbf{x}_k$ and at least one $b_i \neq 0$, and so $\mathbf{x}_i$ is a linear combination of $\mathbf{y}_1, (\mathbf{x}_k)_{k\neq i}$. Thus, this new set also spans V, so $\mathbf{y}_2 = a_1\mathbf{y}_1 + \sum_{k\neq i} b_j\mathbf{x}_k$ for some $a_1, (b_k)_{k\neq i}$. If all b_k were zero then we would have $\mathbf{y}_2 = a_1\mathbf{y}_2$, which is impossible (linear independence). Thus at least one x_k from the remaining ones, say $\mathbf{x}_j$, is a linear combination of $\mathbf{y}_1, \mathbf{y}_2$ and all $\mathbf{x}_k$ for $k \neq i, j$, and so these vectors span V. Continuing this procedure we see that $\mathbf{y}_1, \ldots, \mathbf{y}_r$ span V and so $\mathbf{y}_{r+1}$ can be written as their linear combination, which leads to contradiction. □

Corollary 5.1. *The dimension of $\mathbb{R}^k$ is k.*

Proof. Since $\mathbf{e}_1, \ldots, \mathbf{e}_k$ spans $\mathbb{R}^k$, Theorem 5.2 implies that $\dim(\mathbb{R}^k) \le k$. But $\mathbf{e}_1, \ldots, \mathbf{e}_k$ are linearly independent and so $\dim(\mathbb{R}^k) \ge k$. □

Theorem 5.3. *Suppose that* $\dim(V) = k$, *then*

(a) $\{\mathbf{x}_1, \ldots, \mathbf{x}_k\} \subset V$ spans V if and only if it is linearly independent.
(b) V has a basis and every basis has k vectors.
(c) If $r \le k$ and $\{\mathbf{y}_1, \ldots, \mathbf{y}_r\}$ is an independent set then V has a basis containing $\mathbf{y}_1, \ldots, \mathbf{y}_r$.

Proof. (a) We first show the left implication. If $\{\mathbf{x}_1, \ldots, \mathbf{x}_k\}$ is independent then for every other vector $\mathbf{y} \in V$ the set $\{\mathbf{y}, \mathbf{x}_1, \ldots, \mathbf{x}_k\}$ is linearly dependent, there exist coefficients $\lambda_0, \lambda_1, \ldots, \lambda_k$, not all zero, such that

$$\lambda_0\mathbf{y} + \lambda_1\mathbf{x}_1 + \cdots + \lambda_k\mathbf{x}_k \;=\; \mathbf{0}.$$

We cannot have $\lambda_0 = 0$ because $\mathbf{x}_i$ are linearly independent. Dividing by λ_0 we show that $\mathbf{y}$ is a linear combination of $\mathbf{x}_1, \ldots, \mathbf{x}_k$. This proves that

$V \subset \text{span}\{\mathbf{x}_1, \ldots, \mathbf{x}_k\}$. To show the right implication we prove the contrapositive statement. If $\{\mathbf{x}_1, \ldots, \mathbf{x}_k\}$ is dependent then one vector can be removed without changing the span, so, by Theorem 5.2 the dimension of the span of $\{\mathbf{x}_1, \ldots, \mathbf{x}_k\}$ is less or equal to $k-1$, so $\{\mathbf{x}_1, \ldots, \mathbf{x}_k\}$ it cannot span V.

(b) Since $\dim(V) = k$, there exist k independent vectors. By (a) they span V so they form a basis. For any basis, by Theorem 5.2, the number of vectors is greater or equal to k (spans so $\geq k$ elements). By the definition of dimension, there is no independent set of size strictly greater than k so the number of vectors in a basis is less or equal to k (independent so $\leq k$ elements). This implies the equality.

(c) Let $\{\mathbf{x}_1, \ldots, \mathbf{x}_k\}$ be a basis. Then $\{\mathbf{x}_1, \ldots, \mathbf{x}_k, \mathbf{y}_1, \ldots, \mathbf{y}_r\}$ spans V. The same way as in the proof of Theorem 5.2 one of the $\mathbf{x}_i$'s is a linear combination of the rest and can be dropped. This procedure can be repeated r times. □

Exercise 5.2. Show that $V = \{(x, y, z) \in \mathbb{R}^3 : \ x + y + z = 0\}$ forms a vector space. Find a basis of V. Express the vector $(23, -10, -13)$ in this basis.

If $V \subset \mathbb{R}^k$ then we define the scalar product on V exactly in the same way as in Section 1.3. In some situations it will be useful to consider a generalisation of this concept.

Definition 5.5. An **inner product** of a vector space is a map $\langle \cdot, \cdot \rangle : V \times V \to \mathbb{R}$ that satisfies the following three conditions for all $\mathbf{x}, \mathbf{y}, \mathbf{z} \in V$

1. $\langle \mathbf{x}, \mathbf{y} \rangle = \langle \mathbf{y}, \mathbf{x} \rangle$ (symmetry)
2. $\langle \alpha\mathbf{x} + \beta\mathbf{y}, \mathbf{z} \rangle = \alpha\langle \mathbf{x}, \mathbf{z} \rangle + \beta\langle \mathbf{y}, \mathbf{z} \rangle$ (linearity in the first argument)
3. $\langle \mathbf{x}, \mathbf{x} \rangle \geq 0$ and is zero only if $\mathbf{x} = \mathbf{0}$. (positive definiteness)

Exercise 5.3. Formulate and proof a version of the Cauchy-Schwarz inequality (Theorem 1.7) that holds for any inner product.

Exercise 5.4. Suppose that S is an $k \times k$ positive definite matrix. Show that $\langle \mathbf{x}, \mathbf{y} \rangle := \mathbf{x}^T S \mathbf{y}$ defines an inner product on $\mathbb{R}^k$.

Exercise 5.5. In the space $\mathbb{R}^{m \times n}$ of $m \times n$ matrices define $\langle A, B \rangle := \text{trace}(AB^T)$. Show that it forms an inner product.

5.2 Linear transformations and matrices

A map $T : \mathbb{R}^n \to \mathbb{R}^m$ is called a **linear transformation** if

(i) $T(\mathbf{0}_n) = \mathbf{0}_m$,

(ii) $T(\mathbf{x} + \mathbf{y}) = T(\mathbf{x}) + T(\mathbf{y})$ for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, and

(iii) $T(\lambda\mathbf{x}) = \lambda T(\mathbf{x})$ for all $\mathbf{x} \in \mathbb{R}^n$ and $\lambda \in \mathbb{R}$.

The set $\mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$ of linear transformations $T : \mathbb{R}^n \to \mathbb{R}^m$ forms a vector space. As for any functions, $U = T + S$ is defined by $U(\mathbf{x}) = T(\mathbf{x}) + S(\mathbf{x})$, and λT being defined by $(\lambda T)(\mathbf{x}) = \lambda T(\mathbf{x})$. It is an elementary exercise to check that (i) $T + S$ is a linear transformation if T and S are, and (ii) λT is a linear transformation for every $\lambda \in \mathbb{R}$ if T is.

The vector space $\mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$ is abstract in the sense that it is not immediately clear what is its dimension, or the basis. On the other hand, there is another very concrete vector space, which we will show, is essentially equal to $\mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$. This is the space $\mathbb{R}^{m\times n}$ of all $m \times n$ matrices. Two matrices are added by adding the corresponding entries, $A + B = C$ where $c_{ij} = a_{ij} + b_{ij}$. Similarly, if $\lambda \in \mathbb{R}$ is a scalar, λA is a matrix with the entries λa_{ij}. The vector space $\mathbb{R}^{m\times n}$ has dimension mn, and its canonical basis is given by the elementary matrices E_{ij}. Moreover, this space admits a natural inner product

$$\langle A, B\rangle \;=\; \operatorname{trace}(AB^T) \;=\; \sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij}b_{ij},$$

which gives the metric space structure.

The theory of differentiation is build around the concept of the linear transformation. The aim of this and the next section is to show that, on many different levels, linear operations and matrices are really the same objects. This important observation gives us many powerful techniques and insight that come from the matrix algebra.

As a rule of thumb, think with linear transformations and compute with matrices.

Definition 5.6. Two vector spaces V and W are **isomorphic** if there is a bijection $f : V \to W$ which is linear. Such a map is called a **linear isomorphism**.

Exercise 5.6. Show that if $f : V \to W$ is an isomorphism then the inverse map f^{-1} is linear.

Having an isomorphism between two vector spaces is particularly useful if one of the spaces is more concrete and natural to work with. Our example of linear transformations is one instance of that phenomenon but other examples come into mind.

Exercise 5.7. Let V be the set of all quadratic polynomials, that is, expressions of the form $ax^2 + bx + c$ for $a, b, c \in \mathbb{R}$. Show that this set forms a vector space. Further, show it is isomorphic to $\mathbb{R}^3$.

An isomorphism preserves the vector structure. This can be used in a number of ways if one of the spaces is easier to work with. Suppose that in V we have a natural candidate for a vector basis and a natural notion of the inner product (and so also of a distance). This can then be translated to W in a direct way as explained in the exercises below.

Exercise 5.8. Suppose that $f : V \to W$ is an isomorphism. Show that $\{\mathbf{x}_1, \dots, \mathbf{x}_k\}$ is a basis of V if and only if $\{f(\mathbf{x}_1), \dots, f(\mathbf{x}_k)\}$ is a basis of W. Moreover, every k-dimensional vector space is isomorphic to $\mathbb{R}^k$.

Exercise 5.9. Suppose that $f : V \to W$ is an isomorphism. Given an inner product $\langle \cdot, \cdot \rangle$ on V define an inner product on W through $\langle \mathbf{x}, \mathbf{y} \rangle := \langle f^{-1}(\mathbf{x}), f^{-1}(\mathbf{y}) \rangle$. Show that it satisfies the conditions of Definition 5.5.

The space $\mathbb{R}^{m \times n}$ and $\mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$ are linearly isomorphic. The bijection f takes an $m \times n$ matrix $A = [a_{ij}]$ and associates to it a linear transformation $T_A : \mathbb{R}^n \to \mathbb{R}^m$ given by

$$T_A(\mathbf{x}) \;=\; A\mathbf{x}, \tag{5.1}$$

where $\mathbf{x} = (x_1, \dots, x_n)$. The inverse of f is a linear map that takes a linear transformation $T \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$ and associates to it a matrix $[T] \in \mathbb{R}^{m \times n}$ whose columns are vectors $\mathbf{a}_1, \dots, \mathbf{a}_n \in \mathbb{R}^m$ given by $\mathbf{a}_i = T(\mathbf{e}_i)$, where $\mathbf{e}_1, \dots, \mathbf{e}_n$ is the *standard basis* of $\mathbb{R}^n$. To shows that f is a linear isomorphism you may find the following basic observation useful.

Remark 5.1 (Fundamental observation of matrix algebra). If $A \in \mathbb{R}^{m \times n}$ and $\mathbf{x} \in \mathbb{R}^n$ then $A\mathbf{x} \in \mathbb{R}^m$ is a linear combination of the columns $\mathbf{a}_1, \dots, \mathbf{a}_n$ of A with coefficients given by the entries of $\mathbf{x} = (x_1, \dots, x_n)$. In other words, $A\mathbf{x} = x_1\mathbf{a}_1 + \cdots + x_n\mathbf{a}_n$.

We could not stress more the importance of Remark 5.1. We will see many of its applications.

Exercise 5.10. Show that every $T \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$ is continuous. Conclude that the map $f(\mathbf{x}) = \|A\mathbf{x}\|$ is continuous.

Exercise 5.11. Show that $f : \mathbb{R}^{m \times n} \to \mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$ is a linear isomorphism.

The link between $\mathbb{R}^{m \times n}$ and $\mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$ goes further.

Proposition 5.1. *The composition $T \circ S$ of two linear transformations $S : \mathbb{R}^n \to \mathbb{R}^m$ and $T : \mathbb{R}^m \to \mathbb{R}^l$ is represented by the matrix obtained by matrix multiplication of $[S]$ and $[T]$, that is*

$$[T \circ S] \;=\; [T] \cdot [S].$$

Proof. We will show that the j-th column of $[T \circ S]$ is equal to the j-th column of $[T] \cdot [S]$. Denote by $\mathbf{s}_j = S(\mathbf{e}_j)$ the j-th column of $[S]$ and by $\mathbf{t}_i = T(\mathbf{e}_i)$ the i-th column of $[T]$. We have

$$\begin{aligned}(T \circ S)(\mathbf{e}_j) = T(S(\mathbf{e}_j)) = T(\mathbf{s}_j) = T(\sum_{i=1}^{n} s_{ij}\mathbf{e}_i) = \\ = \sum_{i=1}^{n} s_{ij} T(\mathbf{e}_i) = \sum_{i=1}^{n} s_{ij}\mathbf{t}_i = [T]\mathbf{s}_j = [T] \cdot [S]\mathbf{e}_j.\end{aligned}$$

□

Definition 5.7. For a matrix $\mathbb{R}^{m\times n}$ its kernel is

$$\ker(A) = \{\mathbf{x} : \ A\mathbf{x} = \mathbf{0}_m\} \subset \mathbb{R}^n.$$

Similarly we define the image $\mathrm{Im}(A)$ of A as

$$\mathrm{Im}(A) \ = \ \{A\mathbf{x} : \ \mathbf{x} \in \mathbb{R}^n\} \subset \mathbb{R}^m.$$

It is easy to see that the kernel and the image of A are vector spaces. By Remark 5.1, $\mathrm{Im}(A)$ is spanned by the columns of A.

Definition 5.8. A matrix $A \in \mathbb{R}^{n\times n}$ is *invertible* if there exists a matrix $B \in \mathbb{R}^{n\times n}$ such that $AB = \mathbb{I}_n$. Then B is denoted by A^{-1}.

The following result shows that the inverse is well defined.

Proposition 5.2. *The matrix B in the definition of the inverse, if exists, is unique.*

Proof. Let $\mathbf{b}_1, \ldots, \mathbf{b}_n$ be the columns of B. The matrix equation $AB = \mathbb{I}_n$ is equivalent to $A\mathbf{b}_1 = \mathbf{e}_1, \ldots, A\mathbf{b}_n = \mathbf{e}_n$. Using Remark 5.1 we see that each canonical unit vector is expressed as a linear combination of the columns of A with coefficients give by entries of the matrix B. This implies that the columns of A span $\mathbb{R}^n$ and so they form a basis of $\mathbb{R}^n$ by Theorem 5.3(a). But then the vectors $\mathbf{b}_1, \ldots, \mathbf{b}_n$ are uniquely defined by Theorem 5.1. □

Exercise 5.12. $A \in \mathbb{R}^{n\times n}$ is invertible if and only if

$$\ker(A) = \{\mathbf{0}_n\}. \tag{5.2}$$

Exercise 5.13. Show that if $A \in \mathbb{R}^{n\times n}$ is invertible if and only if its columns are linearly independent. Show that then also $A^{-1}A = \mathbb{I}_n$ (Hint: Show that $AB = \mathbb{I}_n$ implies that $\ker(A) = \{\mathbf{0}_n\}$).

Proposition 5.3. *A linear transformation $T \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^n)$ is invertible if and only if $[T]$ is an invertible matrix. The matrix of T^{-1} is the inverse of $[T]$.*

We leave the proof as an exercise.

In statistics and econometrics we often work in the space of matrices. The following exercises show that these vector spaces are isomorphic to $\mathbb{R}^k$ for some k.

Exercise 5.14. Show that the set of $m \times n$ matrices is isomorphic to $\mathbb{R}^{mn}$.

Exercise 5.15. Show that the set of symmetric $n\times n$ matrices forms a vector space and that is isomorphic to $\mathbb{R}^k$ for $k = \binom{n+1}{2}$.

The main motivation to work with linear transformations rather than the representing matrices is that the notation is simpler. In more abstract situations, including the ones covered in the above exercises, working with linear transformations gives as freedom of not choosing the underlying canonical basis. For example, fixing $A \in \mathbb{R}^{n\times n}$, the map $T(X) = AX + XA$ is a simple linear map from $\mathbb{R}^{n\times n}$ to $\mathbb{R}^{n\times n}$. We could represent it as an element of $\mathcal{L}(\mathbb{R}^{n^2}, \mathbb{R}^{n^2})$ but writing the associated $n^2 \times n^2$ matrix complicates the situation.

5.3 Orthogonal complements and projections

We say that two vectors $\mathbf{x}, \mathbf{y}$ in $\mathbb{R}^k$ are **orthogonal** if the scaler product $\langle \mathbf{x}, \mathbf{y}\rangle$ is equal to zero. A set $\{\mathbf{x}_1, \ldots, \mathbf{x}_r\}$ in $\mathbb{R}^k$ is orthogonal if all its elements are mutually orthogonal.

Definition 5.9. If $V \subset \mathbb{R}^k$ is a vector space, then the **orthogonal complement** of V is

$$V^\perp \;:=\; \{\mathbf{x} \in \mathbb{R}^k :\; \langle \mathbf{x}, \mathbf{v}\rangle = 0 \text{ for all } \mathbf{v} \in V\}.$$

It is easy to see that $V^\perp$ forms a vector space. It can be more efficiently described if a basis for V is given.

Exercise 5.16. Suppose that $\mathbf{v}_1, \ldots, \mathbf{v}_r$ spans V. Show that

$$V^\perp \;:=\; \{\mathbf{x} \in \mathbb{R}^k :\; \langle \mathbf{x}, \mathbf{v}_i\rangle = 0 \text{ for all } i = 1, \ldots, r\}.$$

Let $\mathbf{V}$ be the matrix whose columns are $\mathbf{v}_1, \ldots, \mathbf{v}_r$. Conclude that $V^\perp = \{\mathbf{x} \in \mathbb{R}^k : \mathbf{V}^T\mathbf{x} = \mathbf{0}\}$.

Exercise 5.17. Suppose that $V = \{\mathbf{x} \in \mathbb{R}^3 : x_1 + x_2 + x_3 = 0\}$. What is $V^\perp$?

Example 5.3. Recall the definition of the image and the kernel of a matrix in Definition 5.7. The condition $A\mathbf{x} = \mathbf{0}_m$ means that $\mathbf{x}$ is orthogonal to the space spanned by the rows of A, or equivalently, by the columns of A^T, which gives the following fundamental equivalence

$$\ker(A) \;=\; \mathrm{Im}(A^T)^\perp. \tag{5.3}$$

Suppose that given a vector $\mathbf{y} \in \mathbb{R}^2$ we want to find the closest point to $\mathbf{y}$ on a line $L = \mathrm{span}\{\mathbf{x}\}$. It turns out that this optimal point $\hat{\mathbf{y}}$ is the orthogonal projection of $\mathbf{y}$ onto L, that is, the unique vector in L such that $\mathbf{y} - \hat{\mathbf{y}} \in L^\perp$; see Figure 5.1. We can find $\hat{\mathbf{y}}$ using elementary techniques. Since it lies in L it is of the form $\hat{\mathbf{y}} = \lambda\mathbf{x}$ for some $\lambda \in \mathbb{R}$ and the condition on $\mathbf{y} - \hat{\mathbf{y}}$ gives

$$\langle \mathbf{y} - \hat{\mathbf{y}}, \mathbf{x}\rangle \;=\; \langle \mathbf{y}, \mathbf{x}\rangle - \lambda\langle \mathbf{x}, \mathbf{x}\rangle \;=\; 0,$$

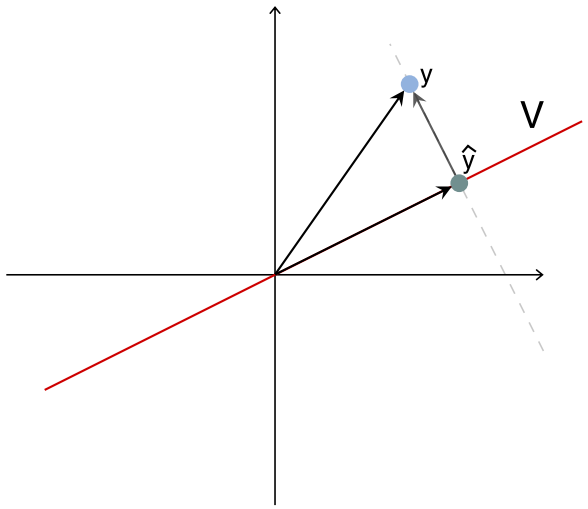

Fig. 5.1 Orthogonal projection in $\mathbb{R}^2$ on a one-dimensional vector space V.

which implies that

$$\hat{\mathbf{y}} = \frac{\langle \mathbf{y}, \mathbf{x} \rangle}{\langle \mathbf{x}, \mathbf{x} \rangle} \mathbf{x}.$$

Exercise 5.18. Show that $\hat{\mathbf{y}}$ is the unique minimizer of $\|\mathbf{y} - \mathbf{u}\|$ over all $\mathbf{u} \in L$ by minimizing the function $f(\lambda) = \|\mathbf{y} - \lambda \mathbf{x}\|^2$.

Theorem 5.4 (The orthogonal decomposition theorem). *Let V be a subspace of $\mathbb{R}^k$. Then each $\mathbf{y} \in \mathbb{R}^k$ can be written uniquely in the form*

$$\mathbf{y} = \hat{\mathbf{y}} + \mathbf{z},$$

where $\hat{\mathbf{y}} \in V$ and $\mathbf{z} \in V^\perp$.

The vector $\hat{\mathbf{y}}$ is called the orthogonal projection of $\mathbf{y}$ onto V and it is often denoted by $\text{proj}_V(\mathbf{y})$.

Proof. Fix *any* basis $\mathbf{v}_1, \ldots, \mathbf{v}_r$ of V and let $\mathbf{V} \in \mathbb{R}^{k \times r}$ be the matrix with these vectors as columns. The condition $\hat{\mathbf{y}} \in V$ translates to $\hat{\mathbf{y}} = \mathbf{V}\lambda$ for some $\lambda \in \mathbb{R}^r$; c.f. Remark 5.1. The condition $\mathbf{y} - \hat{\mathbf{y}} \in V^\perp$ translates to $\mathbf{V}^T(\mathbf{y} - \hat{\mathbf{y}}) = \mathbf{0}$. These two conditions together uniquely identify $\lambda = (\mathbf{V}^T\mathbf{V})^{-1}\mathbf{V}^T\mathbf{y}$. (rings a bell?) The fact that $\mathbf{V}^T\mathbf{V}$ is invertible can be argued as follows: By Exercise 5.12 it is enough to show that $\ker(\mathbf{V}^T\mathbf{V}) = \{\mathbf{0}\}$. If $\mathbf{V}^T\mathbf{V}\mathbf{x} = \mathbf{0}$ then also $\mathbf{x}^T\mathbf{V}^T\mathbf{V}\mathbf{x} = \mathbf{0}$ or equivalently $\|\mathbf{V}\mathbf{x}\| = 0$. This is possible only if $\mathbf{V}\mathbf{x} = \mathbf{0}$. Since the columns of $\mathbf{V}$ are linearly independent we conclude that $\mathbf{x} = \mathbf{0}$. □

Remark 5.2. The uniqueness of the orthogonal decomposition in Theorem 5.4 shows that this decomposition depends only on V and not on a particular basis chosen.

The following will be used in the proof of Theorem 9.4.

Lemma 5.3. *For two vector spaces $U, V \subset \mathbb{R}^k$ we have $U \subset V$ if and only if $V^\perp \subset U^\perp$.*

Proof. For the forward direction, let $\mathbf{x} \in V^\perp$, then $\langle \mathbf{x}, \mathbf{v} \rangle = 0$ for all $\mathbf{v} \in V$. Since $U \subset V$, also $\langle \mathbf{x}, \mathbf{u} \rangle = 0$ for all $\mathbf{u} \in U$ and so $\mathbf{x} \in U^\perp$. To prove the opposite direction, first note that the forward direction implies that

$$V^\perp \subset U^\perp \qquad \Longrightarrow \qquad (U^\perp)^\perp \subset (V^\perp)^\perp.$$

Hence, to finish the proof, it is enough to show that for every $W \subset \mathbb{R}^k$ we have $(W^\perp)^\perp = W$. This equality will follow by two claims (i) $W \subset (W^\perp)^\perp$, and (ii) $\dim W = \dim(W^\perp)^\perp$. The first claim follows directly by the definition. For the second statement, let $\mathbf{x}_1, \ldots, \mathbf{x}_r$ be a basis of W, $\dim W = r$. By Theorem 5.4 each $\mathbf{y} \in \mathbb{R}^k$ can be uniquely written as $\hat{\mathbf{y}} + \mathbf{z}$ where $\hat{\mathbf{y}} \in W$ and $\mathbf{z} \in W^\perp$. Moreover, $\hat{\mathbf{y}}$ can be uniquely written as a linear combination of $\mathbf{x}_1, \ldots, \mathbf{x}_r$ and $\mathbf{z}$ can be uniquely written as a linear combination of the basis of $W^\perp$. By Theorem 5.1 the basis of W complemented with the basis of $W^\perp$ forms the basis of $\mathbb{R}^k$. In particular, $\dim W^\perp = k - r$. Using a similar argument for $W^\perp$ and $(W^\perp)^\perp$ we conclude that $\dim(W^\perp)^\perp = k - (k-r) = r$. □

5.4 Matrix norms

There are several norms that we can define on $\mathbb{R}^{m \times n}$ to give it a metric space structure. A trivial choice is the one coming from the identification with $\mathbb{R}^{mn}$ (c.f. Exercise 5.14), where each $m \times n$ matrix can be transformed to a vector in $\mathbb{R}^{m \times n}$ by the operation called vectorization, where $\operatorname{vec}(A)$ is a long vector obtained from the columns of A stacked one below another. This gives the Frobenius norm

$$\|A\|_F \;:=\; \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij}^2}.$$

This norm is simply the Euclidean norm of $\operatorname{vec}(A)$ in $\mathbb{R}^{mn}$. The canonical choice in this course is the norm induced from vector norms on $\mathbb{R}^m$ and $\mathbb{R}^n$

$$\|A\| \;:=\; \sup_{\mathbf{x} \neq \mathbf{0}_n} \tfrac{\|A\mathbf{x}\|}{\|\mathbf{x}\|} \;=\; \sup_{\|\mathbf{x}\|=1} \|A\mathbf{x}\|.$$

This norm is called the *operator norm*. The fact that $\|A\|$ is indeed a norm, will follow from the next theorem.

Exercise 5.19. Use Exercise 5.10 to conclude that $\sup_{\|\mathbf{x}\|=1}$ in the definition can be replaced with $\max_{\|\mathbf{x}\|=1}$.

As we will see, there are many interesting properties that link the operator norm with the vector norm inducing it. Observe, for example, that for any $\mathbf{x} \in \mathbb{R}^n$ we have

$$\|A\mathbf{x}\| \;\leq\; \|A\|\,\|\mathbf{x}\|. \tag{5.4}$$

Indeed, this clearly holds if $\mathbf{x} = \mathbf{0}_n$. If $\mathbf{x} \neq \mathbf{0}_n$ we use the definition of $\|A\|$ to conclude that

$$\frac{\|A\mathbf{x}\|}{\|\mathbf{x}\|} \;\leq\; \|A\|,$$

which immediately implies (5.4). This important inequality will be repeatedly used in the proof of the following theorem and throughout.

Theorem 5.5. *The following statements hold:*

(a) If $A \in \mathbb{R}^{m\times n}$ then $\|A\| < \infty$.
(b) $\|A + B\| \leq \|A\| + \|B\|$ and $\|cA\| = |c|\,\|A\|$ for all $c \in \mathbb{R}$.
(c) If $A \in \mathbb{R}^{m\times n}$ and $B \in \mathbb{R}^{k\times m}$ then $\|B \cdot A\| \leq \|B\| \cdot \|A\|$.

Proof of Theorem 5.5. (a) Follows from Exercise 5.19.

(b) We have $\|(A+B)\mathbf{x}\| = \|A\mathbf{x}+B\mathbf{x}\| \overset{\triangle}{\leq} \|A\mathbf{x}\|+\|B\mathbf{x}\| \leq (\|A\|+\|B\|)\|\mathbf{x}\|$ and so $\|A + B\| \leq \|A\| + \|B\|$ by dividing by $\|\mathbf{x}\|$ on both sides. The second part follows directly from the definition.

(c) This follows from $\|BA\mathbf{x}\| \leq \|B\| \cdot \|A\mathbf{x}\| \leq \|B\| \cdot \|A\| \cdot \|\mathbf{x}\|$ and by dividing by $\|\mathbf{x}\|$ on both sides. □

At this point $\|\cdot\|$ denotes two different norms. If $\mathbf{x} \in \mathbb{R}^k$ then $\|\mathbf{x}\|$ denotes the Euclidean norm in $\mathbb{R}^k$. If $A \in \mathbb{R}^{m\times n}$, then $\|A\|$ denotes the operator norm of the matrix A, equivalently we have norms on the set of linear transformations. However, if A consists of a single row or a single column we also think about it as a vector. The following result shows that this leads to no ambiguity.

Proposition 5.4. *Let $\mathbf{y} \in \mathbb{R}^k$. Then the Euclidean norm of $\mathbf{y}$ is equal to its operator norm when taken as a matrix in $\mathbb{R}^{k\times 1}$ or in $\mathbb{R}^{1\times k}$.*

Proof. Let $\mathbf{y} \in \mathbb{R}^{1\times k}$. By definition, the operator norm of $\mathbf{y}$ is $\sup_{\|\mathbf{x}\|=1}\langle \mathbf{y}, \mathbf{x}\rangle$. By the Cauchy-Schwarz inequality $\langle \mathbf{y}, \mathbf{x}\rangle \leq \|\mathbf{x}\|\,\|\mathbf{y}\|$ with equality only $\mathbf{x} = \frac{1}{\|\mathbf{y}\|}\mathbf{y}$. So the supremum is $\frac{1}{\|\mathbf{y}\|}\|\mathbf{y}\|^2 = \|\mathbf{y}\|$. Now if $\mathbf{y} \in \mathbb{R}^{k\times 1}$ then the operator norm norm is $\sup_{|x|=1} \|\mathbf{y}x\| = \sup_{|x|=1} |x|\|\mathbf{y}\| = \|\mathbf{y}\|$. □

Note that Theorem 5.5(a) and (b) imply that $\mathbb{R}^{m\times n}$ forms a metric space with the distance function

$$d(A, B) \;=\; \|A - B\|. \tag{5.5}$$

As we noted earlier, the Frobenius norm defines another natural metric $d_E(A, B) \;=\; \sqrt{\sum_{i,j}(A_{ij} - B_{ij})^2}$. In some applications it may be easier to

work with this metric, or any of the equivalent metrics defined in Section 2.3. A natural question is whether d is equivalent to d_E.

Exercise 5.20. Show that for any $A \in \mathbb{R}^{m\times n}$ it holds that

$$\|A\| \;\leq\; \|A\|_F \;\leq\; \max\{m,n\}\|A\|.$$

Conclude that d_E and d are equivalent.

We conclude this section with an example of how metric structure for matrices can be exploited. We will show that the set of symmetric positive definite matrices is open in the space of all symmetric matrices.

Example 5.4. Recall that a symmetric $n \times n$ matrix is positive definite if $\mathbf{x}^T A\mathbf{x} > 0$ for all $\mathbf{x} \neq \mathbf{0}_n$. Equivalently, $\mathbf{x}^T A\mathbf{x} > 0$ for all $\mathbf{x}$ such that $\|\mathbf{x}\| = 1$. Since $\mathbf{x}^T A\mathbf{x}$ attains its minimum over the unit sphere (compact), we conclude that $\mathbf{x}^T A\mathbf{x} > \delta$ for some $\delta > 0$ (if you know eigenvalues, you can make all this more concrete). Let B be a symmetric matrix such that $\|B - A\| < \frac{\delta}{2}$. By the Cauchy-Schwarz inequality, for any $\mathbf{x}$ such that $\|\mathbf{x}\| = 1$ we get

$$|\mathbf{x}^T(B - A)\mathbf{x}| \;\leq\; \|(B - A)\mathbf{x}\| \;\leq\; \|B - A\| \;<\; \tfrac{\delta}{2}.$$

Then for every $\mathbf{x}$ such that $\|\mathbf{x}\| = 1$ we have

$$\mathbf{x}^T B\mathbf{x} = \mathbf{x}^T(B - A)\mathbf{x} + \mathbf{x}^T A\mathbf{x} > \tfrac{\delta}{2}.$$

This shows that B is also positive definite and so A is an interior point of positive definite matrices.

Chapter 6
Differentiation in one dimension (1 lecture)

Recall the following definitions.

Definition 6.1. The function $f : U \to \mathbb{R}$ defined on an open set $U \subset \mathbb{R}$ is **differentiable at** $a \in U$ with derivative $f'(a)$ if the limit

$$f'(a) = \lim_{h \to 0} \frac{f(a+h)-f(a)}{h} \qquad \text{exists.}$$

We say that f is **differentiable** if it is differentiable at every point. The expression $\frac{f(a+h)-f(a)}{h}$ is called a **difference quotient**.

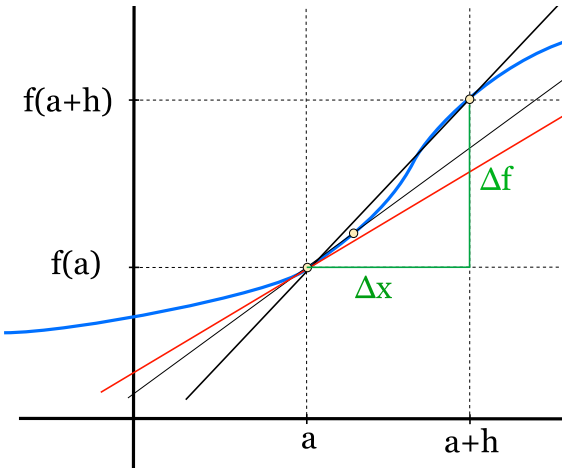

Fig. 6.1 Derivative as the slope of the tangent line.

Figure 6.1 gives a geometric intuition behind this definition. For each $h \neq 0$, the corresponding difference quotient gives the slope of the secant line crossing two points $(a, f(a))$ and $(a+h, f(a+h))$ on the graph of f. Hence,

if f is differentiable, $f'(a)$ is the limit of all these slopes as $h \to 0$. The limit of the secant lines the tangent line coloured in red.

The first basic result assures that differentiability is a stronger condition than continuity.

Theorem 6.1. *Differentiability implies continuity.*

Proof. Continuity is equivalent to $\lim_{h\to 0} f(a+h) = f(a)$. This equality must hold if $f'(a)$ exists. □

It is easy to come up with examples of continuous functions that are not differentiable.

Exercise 6.1. Show that $f(x) = |x|$ is differentiable for all $x \neq 0$ but is *not* differentiable at the origin. Argue the same for

$$f(x) = \begin{cases} x\sin(\frac{1}{x}) & x \neq 0 \\ 0 & x = 0 \end{cases}.$$

In Figure 6.1 we write $\frac{\Delta f}{\Delta x}$ to the denote the difference quotient (for a fixed h). Then $f'(a) = \lim_{\Delta x\to 0} \frac{\Delta f}{\Delta x}$. This notation will be convenient in the rest of this section. We now collect some of the known results from calculus.

Theorem 6.2 (The rules of differentiation).

(a) If f and g are differentiable at a then so is $f + g$ with the derivative

$$(f + g)'(a) \;=\; f'(a) + g'(a).$$

(b) If f and g are differentiable at a then so is $f \cdot g$ with derivative

$$(f \cdot g)'(a) \;=\; f'(a)g(a) + f(a)g'(a).$$

(c) The derivative of a constant function is zero.

(d) If f and g are differentiable at a, and $g(a) \neq 0$, then so is f/g with derivative

$$(f/g)'(a) \;=\; \tfrac{f'(a)g(a)-f(a)g'(a)}{g^2(a)}.$$

(e) If f is differentiable at a and g is differentiable at $f(a)$, then so is $g \circ f$ with derivative

$$(g \circ f)'(a) \;=\; g'(f(a))f'(a).$$

The formula in (b) is called the Leibniz rule and the formula in (e) is called the chain rule.

Proof. (a) The difference quotient for $(f + g)(x)$ is

$$\tfrac{\Delta(f+g)}{\Delta x} = \tfrac{\Delta f}{\Delta x} + \tfrac{\Delta g}{\Delta x}$$

and the limit of the right-hand side is $f'(a) + g'(a)$ by Proposition 3.3 and because g is continuous at a.

(b) The difference quotient for $(f \cdot g)(x)$ is

$$\frac{\Delta(fg)}{\Delta x} = \frac{\Delta f}{\Delta x} g(a+h) + f(a) \frac{\Delta g}{\Delta x}$$

and the limit of the right-hand side is $f'(a)g(a)+f(a)g'(a)$ by Proposition 3.3.

(c) The difference quotient is zero for all h.

(d) The difference quotient for $(f/g)(x)$ is

$$\frac{\Delta(f/g)}{\Delta x} = \frac{\frac{\Delta f}{\Delta x} g(a) - f(a) \frac{\Delta g}{\Delta x}}{g(a+h)g(a)}$$

and the limit of the right-hand side is $\frac{f'(a)g(a)-f(a)g'(a)}{g^2(a)}$ by Proposition 3.3.

(e) By definition of $f'(a)$

$$\frac{f(a+h)-f(a)}{h} = f'(a) + u(h),$$

where $u(h) \to 0$ as $h \to 0$. Similarly

$$\frac{g(f(a)+s)-g(f(a))}{s} = g'(f(a)) + v(s),$$

where $v(s) \to 0$ as $s \to 0$. Take $s = f(a+h) - f(a)$. Then

$$\frac{g(f(a+h))-g(f(a))}{h} = (g'(f(a)) + v(s)) \frac{f(a+h)-f(a)}{h}.$$

As $h \to 0$ also $s \to 0$ (by continuity of f) and the limit of the right-hand side is $g'(f(a))f'(a)$. □

Recall that $\theta \in (a,b)$ is a local minimum (maximum) of $f : (a,b) \to \mathbb{R}$ if there exists a neighborhood $N_r(\theta)$ such that $f(x) \geq f(\theta)$ ($f(x) \geq f(\theta)$) for all $x \in N_r(\theta)$. The following simple lemma will be used in the proof of Theorem 6.4. It gives a necessary condition for f to have a local optimum at $x = \theta$.

Lemma 6.1. *If $f : (a,b) \to \mathbb{R}$ is differentiable and achieves a local minimum or maximum at some $\theta \in (a,b)$ then $f'(\theta) = 0$.*

Proof. Suppose that θ is a local minimum of f and so, for some $r > 0$, $f(x) \geq f(\theta)$ for all $x \in N_r(\theta)$. The difference quotient $(f(\theta+h) - f(\theta))/h$ is positive for $h \in (0,r)$ and negative for $h \in (-r,0)$. Hence the derivative $f'(\theta)$ is the limit of both negative and positive sequences and so it has to be zero. □

Theorem 6.3 (Ratio Mean Value Theorem). *Suppose that f, g are continuous on $[a,b]$ and differentiable on (a,b). Then there is $\theta \in (a,b)$ such*

that

$$(f(b) - f(a))g'(\theta) \;=\; (g(b) - g(a))f'(\theta).$$

Proof. Put $h(x) = (f(b)-f(a))g(x)-(g(b)-g(a))f(x)$. Then h is continuous on $[a, b]$, differentiable on (a, b), and $h(a) = h(b) = f(b)g(a) - f(a)g(b)$. Since $[a, b]$ is compact, Theorem 4.11 implies that h takes on maximum and minimum values, and since it has the same value at both endpoints, h has a maximum or a minimum that occurs at an interior point $\theta \in (a, b)$. By Lemma 6.1, we have $h'(\theta) = 0$ or equivalently $(f(b) - f(a))g'(\theta) \;=\; (g(b) - g(a))f'(\theta)$. □

The following theorem is fundamental for many subsequent results.

Theorem 6.4 (Mean value theorem). *A continuous function $f : [a, b] \to \mathbb{R}$ that is differentiable on (a, b) has the* ***mean value property****: There exists a point $\theta \in (a, b)$ such that*

$$f(b) - f(a) \;=\; f'(\theta)(b - a).$$

Proof. Take $g(x) = x$ in Theorem 6.3. □

Corollary 6.1. *If f is differentiable and $|f'(x)| \le M$ for all $x \in (a, b)$ then f satisfies the* ***Lipschitz condition****: for all $t, x \in (a, b)$ we have*

$$|f(t) - f(x)| \le M|t - x|.$$

In particular, if $f'(x) = 0$ for all $x \in (a, b)$ then $f(x)$ is constant.

Proof. $|f(t) - f(x)| = |f'(\theta)(t - x)|$ for some θ between x and t. □

Exercise 6.2. Let f be differentiable on (a, b). Show that $f'(x) \ge 0$ for all $x \in (a, b)$ implies that f is monotonically increasing in this interval.

Theorem 6.5 (L'Hôpital's Rule). *If f, g are differentiable on (a, b), with $-\infty \le a < b \le +\infty$, and $\lim_{x\to b} f(x) = \lim_{x\to b} g(x) = 0$ then*

$$\lim_{x\to b} \tfrac{f'(x)}{g'(x)} = L \qquad \Longrightarrow \qquad \lim_{x\to b} \tfrac{f(x)}{g(x)} = L.$$

(We assume $g(x), g'(x) \ne 0$ on $x \in (a, b)$.)

We only sketch the proof of this result. Before that we provide a baby version of Theorem 6.5, whose simple proof is left as an exercise.

Theorem 6.6 (Baby L'Hôpital's Rule). *If f, g are differentiable on a, $f(a) = g(a) = 0$, and $g'(a) \ne 0$ then*

$$\tfrac{f'(a)}{g'(a)} = L \qquad \Longrightarrow \qquad \lim_{x\to a} \tfrac{f(x)}{g(x)} = L.$$

Sketch of the proof of Theorem 6.5. Let the sequence $x_n \in (a,b)$ tend to b. Imagine a sequence $t_n \in (a,b)$ tending to b much faster than (x_n) does. Then $f(t_n)/f(x_n)$ and $g(t_n)/g(x_n)$ are as small as we wish (because $\lim_{x\to b} f(x) = \lim_{x\to b} g(x) = 0$), and by Theorem 6.3 there is $\theta_n \in (x_n, t_n)$ such that

$$\tfrac{f(x_n)}{g(x_n)} \;=\; \tfrac{f(x_n)-0}{g(x_n)-0} \;\approx\; \tfrac{f(x_n)-f(t_n)}{g(x_n)-g(t_n)} \;=\; \tfrac{f'(\theta_n)}{g'(\theta_n)}.$$

The latter tends to L because θ_n is sandwiched between x_n and t_n as they tend to b. A rigorous proof relies on a careful construction of the sequence (t_n). □

The derivative of $f'(x)$, if exists, is the second derivative of $f(x)$,

$$f''(a) \;:=\; (f')'(a) \;=\; \lim_{h\to 0} \tfrac{f'(a+h)-f'(a)}{h}.$$

Higher derivatives are defined inductively and written $f^{(r)} = (f^{(r-1)})'$. If $f^{(r)}(a)$ exists then f is **r-th order differentiable at a**. If $f^{(r)}$ exists for each a then f is **r-th order differentiable**. If $f^{(r)}(a)$ exists for all r and all a then f is **smooth**.

The **r-th order Taylor polynomial** of an r-th order differentiable function f at $x = a$ is

$$P(h) \;=\; f(a) + f'(a)h + \tfrac{f''(a)}{2!}h^2 + \cdots + \tfrac{f^{(r)}(a)}{r!}h^r \;=\; \sum_{k=0}^{r} \tfrac{f^{(k)}(a)}{k!}h^k.$$

Exercise 6.3. Show that the r-th order Taylor polynomial of f at $x = a$ satisfies $P^{(k)}(0) = f^{(k)}(a)$ for all $k = 0, 1, \ldots, r$.

Theorem 6.7 (Taylor Approximation Theorem). *Assume that f is a real valued function defined in a neighbourhood of $x = a$ and r-th order differentiable at a. Then*

(a) P approximates f to order r at $x = a$ in the sense that the Taylor remainder

$$R(h) \;=\; f(a+h) - P(h)$$

is r-th order flat at $h = 0$, that is, $R(h)/h^r \to 0$ as $h \to 0$.

(b) The Taylor polynomial is the only polynomial of degree $\le r$ with this approximation property.

(c) If, in addition, f is $(r+1)$-th order differentiable in a neighbourhood of $x = a$ then for some θ between a and $a + h$ we have

$$R(h) \;=\; \tfrac{f^{(r+1)}(\theta)}{(r+1)!}h^{r+1}.$$

Proof. (a) The first r derivatives of $R(h)$ exist and are equal to 0 at $h = 0$ by Exercise 6.3. If $h > 0$ then repeated applications of the Mean Value Theorem

give

$$R(h) = R(h) - 0 \;=\; R'(\theta_1)h \;=\; (R'(\theta_1) - 0)h \;=\; R''(\theta_2)\theta_1 h \;= \\ = \cdots \;=\; R^{(r-1)}(\theta_{r-1})\theta_{r-2}\cdots\theta_1 h$$

where $0 < \theta_{r-1} < \cdots < \theta_1 < h$. Thus

$$\left|\tfrac{R(h)}{h^r}\right| \;=\; \left|\tfrac{R^{(r-1)}(\theta_{r-1})\theta_{r-2}\cdots\theta_1 h}{h^r}\right| \;\le\; \left|\tfrac{R^{(r-1)}(\theta_{r-1})-0}{\theta_{r-1}}\right| \;\to\; R^{(r)}(0) \;=\; 0$$

as $h \to 0$. If $h < 0$ the same is true with $h < \theta_1 < \cdots < \theta_{r-1} < 0$.

(b) If $Q(h)$ is a polynomial of degree $\le r$, $Q \neq P$, then $Q - P$ is not r-th order flat at $h = 0$, so $f(a+h) - Q(h)$ cannot be r-th order flat either.

(c) Fix $h > 0$ and define

$$g(t) = f(a+t) - P(t) - \tfrac{R(h)}{h^{r+1}}t^{r+1} = R(t) - R(h)\tfrac{t^{r+1}}{h^{r+1}}$$

for $0 \le t \le h$. Based on previous calculations we easily verify that $g(0) = g'(0) = \cdots = g^{(r)}(0) = 0$. Directly by construction we also have $g(h) = R(h) - R(h) = 0$. Since $g(0) = g(h) = 0$, the Mean Value Theorem gives a $t_1 \in (0, h)$ such that $g'(t_1) = 0$. Since $g'(0) = g'(t_1) = 0$, the Mean Value Theorem gives a $t_2 \in (0, t_1)$ such that $g''(t_2) = 0$. Continuing, we get a sequence $t_1 > t_2 > \cdots > t_{r+1} > 0$ such that $g^{(k)}(t_k) = 0$ for $k = 1, \ldots, r+1$ and, in particular, $g^{(r+1)}(t_{r+1}) = 0$. On the other hand, since $P(t)$ is a polynomial of degree r, $P^{(r+1)}(t) = 0$ for all t, and so

$$g^{(r+1)}(t) \;=\; f^{(r+1)}(a+t) - (r+1)!\tfrac{R(h)}{h^{r+1}}.$$

Taking $t = t_{r+1}$ gives that

$$R(h) \;=\; \tfrac{f^{(r+1)}(a+t_{r+1})}{(r+1)!}h^{r+1}$$

and thus, $\theta = a + t_{r+1}$ makes the equation in (c) true. If $h < 0$ the argument is symmetric. □

Chapter 7
Differentiation (3 lectures)

7.1 Restricting function to a line

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a function of several variables $\mathbf{x} = (x_1, \ldots, x_k)$. Given $\mathbf{a} \in \mathbb{R}^n$ and a direction vector $\mathbf{u} \in \mathbb{R}^n$ define a line

$$L = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{x} = \mathbf{a} + t\mathbf{u} \text{ for } t \in \mathbb{R}\}.$$

A restriction of $f : \mathbb{R}^n \to \mathbb{R}$ to the line is L is the function g of a single variable t defined as $g(t) = f(\mathbf{a} + t\mathbf{u})$. In particular, $g(0) = f(\mathbf{a})$ and $g(t)$ represents the value of f at points in L. Figure 7.1 shows restriction of $f(x, y) = x^2 - y^2$ to three lines through the origin: $x = y$, $x = 0$ and $y = 0$. In the first case, the restriction is $g(t) = f(t, t) = 0$. In the second case, it is $g(t) = f(0, t) = -t^2$. In the third case, $g(t) = f(t, 0) = t^2$.

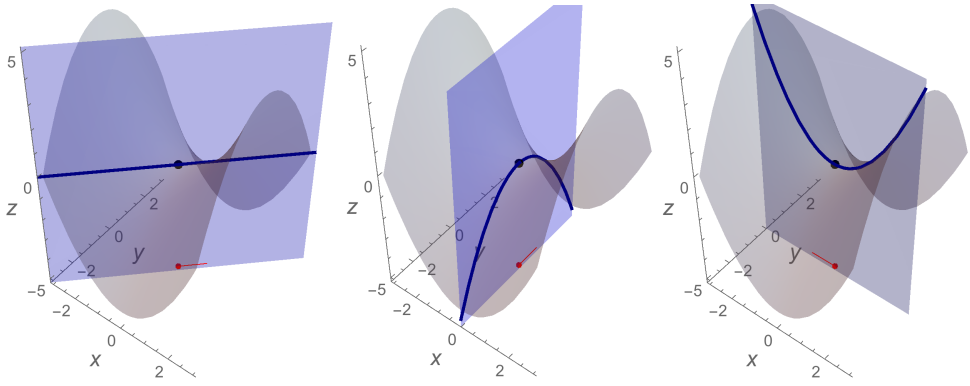

Fig. 7.1 Restriction of $f(x, y) = x^2 - y^2$ to three lines through the origin: $x = y$, $x = 0$, and $y = 0$.

To understand the local behaviour of $\mathbf{a}$ in a neighborhood of $\mathbf{x} = \mathbf{a}$ we can study the local behaviour of all line restrictions in the neighborhood of $t = 0$. This motivates the following definition.

Definition 7.1 (Directional derivative). Let $U \subset \mathbb{R}^n$ be an open set and $\mathbf{a} \in U$. If $\mathbf{u}$ is a vector in $\mathbb{R}^n$, then the directional derivative of $\mathbf{f} : U \to \mathbb{R}^m$ at $\mathbf{a}$ in the direction $\mathbf{u}$ is

$$\mathrm{D}_{\mathbf{u}}\mathbf{f}(\mathbf{a}) \;=\; \lim_{\mathrm{t}\to 0} \frac{\mathbf{f}(\mathbf{a}+\mathrm{t}\mathbf{u}) - \mathbf{f}(\mathbf{a})}{\mathrm{t}}.$$

Example 7.1. Let $f(x, y, z) = e^{x^2+y^2} + \sin(z)$, $\mathbf{a} = \mathbf{0}_3$, and $\mathbf{u} = (2, 2, 1)$. Then

$$g(t) = f(\mathbf{0}_3 + t\mathbf{u}) = f(t(2, 2, 1)) = e^{8t^2} + \sin(t)$$

and

$$g'(t) = 16te^{8t^2} + \cos(t)$$

which implied $\mathrm{D}_{\mathbf{u}}\mathrm{f}(\mathbf{0}_3) = \mathrm{g}'(0) = 1$.

Exercise 7.1. Prove that $\mathrm{D}_{-\mathbf{u}}\mathrm{f}(\mathbf{a}) = -\mathrm{D}_{\mathbf{u}}\mathrm{f}(\mathbf{a})$.

Recall that in the one-dimensional case the derivative of $f'(a)$ gives the instantaneous rate of change of f at a, or in other words it is the limit of $\frac{\Delta f}{\Delta x}$ as $\Delta x \to 0$; c.f. Figure 6.1. In other words, if $h > 0$ is small then

$$\tfrac{f(a+h)-f(a)}{(a+h)-a} \;\to\; f'(a).$$

In the case of a function $f : \mathbb{R}^n \to \mathbb{R}$ we can think of the the instantaneous rate of change of f at $\mathbf{a}$ when we move in the direction $\mathbf{u}$ as the limit with $t > 0$

$$\tfrac{f(\mathbf{a}+t\mathbf{u})-f(\mathbf{a})}{\|(\mathbf{a}+t\mathbf{u})-\mathbf{a}\|} \;=\; \tfrac{f(\mathbf{a}+t\mathbf{u})-f(\mathbf{a})}{t\|\mathbf{u}\|} \;\to\; \tfrac{1}{\|\mathbf{u}\|}\mathrm{D}_{\mathbf{u}}\mathrm{f}(\mathbf{a}).$$

Because of the dependence on the norm of $\mathbf{u}$ it is customary to define the directional derivative for $\mathbf{u} \in \mathbb{R}^n$ such that $\|\mathbf{u}\| = 1$. In that case the directional derivatives represent precisely the instantaneous rate of change of f at $\mathbf{a}$ when we move in the direction $\mathbf{u}$.

Example 7.2. For the function in Example 7.1, the rate of change at $\mathbf{a} = \mathbf{0}_3$ in the direction $\mathbf{u} = (2, 2, 1)$ is $\frac{1}{\|\mathbf{u}\|}\mathrm{D}_{\mathbf{u}}\mathrm{f}(\mathbf{a}) = \frac{1}{3}$.

Exercise 7.2. Consider the function in Figure 7.1. Directly from the picture we see that the rate of change of f at $\mathbf{a} = (0, 0)$ in any of the three directions $(1, 1)$, $(0, 1)$, and $(1, 0)$ is zero. Show that the same is true for *any* direction $\mathbf{u}$.

Definition 7.2 (Partial derivative). Suppose that $U \subset \mathbb{R}^n$ is open. The **i-th partial derivative** of $f : U \to \mathbb{R}$ at a point $\mathbf{a} \in U$, if it exists, is

$$\mathrm{D_i f(a)} \;:=\; \mathrm{D_{e_i} f(a)},$$

where $\mathbf{e}_i$ is the i-th canonical vector in $\mathbb{R}^n$. Extension to higher order derivatives follows by recursion and we write

$$\mathrm{D_{i_1 \cdots i_r}} \;:=\; \mathrm{D_{i_1}} \cdots \mathrm{D_{i_r}}.$$

The vector of partial derivatives is called the **gradient** of f and denoted by $\nabla f(\mathbf{x})$, that is,

$$\nabla f \;=\; (\mathrm{D_1 f}, \ldots, \mathrm{D_n f}) \;\in\; \mathbb{R}^{\mathrm{n}}.$$

The matrix of all second order partial derivatives is called the **Hessian** and it is denoted by $\nabla\nabla^T f(\mathbf{x})$

$$(\nabla\nabla^T f(\mathbf{x}))_{ij} \;=\; \mathrm{D_{ij} f}.$$

As we will see in Theorem 9.2, the Hessian is a symmetric matrix under relatively mild conditions on f.

We finish this section defining the Jacobian, which is a generalization of the gradient to vector-valued maps.

Definition 7.3. The **Jacobian matrix** of a function $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^m$ is the $m \times n$ matrix composed of the n partial derivatives of $\mathbf{f}$ evaluated at $\mathbf{a}$

$$\mathbf{Jf(a)} \;=\; \begin{bmatrix} \mathrm{D_1 f_1(a)} & \cdots & \mathrm{D_n f_1(a)} \\ \vdots & & \vdots \\ \mathrm{D_1 f_m(a)} & \cdots & \mathrm{D_n f_m(a)} \end{bmatrix}.$$

Note that the rows are the gradient vectors of $f_1, \ldots, f_m$.

7.2 Differentiation as a linear operation

The starting point of this section is reformulation of the derivative of a function of one variable. If U is an open interval in $\mathbb{R}$, $a \in U$, and $f : U \to \mathbb{R}$. Then Definition 6.1 can be reformulated as

$$f(a+h) \;=\; f(a) + f'(a) \cdot h + r(h),$$

where the remainder $r(h)$ is small in the sense that

$$\lim_{h \to 0} \tfrac{r(h)}{h} = 0,$$

which we also can write as $r(h) = o(h)$, and we say that $r(h)$ is sublinear. This shows that that $f'(a)h$ is the best linear approximation to $f(a+h) - f(a)$ at $x = a$. We now easily generalize this to the vector case.

Definition 7.4. Suppose U is open in $\mathbb{R}^n$, $\mathbf{f} : U \to \mathbb{R}^m$. The function $\mathbf{f}$ is differentiable at $\mathbf{a} \in U$ with derivative $\mathrm{D}\mathbf{f}(\mathbf{a}) = \mathrm{T}$ if $T : \mathbb{R}^n \to \mathbb{R}^m$ is a linear function and

$$\mathbf{f}(\mathbf{a}+\mathbf{h}) \;=\; \mathbf{f}(\mathbf{a}) + T(\mathbf{h}) + \mathbf{r}(\mathbf{h}),$$

where the remainder $\mathbf{r}(\mathbf{h})$ satisfies

$$\lim_{\mathbf{h}\to\mathbf{0}_n} \tfrac{\mathbf{r}(\mathbf{h})}{\|\mathbf{h}\|} \;=\; \mathbf{0}_m.$$

If $\mathbf{f}$ is differentiable at every $\mathbf{a} \in U$, we say that $\mathbf{f}$ is differentiable in U.

Note that if $\|\mathbf{h}\|$ is small enough, then $\mathbf{a}+\mathbf{h}$ lies in U and so $\mathbf{f}(\mathbf{a}+\mathbf{h})$ is well-defined.

Proposition 3.1 implies that the function $\mathbf{x} \to \|\mathbf{x}\|$ is continuous. The following exercise will be used throughout this chapter.

Exercise 7.3. Let $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^m$. Show that $\lim_{\mathbf{x}\to\mathbf{c}} \mathbf{f}(\mathbf{x}) = \mathbf{0}$ if and only if $\lim_{\mathbf{x}\to\mathbf{c}} \|\mathbf{f}(\mathbf{x})\| = 0$. Conclude that the condition on $\mathbf{r}$ in Definition 7.4 is equivalent to

$$\lim_{\mathbf{h}\to\mathbf{0}_n} \frac{\|\mathbf{r}(\mathbf{h})\|}{\|\mathbf{h}\|} \;=\; 0.$$

Proposition 7.1. *If* $\mathbf{f}$ *is differentiable at* $\mathbf{a}$ *then* $\mathbf{f}$ *is continuous at* $\mathbf{a}$.

Proof. Denote $T = \mathrm{D}\mathbf{f}(\mathbf{a})$. Note that $\mathbf{x} \to \mathbf{a}$ is equivalent to $\mathbf{x} = \mathbf{a}+\mathbf{h}$, where $\mathbf{h} \to \mathbf{0}$. The result follows because

$$\|\mathbf{f}(\mathbf{a}+\mathbf{h}) - \mathbf{f}(\mathbf{a})\| \;\le\; \|\mathbf{T}\| \cdot \|\mathbf{h}\| + \|\mathbf{r}(\mathbf{h})\|.$$

If $\mathbf{h} \to \mathbf{0}$ then the right hand side goes to zero, and so $\|\mathbf{f}(\mathbf{a}+\mathbf{h}) - \mathbf{f}(\mathbf{a})\| \to 0$, or equivalently by Remark 3.2, $\mathbf{f}(\mathbf{a}+\mathbf{h}) \to \mathbf{f}(\mathbf{a})$. □

Remark 7.1. We defined derivative at a point but, since for each $\mathbf{x}$, $\mathrm{D}\mathbf{f}(\mathbf{x}) \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$, we can consider the **derivative function** $\mathrm{Df} : \mathrm{U} \to \mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$.

Each linear function in $\mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$ can be associated with a matrix. What is this matrix for the derivative $T = \mathrm{D}\mathbf{f}(\mathbf{a})$? By the discussion in Section 5.2 the columns of $[T]$ are the vectors $T(\mathbf{e}_j) \in \mathbb{R}^m$ where $\mathbf{e}_j$ are the unit vectors in $\mathbb{R}^n$. To compute $T(\mathbf{e}_j)$ we use the definition of the derivative with $\mathbf{h} = t\mathbf{e}_j$

$$\mathbf{f}(\mathbf{a}+t\mathbf{e}_j) - \mathbf{f}(\mathbf{a}) \;=\; T(t\mathbf{e}_j) + \mathbf{r}(t\mathbf{e}_j) \;=\; tT(\mathbf{e}_j) + \mathbf{r}(t\mathbf{e}_j).$$

Now dividing by $t > 0$ and taking the limit $t \to 0$ implies that $T(\mathbf{e}_j)$ is equal to

$$T(\mathbf{e}_j) \;=\; \lim_{t\to 0} \tfrac{\mathbf{f}(\mathbf{a}+t\mathbf{e}_j)-\mathbf{f}(\mathbf{a})}{t}.$$

The i-th coordinate of this limit is $\mathrm{D_j f_i}(\mathbf{a})$, and so $T(\mathbf{e}_j)$ is the j-th column of the Jacobian $\mathrm{J}\mathbf{f}(\mathbf{a})$. This is an important result to remember

$$[\mathrm{D}\mathbf{f}(\mathbf{a})] \;=\; \mathrm{J}\mathbf{f}(\mathbf{a}), \tag{7.1}$$

which implies, in particular, that the derivative is unique (if exists).

More generally, taking $\mathbf{h} = t\mathbf{u}$ in the definition of the derivative implies the following result.

Proposition 7.2. *Let $\mathbf{f} : U \to \mathbb{R}^m$, where U is an open subset in $\mathbb{R}^n$, be differentiable at $\mathbf{a} \in U$. If $\mathbf{u} \in \mathbb{R}^n$ then*

$$\mathrm{D}_{\mathbf{u}}\mathbf{f}(\mathbf{a}) \;=\; \mathrm{T}(\mathbf{u}) \;=\; \mathrm{J}\mathbf{f}(\mathbf{a}) \cdot \mathbf{u}.$$

Proof. Since $T = \mathrm{D}\mathbf{f}(\mathbf{a})$ exists

$$f(\mathbf{a} + t\mathbf{u}) - \mathbf{f}(\mathbf{a}) \;=\; \mathbf{T}(t\mathbf{u}) + \mathbf{r}(t\mathbf{u})$$

with $\frac{\mathbf{r}(t\mathbf{u})}{t} \to \mathbf{0}$ as $t \to 0$. Dividing by t and taking the limit, we get that

$$T(\mathbf{u}) \;=\; \lim_{t\to 0} \tfrac{\mathbf{f}(\mathbf{a}+t\mathbf{u})-\mathbf{f}(\mathbf{a})}{t},$$

which proves the first equality. The second equality follows by (7.1). □

We briefly discuss the case of real-valued functions. If $f : U \to \mathbb{R}$ is differentiable at $\mathbf{a} \in U \subset \mathbb{R}^n$ then, by Proposition 7.2, the derivative $\mathrm{Df}(\mathbf{a}) \in \mathcal{L}(\mathbb{R}^n, \mathbb{R})$ is represented by the gradient vector

$$\mathrm{D}_{\mathbf{u}}\mathrm{f}(\mathbf{a}) \;=\; \langle \nabla \mathrm{f}(\mathbf{a}), \mathbf{u} \rangle. \tag{7.2}$$

Corollary 7.1. *Suppose that $f : U \to \mathbb{R}$ is differentiable at $\mathbf{a} \in U \subset \mathbb{R}^n$. Then $\nabla f(\mathbf{a})$ is the direction of the quickest rate of increase of f.*

Proof. The rate of change in the direction $\mathbf{u}$ is given by the directional derivative $D_{\mathbf{u}}f(\mathbf{a})$ divided by the norm of $\mathbf{u}$. By the Cauchy-Schwarz inequality $\mathrm{D}_{\mathbf{u}}\mathrm{f}(\mathbf{a}) = \langle \nabla \mathrm{f}(\mathbf{a}), \mathbf{u} \rangle \le \|\nabla \mathrm{f}(\mathbf{a})\| \|\mathbf{u}\|$ and we get equality only if $\mathbf{u}$ is proportional to $\nabla f(\mathbf{a})$. □

Definition 7.5. Let $f : E \to \mathbb{R}$, $E \subset \mathbb{R}^n$. A point $\mathbf{a} \in E$ is a local maximum (minimum) of f if there exists a neighborhood $U \subset E$ of $\mathbf{a}$ such that $f(\mathbf{x}) \le f(\mathbf{a})$ (resp. $f(\mathbf{x}) \ge f(\mathbf{a})$) for all $\mathbf{x} \in U$. We say $\mathbf{a}$ is a local optimum if it is either a local minimum or a local maximum.

Corollary 7.2. *Let $f : U \to \mathbb{R}$, $U \subset \mathbb{R}^n$ open, f differentiable at $\mathbf{a} \in U$. If f has a local optimum at $\mathbf{a}$ then $\nabla f(\mathbf{a}) = \mathbf{0}_n$.*

A natural question is the following: suppose that all partial derivatives exist so that the Jacobian matrix $\mathrm{J}\mathbf{f}(\mathbf{a})$ can be computed. Does it imply that the derivative $\mathrm{D}\mathbf{f}(\mathbf{a})$ exists? The answer, in general, is no.

Exercise 7.4. Show that both partial derivatives of the function

$$f(x,y) = \begin{cases} \frac{xy}{x^2+y^2} & \text{if } (x,y) \neq 0, \\ 0 & \text{if } (x,y) = (0,0). \end{cases}$$

exist at the origin, but the function is not differentiable there.

This shows that the concept of differentiability is more subtle. It turns out however, that if all partial derivatives exist and are continuous, $\mathbf{f}$ must be diferentiable. This is part of Theorem 7.7 in Section 7.4.

7.3 Rules for computing the derivatives

We start with two basic result that treat the case of constant and linear functions.

Exercise 7.5. Let $U \subset \mathbb{R}^n$ be open. Show that if $\mathbf{f} : U \to \mathbb{R}^m$ is a constant function, then $\mathbf{f}$ is differentiable, and its derivative is $\mathbf{0} \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$.

Exercise 7.6. Show that if $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^m$ is linear, then it is differentiable everywhere, and its derivative at all points $\mathbf{a}$ is $\mathbf{f}$, that is $(\mathrm{D}\mathbf{f}(\mathbf{a}))(\mathbf{v}) = \mathbf{f}(\mathbf{v})$.

The following result shows that differentiability of vector valued functions can be checked componentwise.

Theorem 7.1. *Let $U \subset \mathbb{R}^n$ be open. The function $\mathbf{f} = (f_1, \ldots, f_m) : U \to \mathbb{R}^m$ is differentiable at $\mathbf{a}$ if and only if each $f_i : U \to \mathbb{R}$ is. Moreover, the components of $\mathrm{D}\mathbf{f}(\mathbf{a}) \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$ are the derivatives $\mathrm{D}f_1(\mathbf{a}), \ldots, \mathrm{D}f_m(\mathbf{a}) \in \mathcal{L}(\mathbb{R}^n, \mathbb{R})$.*

Proof. Note that the assumption that $\mathbf{f} = (f_1, \ldots, f_m)$ is differentiable, can be written as

$$\lim_{\mathbf{h} \to \mathbf{0}_n} \frac{1}{\|\mathbf{h}\|} \left(\begin{bmatrix} f_1(\mathbf{a}+\mathbf{h}) \\ \cdots \\ f_m(\mathbf{a}+\mathbf{h}) \end{bmatrix} - \begin{bmatrix} f_1(\mathbf{a}) \\ \cdots \\ f_m(\mathbf{a}) \end{bmatrix} - \begin{bmatrix} T_1(\mathbf{h}) \\ \cdots \\ T_m(\mathbf{h}) \end{bmatrix} \right) = \mathbf{0}_m$$

for some linear function $T = (T_1, \ldots, T_m) \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$. The assumption that $f_1, \ldots, f_m$ are differentiable, can be written

$$\begin{bmatrix} \lim_{\mathbf{h} \to \mathbf{0}_n} \frac{1}{\|\mathbf{h}\|} (f_1(\mathbf{a}+\mathbf{h}) - f_1(\mathbf{a}) - T_1(\mathbf{h})) \\ \cdots \\ \lim_{\mathbf{h} \to \mathbf{0}_n} \frac{1}{\|\mathbf{h}\|} (f_m(\mathbf{a}+\mathbf{h}) - f_m(\mathbf{a}) - T_m(\mathbf{h})) \end{bmatrix} = \mathbf{0}_m$$

for some $T_1, \ldots, T_m \in \mathcal{L}(\mathbb{R}^n, \mathbb{R})$. By Theorem 3.2, the expressions on left-hand sides are equal and so both equations are equivalent. □

It is an elementary check that derivative is linear in the following sense.

Theorem 7.2. *Let $U \subset \mathbb{R}^n$ be open. If $\mathbf{f}, \mathbf{g} : U \to \mathbb{R}^m$ are differentiable at $\mathbf{a}$ then so is $s\mathbf{f} + t\mathbf{g}$ for any $s, t \in R$, and*

$$\mathrm{D}(\mathrm{s}\mathbf{f} + \mathrm{t}\mathbf{g})(\mathbf{a}) \;=\; \mathrm{sD}\mathbf{f}(\mathbf{a}) + \mathrm{tD}\mathbf{g}(\mathbf{a}).$$

The next result collects some basic differentiation formulas, which will not be discussed in class in detail. To make things easier to parse we formulate the Jacobian versions of these results.

Theorem 7.3. *Let $U \subset \mathbb{R}^n$ be open.*

1. *If $f : U \to \mathbb{R}$ and $\mathbf{g} : U \to \mathbb{R}^m$ are differentiable at $\mathbf{a}$, then so is $f\mathbf{g}$, and the Jacobian matrix is given by*

$$\mathrm{J}(\mathrm{f}\mathbf{g})(\mathbf{a}) \;=\; \underbrace{f(\mathbf{a})}_{\mathbb{R}} \cdot \underbrace{\mathrm{J}\mathbf{g}(\mathbf{a})}_{\mathbb{R}^{\mathrm{m}\times\mathrm{n}}} + \underbrace{\mathbf{g}(\mathbf{a})}_{\mathbb{R}^{\mathrm{m}\times 1}} \cdot \underbrace{\mathrm{J}\mathrm{f}(\mathbf{a})}_{\mathbb{R}^{1\times\mathrm{n}}}.$$

2. *If $f : U \to \mathbb{R}$ and $\mathbf{g} : U \to \mathbb{R}^m$ are differentiable at $\mathbf{a}$ and $f(\mathbf{a}) \neq 0$, then so is $\mathbf{g}/f$, and the Jacobian is given by*

$$\mathrm{J}(\mathbf{g}/\mathrm{f})(\mathbf{a}) \;=\; \tfrac{1}{\mathrm{f}(\mathbf{a})} \cdot \mathrm{J}\mathbf{g}(\mathbf{a}) - \tfrac{1}{(\mathrm{f}(\mathbf{a}))^2} \cdot \mathbf{g}(\mathbf{a}) \cdot \mathrm{J}\mathrm{f}(\mathbf{a}).$$

3. *If $\mathbf{f}, \mathbf{g} : U \to \mathbb{R}^m$ are differentiable at $\mathbf{a}$ then so is the scalar product $\mathbf{f}^T\mathbf{g} : U \to \mathbb{R}$, and the Jacobian is given by*

$$\mathrm{J}\langle \mathbf{f}, \mathbf{g}\rangle(\mathbf{a}) \;=\; \mathbf{g}(\mathbf{a})^{\mathrm{T}}\mathrm{J}\mathbf{f}(\mathbf{a}) + \mathbf{f}(\mathbf{a})^{\mathrm{T}}\mathrm{J}\mathbf{g}(\mathbf{a}).$$

Proof. To prove 1 we need to show that

$$\tfrac{1}{\|\mathbf{h}\|}\big(f(\mathbf{a}+\mathbf{h})\mathbf{g}(\mathbf{a}+\mathbf{h}) - f(\mathbf{a})\mathbf{g}(\mathbf{a}) - (f(\mathbf{a})\mathrm{J}\mathbf{g}(\mathbf{a}) - \mathbf{g}(\mathbf{a})\mathrm{J}\mathrm{f}(\mathbf{a})) \cdot \mathbf{h}\big)$$

converges to $\mathbf{0}$ as $\mathbf{h} \to \mathbf{0}$ rewrite it as

$$\mathbf{g}(\mathbf{a})\tfrac{f(\mathbf{a}+\mathbf{h}) - f(\mathbf{a}) - \mathrm{J}\mathrm{f}(\mathbf{a})\cdot\mathbf{h}}{\|\mathbf{h}\|} + f(\mathbf{a}+\mathbf{h})\tfrac{\mathbf{g}(\mathbf{a}+\mathbf{h}) - \mathbf{g}(\mathbf{a}) - \mathrm{J}\mathbf{g}(\mathbf{a})\cdot\mathbf{h}}{\|\mathbf{h}\|} + \\ +(f(\mathbf{a}+\mathbf{h}) - f(\mathbf{a}))\tfrac{\mathrm{J}\mathbf{g}(\mathbf{a})\cdot\mathbf{h}}{\|\mathbf{h}\|}.$$

Now the first term converges to zero by the definition of $\mathrm{J}\mathrm{f}(\mathbf{a})$. The second term converges to zero by continuity of f and the definition of $\mathrm{J}\mathbf{g}(\mathbf{a})$. The third term also converges to zero by continuity of f and the fact that $\|\mathrm{J}\mathbf{g}(\mathbf{a}) \cdot \mathbf{h}\|/\|\mathbf{h}\| \leq \|\mathrm{J}\mathbf{g}(\mathbf{a})\|$. The proof of item 2 is left as an exercise. To prove item 3

$$\mathrm{J}\langle \mathbf{f}, \mathbf{g}\rangle(\mathbf{a}) = \sum_{\mathrm{i}=1}^{\mathrm{m}} \mathrm{J}(\mathrm{f_i g_i})(\mathbf{a}) = \sum_{\mathrm{i}=1}^{\mathrm{m}} \mathrm{g_i}(\mathbf{a}) \cdot \mathrm{J}\mathrm{f_i}(\mathbf{a}) + \sum_{\mathrm{i}=1}^{\mathrm{m}} \mathrm{f_i}(\mathbf{a}) \cdot \mathrm{J}\mathrm{g_i}(\mathbf{a}).$$

The result follows because Theorem 7.1 implies that $\mathrm{J}\mathrm{f_i}(\mathbf{a})$ and $\mathrm{J}\mathrm{g_i}(\mathbf{a})$ are the coordinate functions of $\mathrm{J}\mathbf{f}(\mathbf{a})$ and $\mathrm{J}\mathbf{g}(\mathbf{a})$ respectively. □

Example 7.3 (Least squares for nonlinear regression). Consider a regression function $y = f(\mathbf{x}, \theta)$, where $y \in \mathbb{R}$, $\mathbf{x} \in \mathbb{R}^m$ and $\theta \in \mathbb{R}^d$. Given data $(y_i, \mathbf{x}_i)$ for $i = 1, \ldots, n$ we compute the residual vector $\mathbf{r}(\theta) \in \mathbb{R}^n$ whose i-th coordinate is $r_i(\theta) = y_i - f(\mathbf{x}_i, \theta)$. We estimate θ from data by minimizing the norm of $\mathbf{r}(\theta)$, or equivalently, by minimizing $s(\theta) = \frac{1}{2}\mathbf{r}^T(\theta)\mathbf{r}(\theta)$. One of the fundamental quantities to compute for this optimization problem is the gradient $\nabla s(\theta)$. By Theorem 7.3(3)

$$\mathrm{J}s(\theta) \;=\; \tfrac{1}{2}\mathrm{J}(\mathbf{r}^{\mathrm{T}}\mathbf{r})(\theta) \;=\; \mathbf{r}(\theta)^{\mathrm{T}}\mathrm{J}\mathbf{r}(\theta)$$

and so $\nabla s(\theta) = \mathrm{J}\mathbf{r}(\theta)^{\mathrm{T}}\mathbf{r}(\theta)$.

The most fundamental rule is the chain rule and we will state it in a separate theorem.

Theorem 7.4 (Chain rule). *Let $U \subset \mathbb{R}^n$, $V \subset \mathbb{R}^m$ be open sets, let $\mathbf{f} : U \to V$ and $\mathbf{g} : V \to \mathbb{R}^p$ be mappings, and let $\mathbf{a} \in U$. If $\mathbf{f}$ is differentiable at $\mathbf{a}$ and $\mathbf{g}$ is differentiable at $\mathbf{b} = \mathbf{f}(\mathbf{a})$, then the composition $\mathbf{F} = \mathbf{g} \circ \mathbf{f}$ is differentiable at $\mathbf{a}$, and its derivative is given by*

$$\mathrm{D}(\mathbf{g} \circ \mathbf{f})(\mathbf{a}) \;=\; \mathrm{D}\mathbf{g}(\mathbf{f}(\mathbf{a})) \circ \mathrm{D}\mathbf{f}(\mathbf{a}).$$

Proof. The proof follows exactly the same lines as in the univariate case. Let $S = \mathrm{D}\mathbf{f}(\mathbf{a})$, $T = \mathrm{D}\mathbf{g}(\mathbf{b})$. We have

$$\mathbf{f}(\mathbf{a}+\mathbf{h}) - \mathbf{f}(\mathbf{a}) = S(\mathbf{h}) + \mathbf{u}(\mathbf{h})\|\mathbf{h}\|, \qquad \mathbf{g}(\mathbf{b}+\mathbf{k}) - \mathbf{g}(\mathbf{b}) = T(\mathbf{k}) + \mathbf{v}(\mathbf{k})\|\mathbf{k}\|,$$

where

$$u(\mathbf{h}) \underset{\mathbf{h}\to\mathbf{0}_n}{\longrightarrow} 0, \qquad \mathbf{v}(\mathbf{k}) \underset{\mathbf{k}\to\mathbf{0}_m}{\longrightarrow} 0$$

Given $\mathbf{h}$ put $\mathbf{k} = \mathbf{f}(\mathbf{a}+\mathbf{h}) - \mathbf{f}(\mathbf{a})$ (so that $\mathbf{k} \to \mathbf{0}$ if $\mathbf{h} \to \mathbf{0}$). Then

$$\mathbf{g}(\mathbf{f}(\mathbf{a}+\mathbf{h})) - \mathbf{g}(\mathbf{f}(\mathbf{a})) = T(\mathbf{k}) + \mathbf{v}(\mathbf{k})\|\mathbf{k}\| = T(S(\mathbf{h})) + T(\mathbf{u}(\mathbf{h})\|\mathbf{h}\|) + \mathbf{v}(\mathbf{k})\|\mathbf{k}\|.$$

It remains to show that the remainder $T(\mathbf{u}(\mathbf{h})\|\mathbf{h}\|) + \mathbf{v}(\mathbf{k})\|\mathbf{k}\|$ is sublinear

$$\tfrac{T(\mathbf{u}(\mathbf{h})\|\mathbf{h}\|)+\mathbf{v}(\mathbf{k})\|\mathbf{k}\|}{\|\mathbf{h}\|} = T\left(\mathbf{u}(\mathbf{h})\right) + \tfrac{\mathbf{v}(\mathbf{k})\|\mathbf{k}\|}{\|\mathbf{h}\|}.$$

The limit of the first summand is zero by continuity of T. Same is true for the second summand because $\lim_{\mathbf{h}\to\mathbf{0}} \mathbf{v}(\mathbf{k}) = \mathbf{0}$ and by continuity of the norm

$$\lim_{\mathbf{h}\to\mathbf{0}} \tfrac{\|\mathbf{k}\|}{\|\mathbf{h}\|} = \left\| \lim_{\mathbf{h}\to\mathbf{0}} \frac{\mathbf{f}(\mathbf{a}+\mathbf{h}) - \mathbf{f}(\mathbf{a})}{\|\mathbf{h}\|} \right\| = \|\mathrm{D}\mathbf{f}(\mathbf{a})\|.$$

□

Example 7.4. Define $\mathbf{f} : \mathbb{R} \to \mathbb{R}^3$ and $g : \mathbb{R}^3 \to \mathbb{R}$ by

$$\mathbf{f}(t) = \begin{bmatrix} t \\ t^2 \\ t^3 \end{bmatrix}; \qquad g(x,y,z) = x^2 + y^2 + z^2.$$

The derivative of g is a linear transformation $T : \mathbb{R}^3 \to \mathbb{R}$ represented by the vector $[2x, 2y, 2z]$. Evaluated at $\mathbf{f}(t)$ it is $[2t, 2t^2, 2t^3]$. The derivative of $\mathbf{f}$ at t is the linear map $S : \mathbb{R} \to \mathbb{R}^3$ represented by the vector $[1, 2t, 3t^2]^T$. So

$$\mathrm{J}(\mathrm{g} \circ \mathbf{f})(\mathrm{t}) = \mathrm{Jg}(\mathbf{f}(\mathrm{t})) \cdot \mathbf{Jf}(\mathrm{t}) = \begin{bmatrix} 2t & 2t^2 & 2t^3 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ 2t \\ 3t^2 \end{bmatrix} = 2\mathrm{t} + 4\mathrm{t}^3 + 6\mathrm{t}^5.$$

Note that we could also do all calculations directly because $g(\mathbf{f}(t)) = t^2 + t^4 + t^6$.

7.4 Mean Value Theorem and $\mathcal{C}^1$ functions

We start by generalizing Theorem 6.4 to functions of many variables.

Theorem 7.5. *Let $U \subset \mathbb{R}^n$ be open, $f : U \to \mathbb{R}$ be differentiable, and the segment $[\mathbf{a}, \mathbf{b}]$ joining $\mathbf{a}$ and $\mathbf{b}$ be contained in U. Then there exists $\mathbf{c} \in [\mathbf{a}, \mathbf{b}]$ such that*

$$f(\mathbf{b}) - f(\mathbf{a}) \;=\; (\mathrm{Df}(\mathbf{c}))(\mathbf{b} - \mathbf{a}).$$

Proof. Consider the function $g : [0,1] \to \mathbb{R}$ given by $g(t) = f((1-t)\mathbf{a} + t\mathbf{b})$. By the chain rule, g is differentiable, and by the one-variable mean value theorem, there exists $\theta \in (0,1)$ such that

$$g(1) - g(0) \;=\; g'(\theta)(1-0) \;=\; g'(\theta). \tag{7.3}$$

Set $\mathbf{c} = (1-\theta)\mathbf{a} + \theta\mathbf{b}$. We can express $g'(\theta)$ in terms of the derivative of f:

$$g'(\theta) \;=\; \lim_{s\to 0} \tfrac{g(\theta+s)-g(\theta)}{s} \;=\; \lim_{s\to 0} \tfrac{f(\mathbf{c}+s(\mathbf{b}-\mathbf{a}))-f(\mathbf{c})}{s} \;=\; (\mathrm{Df}(\mathbf{c}))(\mathbf{b}-\mathbf{a}),$$

where the last equation follows by Proposition 7.2. Equation (7.3) reads

$$f(\mathbf{b}) - f(\mathbf{a}) = (\mathrm{Df}(\mathbf{c}))(\mathbf{b} - \mathbf{a}).$$

□

We also have a version of Corollary 6.1 for vector-valued functions. It will be useful in the proof of the Inverse Function Theorem.

Theorem 7.6. *Let $\mathbf{f} : U \to \mathbb{R}^m$, $U \subset \mathbb{R}^n$ open and convex. Assume $\mathbf{f}$ is differentiable in U and there exists $M \in \mathbb{R}$ such that $\|\mathrm{D}\mathbf{f}(\mathbf{x})\| \leq \mathrm{M}$ for all $\mathbf{x} \in U$. Then*

$$\|\mathbf{f}(\mathbf{b}) - \mathbf{f}(\mathbf{a})\| \leq M \cdot \|\mathbf{b} - \mathbf{a}\| \qquad \textit{for all } \mathbf{a}, \mathbf{b} \in U.$$

Proof. Put $\mathbf{z} = \mathbf{f}(\mathbf{b}) - \mathbf{f}(\mathbf{a})$ and define $\varphi(\mathbf{t}) = \langle \mathbf{z}, \mathbf{f}(\mathbf{t}) \rangle$ for $\mathbf{t} \in [\mathbf{a}, \mathbf{b}]$. Using Proposition 7.2 (or the chain rule), for every $\mathbf{c} \in U$ and $\mathbf{u} \in \mathbb{R}^n$ we get

$$\begin{aligned}(\mathrm{D}\varphi(\mathbf{c}))(\mathbf{u}) &= \lim_{t\to 0} \tfrac{\varphi(\mathbf{c}+t\mathbf{u})-\varphi(\mathbf{c})}{t} = \\ &= \langle \mathbf{z}, \lim_{t\to 0} \tfrac{\mathbf{f}(\mathbf{c}+t\mathbf{u})-\mathbf{f}(\mathbf{c})}{t} \rangle = \langle \mathbf{z}, (\mathrm{D}\mathbf{f}(\mathbf{c}))(\mathbf{u}) \rangle.\end{aligned}$$

By Theorem 7.5, for some $\mathbf{c}$ on the segment joining $\mathbf{a}$ and $\mathbf{b}$

$$\varphi(\mathbf{b}) - \varphi(\mathbf{a}) = (\mathrm{D}\varphi(\mathbf{c}))(\mathbf{b} - \mathbf{a}) = \langle \mathbf{z}, (\mathrm{D}\mathbf{f}(\mathbf{c}))(\mathbf{b} - \mathbf{a}) \rangle.$$

On the other hand $\varphi(\mathbf{b}) - \varphi(\mathbf{a}) = \|\mathbf{z}\|^2$. By the Cauchy-Schwarz inequality and a basic matrix norm bound

$$\|\mathbf{z}\|^2 = \langle \mathbf{z}, (\mathrm{D}\mathbf{f}(\mathbf{c}))(\mathbf{b} - \mathbf{a}) \rangle \leq \|\mathbf{z}\| \, \|(\mathrm{D}\mathbf{f}(\mathbf{c}))(\mathbf{b} - \mathbf{a}))\| \leq \|\mathbf{z}\| \, \|\mathrm{D}\mathbf{f}(\mathbf{c})\| \, \|\mathbf{b} - \mathbf{a}\|.$$

Since the theorem trivially holds if $\mathbf{z} = \mathbf{0}$ we can assume the norm of $\mathbf{z}$ is positive. Now dividing by $\|\mathbf{z}\|$ gives the claim. □

Definition 7.6. A function $\mathbf{f} : U \to \mathbb{R}^m$ is continuously differentiable on an open set $U \subset \mathbb{R}^n$ if all its partial derivatives exist and are continuous on U. Such a function is known as a $\mathcal{C}^1$ function. We write $\mathbf{f} \in \mathcal{C}^1(U)$.

This definition can be generalized.

Definition 7.7. A $\mathcal{C}^p$ function on $U \subset \mathbb{R}^n$ is a function that is p times continuously differentiable: all its partial derivatives up to order p exist and and are continuous on U.

The following theorem guarantees that a continuously differentiable function is indeed differentiable.

Theorem 7.7. *If U is an open subset of $\mathbb{R}^n$, and $\mathbf{f} : U \to \mathbb{R}^m$ is a $\mathcal{C}^1$ function, then $\mathbf{f}$ is differentiable on U. Moreover, the function $\mathrm{D}\mathbf{f} : \mathrm{U} \to \mathcal{L}(\mathbb{R}^\mathrm{n}, \mathbb{R}^\mathrm{m})$ is continuous on U.*

Proof. By Theorem 7.1(3), to show that $\mathbf{f}$ is differentiable, it suffices to consider the case $m = 1$, $f : U \to \mathbb{R}$. Since the partial derivatives $\mathrm{D_i f}(\mathbf{x})$ exist on U, we need to show that for every $\mathbf{a} \in U$

$$\lim_{\mathbf{h}\to\mathbf{0}} \tfrac{|f(\mathbf{a}+\mathbf{h})-f(\mathbf{a})-\langle \nabla f(\mathbf{a}), \mathbf{h} \rangle|}{\|\mathbf{h}\|} = 0.$$

We can write $\mathbf{h} = \sum_{i=1}^n h_i \mathbf{e}_i$ and let $\mathbf{v}_k = h_1 \mathbf{e}_1 + \cdots + h_k \mathbf{e}_k$ for $k = 1, \ldots, n$ with $\mathbf{v}_0 = \mathbf{0}$. Then $\mathbf{a} + \mathbf{v}_0 = \mathbf{a}$, $\mathbf{a} + \mathbf{v}_n = \mathbf{a} + \mathbf{h}$, and so

$$f(\mathbf{a} + \mathbf{h}) - f(\mathbf{a}) = \sum_{i=1}^{n} \left(f(\mathbf{a} + \mathbf{v}_i) - f(\mathbf{a} + \mathbf{v}_{i-1}) \right).$$

For $i = 1, \ldots, n$ suppose that $h_i > 0$ and let $g_i : \mathbb{R} \to \mathbb{R}$ be defined for $t \in [0, h_i]$ by

$$g_i(t) \;=\; f(\mathbf{a} + \mathbf{v}_{i-1} + t\mathbf{e}_i)$$

so that $g_i(0) = f(\mathbf{a} + \mathbf{v}_{i-1})$ and $g_i(h_i) = f(\mathbf{a} + \mathbf{v}_i)$. Assume that $\mathbf{h}$ is small enough so that the segment joining $\mathbf{a} + \mathbf{v}_{i-1}$ and $\mathbf{a} + \mathbf{v}_i$ lies in U. Because the partial derivatives of f exist in U, $g_i(t)$ is differentiable in $(0, h_i)$ with derivative

$$g_i'(t) = \lim_{s \to 0} \tfrac{f(\mathbf{a}+\mathbf{v}_{i-1}+(s+t)\mathbf{e}_i) - f(\mathbf{a}+\mathbf{v}_{i-1}+t\mathbf{e}_i)}{s} = D_i f(\mathbf{a} + \mathbf{v}_{i-1} + t\mathbf{e}_i).$$

Theorem 6.4 gives that for some $\theta_i \in (0, h_i)$

$$g_i(h_i) - g_i(0) \;=\; g_i'(\theta_i) h_i$$

or, in other words, for $\mathbf{c}_i := \mathbf{a} + \mathbf{v}_{i-1} + \theta_i \mathbf{e}_i$

$$f(\mathbf{a} + \mathbf{v}_i) - f(\mathbf{a} + \mathbf{v}_{i-1}) \;=\; \mathrm{D_i f(c_i) h_i}.$$

Thus we get

$$f(\mathbf{a} + \mathbf{h}) - f(\mathbf{a}) - \sum_{i=1}^{n} \mathrm{D_i f(a) h_i} = \sum_{\mathrm{i=1}}^{\mathrm{n}} \mathrm{h_i((D_i f)(c_i) - (D_i f)(a))}$$

Noting that $\sum_{i=1}^{n} \mathrm{D_i f(\mathbf{a}) h_i} = \langle \nabla \mathrm{f}(\mathbf{a}), \mathbf{h} \rangle$ and using the Cauchy-Schwarz inequality for the expression on the right we get

$$\tfrac{|f(\mathbf{a}+\mathbf{h}) - f(\mathbf{a}) - \langle \nabla f(\mathbf{a}), \mathbf{h} \rangle|}{\|\mathbf{h}\|} \;\leq\; \sqrt{\sum_{i=1}^{n} ((\mathrm{D_i f})(\mathbf{c_i}) - (\mathrm{D_i f})(\mathbf{a}))^2}.$$

Since the partial derivatives $\mathrm{D_i f}$ are continuous, and since $\mathbf{c}_i \to \mathbf{a}$ as $\mathbf{h} \to \mathbf{0}$, the right-hand side above goes to 0 as $\mathbf{h} \to \mathbf{0}$ proving that f is differentiable at $\mathbf{a}$.

We now show that that $\mathrm{D}\mathbf{f}(\mathbf{x})$ is continuous in U. By Theorem 3.3, a function is continuous if and only if its coordinate functions are continuous. It follows that the Jacobian function $\mathbf{x} \mapsto \mathrm{J}\mathbf{f}(\mathbf{x})$ is continuous in U. Recall that the metric structure in $\mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$ was induced from $\mathbb{R}^{m \times n}$ and so continuity of $\mathrm{J}\mathbf{f}(\mathbf{x})$ is equivalent to continuity of $\mathrm{D}\mathbf{f}(\mathbf{x})$. □

7.5 The Jacobian matrix: not always the right approach*

Computing the derivative of a function $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^m$ by computing its Jacobian matrix is usually a lot quicker than computing it from the definition just as in one-variable calculus it is quicker to compute derivatives using handy formulas rather than computing a limit. But it is important to keep in mind the meaning of the derivative: it is the best linear approximation to a function, the linear function T such that

$$\lim_{\mathbf{h}\to\mathbf{0}} \tfrac{1}{\|\mathbf{h}\|} \left(\mathbf{f}(\mathbf{a}+\mathbf{h}) - \mathbf{f}(\mathbf{a}) - T(\mathbf{h})\right) \;=\; \mathbf{0}_m.$$

This is the definition that can always be used to compute a derivative even when the domain and codomain of a function are abstract vector spaces, not $\mathbb{R}^n$. Although it may be possible to identify such a function with a function from $\mathbb{R}^n$ to $\mathbb{R}^m$, and thus compute a Jacobian matrix, doing so may be quite tedious. We will see an example of this in the proof of Theorem 8.5. Two different examples are given below.

Example 7.5 (Squaring function for matrices). Let $\mathbf{f} : \mathbb{R}^{n\times n} \to \mathbb{R}^{n\times n}$ be given by $\mathbf{f}(A) = A^2$. We can of course think about this function as a function from $\mathbb{R}^{n^2} \to \mathbb{R}^{n^2}$ and compute its Jacobian. But the computations become extremely complicated. On the other hand, sticking to matrices we easily see that for every $H \in \mathbb{R}^{n\times n}$

$$\mathbf{f}(A+H) - \mathbf{f}(A) \;=\; AH + HA + H^2.$$

Define $T(H) = AH + HA$. Clearly $T \in \mathcal{L}(\mathbb{R}^{n\times n}, \mathbb{R}^{n\times n})$ and

$$\lim_{H\to\mathbf{0}_{n\times n}} \tfrac{1}{\|H\|}(\mathbf{f}(A+H) - \mathbf{f}(A) - T(H)) \;=\; \lim_{H\to\mathbf{0}_{n\times n}} \tfrac{1}{\|H\|} H^2.$$

This limit is $\mathbf{0}_{n\times n}$ because the

$$\|\tfrac{1}{\|H\|} H^2\| = \tfrac{1}{\|H\|}\|H^2\| \le \|H\|.$$

Another way to see these calculations is to take $H = \epsilon U$, where ϵ is infinitesimally small and so all higher powers of ϵ are treated as zero. So

$$\mathbf{f}(A+\epsilon U) - \mathbf{f}(A) \;=\; \epsilon(AU + UA) + O(\epsilon^2),$$

which again gives the same result.

The following example is useful in statistics and optimization and it will much better illustrate usefulness of the second technique.

Example 7.6. Let $\mathbb{S}^n$ denote the set of $n \times n$ symmetric matrices. Clearly $\mathbb{S}^n$ forms a vector space isomorphic to $\mathbb{R}^k$, $k = \binom{n+1}{2}$. This space is equipped with

the standard inner product $\langle A, B\rangle = \mathrm{tr}(AB)$. Denote by $\mathbb{S}^n_+$ the subset of $\mathbb{S}_n$ that are positive definite. This set is open. Consider the function $\mathbf{f} : \mathbb{S}^n_+ \to \mathbb{R}$ given by $\mathbf{f}(A) = \log\det(A)$. Let $U \in \mathbb{S}^n$ then

$$\mathbf{f}(A+\epsilon U) - \mathbf{f}(A) \;=\; \log\det(A+\epsilon U) - \log\det(A) \;=\; \log\det(\mathbb{I}_n + \epsilon U A^{-1}),$$

where $\mathbb{I}_n$ denotes the $n\times n$ identity matrix. We can rewrite $\det(\mathbb{I}_n+\epsilon U A^{-1}) = \prod_i(1+\epsilon\lambda_i)$, where λ_i are the eigenvalues of UA^{-1}. We now get

$$\log\det(\mathbb{I}_n+\epsilon U A^{-1}) \;=\; \sum_i \log(1+\epsilon\lambda_i) \;=\; \epsilon\sum_i \lambda_i + O(\epsilon^2) \;=\; \epsilon\,\mathrm{tr}(UA^{-1}) + O(\epsilon^2).$$

This shows that the derivative $T = \mathrm{D}\mathbf{f}(\mathrm{A})$ satisfies

$$(\mathrm{D}\mathbf{f}(\mathrm{A}))(\mathrm{U}) = \mathrm{tr}(\mathrm{U}\mathrm{A}^{-1}) = \langle \mathrm{U}, \mathrm{A}^{-1}\rangle.$$

Generalising (7.2) we see that A^{-1} is the element of $\mathbb{S}^n$ representing this derivative (is the gradient of $\mathbf{f}$).

Working with linear transformations also gives quick ways of computing the derivatives without carefully computing all partial derivatives.

Example 7.7. Suppose we want to find the derivative for the quadratic function $f(\mathbf{x}) = \mathbf{x}^T A\mathbf{x} + \mathbf{x}^T\mathbf{b} + c$, where A is a symmetric $n\times n$ matrix, $\mathbf{b}\in\mathbb{R}^n$, and $c\in\mathbb{R}$. We have

$$f(\mathbf{x}+\mathbf{h}) - f(\mathbf{x}) \;=\; \mathbf{h}^T A\mathbf{h} + 2\mathbf{x}^T A\mathbf{h} + \mathbf{b}^T\mathbf{h}.$$

The linear part of this expression is $T(\mathbf{h}) := (2\mathbf{x}^T A + \mathbf{b}^T)\mathbf{h}$. Since $\|\mathbf{h}^T A\mathbf{h}\| \le \|A\|\,\|\mathbf{h}\|^2$ we conclude that the remainder is indeed sublinear.

Exercise 7.7. Let $A \in \mathbb{R}^{m\times n}$. Find the derivative of $f : \mathbb{R}^{m\times n} \to \mathbb{R}$ given by $f(X) = \mathrm{trace}(AX^T)$.

Chapter 8
Solving systems of equations (2 lectures)

8.1 Solving linear equations (self-study)

A system of linear equations is a collection of m linear equations in n variables $x_1, \ldots, x_n$:

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= b_1 \\ &\cdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n &= b_m \end{aligned}$$

where $a_{11}, \ldots, a_{mn}$ and $b_1, \ldots b_m$ are fixed numbers, which we write in form of a matrix A and a vector $\mathbf{b}$. This system can be more compactly written in the matrix notation as $A\mathbf{x} = \mathbf{b}$. In this section we briefly recall the method to solve a system of linear equations.

Denote by $[A|\mathbf{b}]$ the $\mathbb{R}^{m\times(n+1)}$ matrix obtained by adding $\mathbf{b}$ as the last column to A. The canonical way of solving $A\mathbf{x} = \mathbf{b}$ is by the row operations on $[A|\mathbf{b}]$.

Definition 8.1 (Row operations). A row operation on a matrix is one of the following three operations:

1. Multiplying a row by a nonzero number, $\alpha \neq 0$
2. Adding a multiple of a row onto another row
3. Exchanging two rows

The following theorem should be well known.

Theorem 8.1. *If the matrix $[A|\mathbf{b}]$ representing $A\mathbf{x} = \mathbf{b}$ can be turned into $[A'|\mathbf{b}']$ after a sequence of row operations, then the set of solutions to $A\mathbf{x} = \mathbf{b}$ and the set of solutions to $A'\mathbf{x} = \mathbf{b}'$ coincide.*

Proof. Row operation consist of multiplying one equation by a nonzero number, adding a multiple of one equation to another, and exchanging two equations. Any solution to $A\mathbf{x} = \mathbf{b}$ is then also a solution of $A'\mathbf{x} = \mathbf{b}'$. In the

other direction, any row operation can be undone by another row operation (exercise), so any solution to $A'\mathbf{x} = \mathbf{b}'$ is also a solution to $A\mathbf{x} = \mathbf{b}$. □

The idea now is to use row operations to reduce $[A|\mathbf{b}]$ to a much simpler form.

Definition 8.2 (Echelon form). A matrix is in echelon form if

1. In every row the first nonzero entry is 1 and it is called a pivotal 1.
2. In every column that contains a pivotal 1, all other entries are zero.
3. The pivotal 1 of a lower row is always to the right of the pivotal 1 of a higher row.
4. Any rows consisting entirely of 0's are at the bottom.

Given any matrix B there exists a unique matrix $\tilde{B}$ in echelon form that can be obtained from B by row operations. If $[A|\mathbf{b}]$ is reduced to the echelon form, the solutions of $A\mathbf{x} = \mathbf{b}$ are trivial to write down. For example

$$\begin{bmatrix} 1 & 0 & 0 & 3 \\ 0 & 1 & 0 & -2 \\ 0 & 0 & 1 & 1 \end{bmatrix}$$

is in the echelon form and the corresponding linear equation has one solution $(x_1, x_2, x_3) = (3, -2, 1)$. Also

$$\begin{bmatrix} 1 & 2 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

is in the echelon form but the corresponding system of linear equations has no solution as the last equation says $0 = 1$. Finally,

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

represents a linear system with infinitely many solutions that satisfy $x_1 + x_2 = 0$, $x_3 = 0$. For example, $(-1, 1, 0)$, $(2, -2, 0)$ are both solutions of this system. More generally,

Theorem 8.2. *Represent the system $A\mathbf{x} = \mathbf{b}$, involving m linear equations in n unknowns, by the $m \times (n+1)$ matrix $[A|\mathbf{b}]$, which can be row reduced to $[\tilde{A}|\tilde{\mathbf{b}}]$ in the echelon form. Then*

1. *If $\tilde{\mathbf{b}}$ contains a pivotal 1, the system has no solutions.*
2. *If $\tilde{\mathbf{b}}$ contains no pivotal 1, then*
   a. *If each column of $\tilde{A}$ contains a pivotal 1, the system has a unique solution.*

b. If at least one column of $\tilde{A}$ does not contain a pivotal 1, there are infinitely many solutions. You can choose freely the values of unknowns corresponding to nonpivotal columns of $\tilde{A}$, and these values uniquely determine the values of the other unknowns.

We say that a system of linear equations is *homogeneous* if $\mathbf{b} = \mathbf{0}$. The set of solutions of $A\mathbf{x} = \mathbf{0}$ is the kernel of A and it is always nonempty. By Theorem 8.1 we have that $\ker(\tilde{A}) = \ker(A)$.

Our interest in solving systems of linear equations is as the first step to more general non-linear results: the inverse function theorem and the implicit function theorem. We will consider two cases.

Case I: $m = n$ and the general system $A\mathbf{x} = \mathbf{b}$,
Case II: $n > m$ and the homogeneous system $A\mathbf{x} = \mathbf{0}$.

The first case is very simple. If $m = n$ then $A\mathbf{x} = \mathbf{b}$ has a unique solution if and only if A is invertible (see Exercise 5.12) and the solution is $\mathbf{x} = A^{-1}\mathbf{b}$.

In the second case we have more variables than equations and so, there are always columns without a pivotal 1. Reorder the variables as $\mathbf{x} = (\mathbf{y}, \mathbf{z})$ where the r-dimensional vector $\mathbf{y}$ corresponds to the columns with a pivotal 1 in $\tilde{A}$ and $\mathbf{z}$ are the remaining variables. With this reordering $\tilde{A}$ has the following block form

$$\tilde{A} \;=\; \begin{bmatrix} \mathbb{I}_r & B \\ \mathbf{0} & \mathbf{0}, \end{bmatrix} \tag{8.1}$$

where $\mathbb{I}_r$ is an $r \times r$ identity matrix and $B \in \mathbb{R}^{r\times(n-r)}$ for $r \leq m$. Zero rows mean that some of the equations were redundant and could be removed from the system. In practice we often get rid of redundant equations in advance. The following result gives a geometric meaning of the no zero row condition.

Proposition 8.1. *The matrix $\tilde{A}$ has no zero rows if and only if* $\mathrm{Im}(A) = \mathbb{R}^m$.

Proof. We have $\mathrm{Im}(A) = \mathbb{R}^m$ if and only if $A\mathbf{x} = \mathbf{b}$ has a solution for every $\mathbf{b} \in \mathbb{R}^m$. By Theorem 8.2 this means that for every $\mathbf{b}$ the corresponding $\tilde{\mathbf{b}}$ has no pivotal 1. Equivalently $\tilde{A}$ contains no zero rows. □

The following result will be later generalized by the Implicit Function Theorem.

Theorem 8.3. *Suppose that $n > m$ and* $\mathrm{Im}(A) = \mathbb{R}^m$. *Then (up to reordering the columns) $A = [A_{\mathbf{y}}|A_{\mathbf{z}}]$, where $A_{\mathbf{y}} \in \mathbb{R}^{m\times m}$ is invertible and the system of equations*

$$A\,\mathbf{x} \;=\; [A_{\mathbf{y}}|A_{\mathbf{z}}] \begin{bmatrix} \mathbf{y} \\ \mathbf{z} \end{bmatrix} \;=\; \mathbf{0}$$

is equivalent to $\mathbf{y} = -A_{\mathbf{y}}^{-1}A_{\mathbf{z}}\mathbf{z}$.

Proof. By Proposition 8.1, $\tilde{A}$ has no zero rows, and so, by (8.1), the system $A_{\mathbf{y}}\mathbf{y} + A_{\mathbf{z}}\mathbf{z} = \mathbf{0}$ is equivalent to $\mathbf{y} + \tilde{A}_{\mathbf{z}}\mathbf{z} = \mathbf{0}$. In other words, (taking $\mathbf{z} = \mathbf{0}$) $A_{\mathbf{y}}\mathbf{y} = \mathbf{0}$ if and only if $\mathbf{y} = 0$, or equivalently, $A_{\mathbf{y}}$ is invertible. But then $A_{\mathbf{y}}\mathbf{y} + A_{\mathbf{z}}\mathbf{z} = \mathbf{0}$ is equivalent to $\mathbf{y} + A_{\mathbf{y}}^{-1}A_{\mathbf{z}}\mathbf{z} = \mathbf{0}$. □

Geometrically this result says that if $\mathrm{Im}(A) = \mathbb{R}^m$ then $\ker(A)$ is an $n-m$ dimensional linear subspace parameterized by the non-pivotal variables $\mathbf{z}$.

8.2 Geometry of invertible matrices

Denote by Ω_n the set of all $n \times n$ invertible matrices. In this section we study the geometry of this set, or equivalently, the set of invertible linear mappings in $\mathcal{L}(\mathbb{R}^n, \mathbb{R}^n)$.

Proposition 8.2. *For any $C \in \mathbb{R}^{n\times n}$, $\|\mathbb{I}_n - C\| < 1$ implies that C is invertible. In particular, if A is invertible and $B \in \mathbb{R}^{n\times n}$ is such that $d(A,B) < \frac{1}{\|A^{-1}\|}$ then B is invertible.*

Proof. If $\|\mathbb{I}_n - C\| < 1$, then $\|(\mathbb{I}_n - C)\mathbf{x}\| < \|\mathbf{x}\|$ for all $\mathbf{x} \neq \mathbf{0}_n$. Suppose that $C\mathbf{x} = \mathbf{0}_n$ for some $\mathbf{x} \neq \mathbf{0}_n$. Then

$$\|(\mathbb{I}_n - C)\mathbf{x}\| = \|\mathbf{x}\| < \|\mathbf{x}\|,$$

which is a contradiction. This implies that $\ker(C) = \{\mathbf{0}\}$ and therefore C is invertible by Exercise 5.12. For the second statement note that B is invertible if and only if $A^{-1}B$ is invertible. We have

$$\|\mathbb{I}_n - A^{-1}B\| = \|A^{-1}(A - B)\| \leq \|A^{-1}\|\,\|A - B\|.$$

By the first part, if $\|A^{-1}\|\,\|A - B\| < 1$ then $A^{-1}B$ is invertible. □

This shows that every invertible matrix forms an interior point of Ω_n, which implies the following theorem.

Theorem 8.4. *Ω_n is an open subset of $\mathbb{R}^{n\times n}$.*

Remark 8.1. To prove that Ω_n we could also use the following argument: $A \in \Omega_n$ if and only if $\det(A) \neq 0$. The determinant is a polynomial in the entries of A and so a continuous function on $\mathbb{R}^{n\times n}$. Therefore, the set where the determinant vanishes is closed.

Proposition 8.3. *If $C \in \mathbb{R}^{n\times n}$ and $\|C\| = c < 1$ then $\mathbb{I}_n - C$ is invertible and*

$$(\mathbb{I}_n - C)^{-1} \;=\; \mathbb{I}_n + C + C^2 + C^3 + \ldots \;=\; \lim_{k\to\infty} \sum_{i=0}^{k} C^i.$$

Proof. $\mathbb{I}_n - C$ is invertible by Proposition 8.2. Denote $A_k = \mathbb{I}_n + C + C^2 + \cdots + C^k$. For $k > l$ we have

$$\begin{aligned}\|A_k - A_l\| = \|C^{l+1} + \cdots + C^k\| \le c^{l+1}(1 + \cdots + c^{k-l-1}) = \\ = c^{l+1}\tfrac{1-c^{k-l}}{1-c} \le c^{l+1}\tfrac{1}{1-c} \underset{l\to\infty}{\longrightarrow} 0.\end{aligned}$$

This implies that (A_k) forms a Cauchy sequence in $\mathbb{R}^{n\times n}$ with the norm given by the operator norm. The operator norm and the Frobenius norm are equivalent (Exercise 5.20) which together with Theorem 4.13 implies that (A_k) converges to A formally denoted by $\sum_{k=0}^{\infty} C^k$. It follows that the limit of $(\mathbb{I}_n - C)A_k$ also exists and is equal to $(\mathbb{I}_n - C)A$. On the other hand, since $(\mathbb{I}_n - C)A_k = \mathbb{I}_n - C^{k+1}$, this limit is $\mathbb{I}_n$ confirming that $\sum_{k=0}^{\infty} C^k$ is the inverse of $\mathbb{I}_n - C$. □

Theorem 8.5. *The map* $\mathbf{f} : \Omega_n \to \Omega_n$ *given by* $\mathbf{f}(A) = A^{-1}$ *is differentiable (and so also continuous). The derivative* $\mathrm{D}\mathbf{f}(\mathrm{A}) = \mathrm{T}$ *is* $T(H) = -A^{-1}HA^{-1}$.

Proof. For any $H \in \mathbb{R}^{n\times n}$ we have

$$\mathbf{f}(A+H) - \mathbf{f}(A) \;=\; A^{-1}[(\mathbb{I}_n + HA^{-1})^{-1} - \mathbb{I}_n].$$

We will be taking the limit $H \to \mathbf{0}_{n\times n}$ and so with no loss of generality we can assume $\|HA^{-1}\| < 1$, in which case $(\mathbb{I}_n + HA^{-1})$ is invertible with the inverse given by Proposition 8.3

$$(\mathbb{I}_n + HA^{-1})^{-1} \;=\; \mathbb{I}_n - HA^{-1} + (HA^{-1})^2 - (HA^{-1})^3 + \ldots$$

which implies that

$$\mathbf{f}(A+H) - \mathbf{f}(A) \;=\; A^{-1}[-HA^{-1} + (HA^{-1})^2 - (HA^{-1})^3 + \ldots].$$

Define $\mathbf{r}(H) = A^{-1}[(HA^{-1})^2 - (HA^{-1})^3 + \ldots]$ and then by standard properties of the operator norm

$$\|\mathbf{r}(H)\| \le \|A^{-1}\|^3\|H\|^2 \left\|\sum_{k=0}^{\infty}(-1)^k(HA^{-1})^k\right\|.$$

By showing the inequality for finite sums and passing to the limit we confirm that

$$\left\|\sum_{k=0}^{\infty}(-1)^k(HA^{-1})^k\right\| \;\le\; \sum_{k=0}^{\infty}\|HA^{-1}\|^k \;=\; \frac{1}{1-\|HA^{-1}\|}.$$

This gives that

$$\|\mathbf{r}(H)\| \le \frac{\|A^{-1}\|^3\|H\|^2}{1-\|HA^{-1}\|}$$

and so indeed $\mathbf{r}(H)/\|H\| \to 0$ as $H \to \mathbf{0}_{n\times n}$ proving that $T(H) = -A^{-1}HA^{-1}$ is the derivative. □

Remark 8.2. To prove differentiability of the matrix inverse we could use the following argument: The inverse A^{-1} is explicitly given as a rational function of the entries of A

$$A^{-1} = \tfrac{1}{\det(A)}\mathrm{adj}(A),$$

where $\mathrm{adj}(A)$ is the transpose of the matrix of cofactors of A. Differentiability then follows by Theorem 7.3(2).

8.3 Banach's fixed point theorem

For a function $f : X \to X$ we say that $x \in X$ is a **fixed point of** f if $f(x) = x$. In Chapter 12 we provide a more detailed discussion of fixed point theory. In this section we simply formulate a simple fixed point theorem that will be useful in the next section in the proof of the Inverse Function Theorem.

Definition 8.3. Let X be a metric space with metric d. If $\varphi : X \to X$ satisfies that for some $c < 1$

$$\mathrm{d}(\varphi(x), \varphi(y)) \le c \cdot \mathrm{d}(x, y) \qquad \text{for all } x, y \in X, \tag{8.2}$$

then φ is a **contraction** on X.

Exercise 8.1. Show that if $\varphi : X \to X$ is a contraction then it is continuous.

Theorem 8.6 (Banach's fixed point theorem). *If X is a complete metric space, and if $\varphi : X \to X$ is a contraction, then there exists one and only one $x \in X$ such that $\varphi(x) = x$.*

In other words, φ has a unique fixed point.

Proof. To show uniqueness, suppose $\varphi(x) = x$ and $\varphi(y) = y$. Then (8.2) gives $\mathrm{d}(x, y) \le c \cdot \mathrm{d}(x, y)$, which can only happen if $\mathrm{d}(x, y) = 0$. To show the existence of a fixed point, let $x_0 \in X$ arbitrarly, and define $\{x_n\}$ recursively by setting $x_{n+1} = \varphi(x_n)$. Choose $c < 1$ so that (8.2) holds. For $n \ge 1$ we then have

$$d(x_{n+1}, x_n) = d(\varphi(x_n), \varphi(x_{n-1})) \le c\, d(x_n, x_{n-1}),$$

which implies $d(x_{n+1}, x_n) \le c^n d(x_1, x_0)$. If $n < m$ we get

$$d(x_n, x_m) \le \sum_{i=n+1}^{m} d(x_i, x_{i-1}) \le (c^n + c^{n+1} + \cdots + c^m) d(x_1, x_0) \le \tfrac{c^n}{1-c} d(x_1, x_0).$$

Thus $\{x_n\}$ is a Cauchy sequence. Since X is complete, $\lim x_n = x$ for some $x \in X$. Since φ is a contraction, φ is continuous on X. Hence $\varphi(x) = \lim_{n\to\infty} \varphi(x_n) = \lim_{n\to\infty} x_{n+1} = x$. □

We conclude this section with some exercises.

Exercise 8.2. Let X be a complete metric space and consider a map $f : X \to X$ such that $f(X) = X$ and for some $c > 1$ we have that

$$\mathrm{d}(f(x), f(y)) \geq c \cdot \mathrm{d}(x, y) \qquad \text{for all } x, y \in X.$$

Prove that f has a unique fixed point.

Exercise 8.3. Suppose that $f : \mathbb{R} \to \mathbb{R}$ is a contraction. Show that the equation $x = f(x) + a$ has a unique solution.

Exercise 8.4. For any $n \in \mathbb{N}$, an $n \times n$ matrix A is said to be stochastic if $a_{ij} \geq 0$ for all i and j, and $\sum_{j=1}^{n} a_{ij} = 1$ for all i. Prove that for every strictly positive $n \times n$ stochastic matrix A ($a_{ij} > 0$), there is a strictly positive vector $x \in \mathbb{R}^n$ such that $Ax = x$ and $\sum x_i = 1$. Show also that there is a strictly positive vector $x \in \mathbb{R}^n$ such that $A^T x = x$ and $\sum x_i = 1$.

Exercise 8.5. Prove that there is a unique $f \in C[0, 1]$ such that $f \geq 0$ and

$$f(x) \;=\; 1 + \tfrac{3}{4} \ln\left(1 + \int_0^x f(t)\mathrm{d}t\right).$$

Hint: Use Exercise 4.4 and the fact that for all $0 \leq a \leq b$

$$\ln\left(\tfrac{1+b}{1+a}\right) \leq b - a.$$

8.4 Inverse function theorem

In Section 8.1 we completely analysed systems of linear equations. We know when there is a unique solution, if there is no, we know which variables depend on the others. Our tools for answering similar questions for nonlinear systems are the implicit function theorem, and its special case, the inverse function theorem. Both of them rely on the fact that continuously differentiable functions locally behave like their derivatives.

We start with the Inverse Function Theorem. If $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^n$ is a linear mapping then $\mathrm{Df}(\mathbf{x}) = \mathrm{f}(\mathbf{x})$ for all $\mathbf{x}$ and so a linear mapping is invertible if and only if its derivative is. A similar theorem holds locally for more general function.

Theorem 8.7 (Inverse function theorem: short version). *If* $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^n$ *is continuously differentiable in a neighborhood of* $\mathbf{a}$ *and* $\mathrm{D}\mathbf{f}(\mathbf{a})$ *is invertible, then* $\mathbf{f}$ *is locally invertible, with differentiable inverse, in some neighborhood of the point* $\mathbf{b} = \mathbf{f}(\mathbf{a})$. *The derivative of the inverse is the inverse of* $\mathrm{D}\mathbf{f}(\mathbf{a})$.

Example 8.1. Where is the function $\mathbf{f}(x, y) = (\sin(x + y), x^2 - y^2)$ locally invertible? The function is continuously differentiable. The Jacobian matrix is

$$\mathrm{J}\mathbf{f}(\mathrm{x}, \mathrm{y}) \;=\; \begin{bmatrix} \cos(x+y) & \cos(x+y) \\ 2x & -2y \end{bmatrix}.$$

This is invertible as long as $\cos(x + y) \neq 0$ and $x + y \neq 0$.

Remark 8.3. If the derivative of a function $\mathbf{f}$ is invertible at some point $\mathbf{x}_0$ then $\mathbf{f}$ is locally invertible in a neighborhood of $\mathbf{f}(\mathbf{x}_0)$; but it is not true that if the derivative is invertible everywhere, the function is invertible. Consider the function $\mathbf{f} : \mathbb{R}^2 \to \mathbb{R}^2$ given by $\mathbf{f}(t, \theta) = (e^t \cos\theta, e^t \sin\theta)$; the derivative is invertible everywhere, since $\det(\mathrm{J}\mathbf{f}) = \mathrm{e}^{\mathrm{t}} \neq 0$, but $\mathbf{f}$ is not invertible, since it sends (t, θ) and $(t, \theta + 2\pi)$ to the same point.

We now formulate a more concrete version of the inverse function theorem.

Theorem 8.8 (The inverse function theorem). *Suppose that* $\mathbf{f} : E \to \mathbb{R}^n$, $\mathbf{f} \in \mathcal{C}^1(E)$, *where* $E \subset \mathbb{R}^n$ *is a neighborhood of* $\mathbf{a}$. *Set* $\mathbf{b} = \mathbf{f}(\mathbf{a})$. *If the derivative* $\mathrm{D}\mathbf{f}(\mathbf{a})$ *is invertible, then*

(a) there exist open sets U *and* V *in* $\mathbb{R}^n$ *such that* $\mathbf{a} \in U$, $\mathbf{b} \in V$, $\mathbf{f}$ *is one-to-one on* U *and* $\mathbf{f}(U) = V$.
(b) If $\mathbf{g}$ *is the inverse function, defined in* V *by*

$$\mathbf{g}(\mathbf{f}(\mathbf{x})) = \mathbf{x} \qquad (\mathbf{x} \in U),$$

then $\mathbf{g} \in \mathcal{C}^1(V)$ *and*

$$\mathrm{D}\mathbf{g}(\mathbf{y}) \;=\; (\mathrm{D}\mathbf{f}(\mathbf{g}(\mathbf{y})))^{-1}. \tag{8.3}$$

Proof. (a) Let $T = \mathrm{D}\mathbf{f}(\mathbf{a})$ and $\lambda = \frac{1}{2\|T^{-1}\|}$. By Theorem 7.7, $\mathbf{x} \to \mathrm{D}\mathbf{f}(\mathbf{x})$ is continuous at $\mathbf{a}$, and so there exists a neighborhood $U \subset E$ of $\mathbf{a}$ such that $\|\mathrm{D}\mathbf{f}(\mathbf{x}) - \mathrm{T}\| < \lambda$ if $\mathbf{x} \in U$. Also we note that $\|\mathrm{D}\mathbf{f}(\mathbf{x}) - \mathrm{T}\| \leq \lambda$ if $\mathbf{x} \in \overline{U}$.

Claim 1: $\mathbf{f}$ is one-to-one on $\overline{U}$. For each $\mathbf{y} \in \mathbb{R}^n$ define $\varphi_{\mathbf{y}} : E \to \mathbb{R}^n$ by

$$\varphi_{\mathbf{y}}(\mathbf{x}) \;=\; \mathbf{x} + T^{-1}(\mathbf{y} - \mathbf{f}(\mathbf{x})).$$

Observe that $\mathbf{y} = \mathbf{f}(\mathbf{x})$ if and only if $\varphi_{\mathbf{y}}(\mathbf{x}) = \mathbf{x}$, that is, $\mathbf{x}$ is a fixed point of $\varphi_{\mathbf{y}}$. Clearly

$$\mathrm{D}\varphi_{\mathbf{y}}(\mathbf{x}) = \mathrm{Id} - \mathrm{T}^{-1} \circ \mathrm{D}\mathbf{f}(\mathbf{x}) = \mathrm{T}^{-1} \circ (\mathrm{T} - \mathrm{D}\mathbf{f}(\mathbf{x})),$$

where Id denotes the identity transformation $\mathrm{Id}(\mathbf{x}) = \mathbf{x}$. Thus, if $\mathbf{x} \in \overline{U}$,

$$\|\mathrm{D}\varphi_{\mathbf{y}}(\mathbf{x})\| \leq \|\mathrm{T}^{-1}\| \cdot \|\mathrm{T} - \mathrm{D}\mathbf{f}(\mathbf{x})\| \leq \tfrac{1}{2}.$$

By Theorem 7.6,

$$\|\varphi_{\mathbf{y}}(\mathbf{x}_1) - \varphi_{\mathbf{y}}(\mathbf{x}_2)\| \leq \tfrac{1}{2}\|\mathbf{x}_1 - \mathbf{x}_2\| \qquad \text{for all } \mathbf{x}_1, \mathbf{x}_2 \in \overline{U}.$$

Therefore, $\varphi_{\mathbf{y}}$ is a contraction, so it has at most one fixed point in $\overline{U}$ (may have no fixed points because $\varphi_{\mathbf{y}}(\mathbf{x})$ may not lie in $\overline{U}$), so we have $\mathbf{f}(\mathbf{x}) = \mathbf{y}$ for at most one $\mathbf{x} \in \overline{U}$. Thus, $\mathbf{f}$ is one-to-one in $\overline{U}$.

Claim 2: The set $V = \mathbf{f}(U)$ is open. Since $\mathbf{f}$ is continuous bijection from the compact set $\overline{U}$ to $\mathbf{f}(\overline{U})$, by Theorem 4.12, the local inverse $\mathbf{g}$ of $\mathbf{f}$ ($\mathbf{g} : \mathbf{f}(\overline{U}) \to \overline{U}$) is also continuous. By Theorem 3.8, $\mathbf{f}(U)$ must be open.

(b) We show that the local inverse $\mathbf{g}$ of $\mathbf{f}$ ($\mathbf{g} : V \to U$) is continuously differentiable.

Claim 3: $\mathbf{g}$ is differentiable in V. Let $\mathbf{y} \in V$, $\mathbf{y} + k \in V$. Then there exists $\mathbf{x} \in U$ and $\mathbf{x} + \mathbf{h} \in U$ so that $\mathbf{y} = \mathbf{f}(\mathbf{x})$, $\mathbf{y} + \mathbf{k} = \mathbf{f}(\mathbf{x} + \mathbf{h})$. Then

$$\begin{aligned}\varphi_{\mathbf{y}}(\mathbf{x} + \mathbf{h}) - \varphi_{\mathbf{y}}(\mathbf{x}) &= \mathbf{x} + \mathbf{h} + T^{-1}(\mathbf{y} - \mathbf{f}(\mathbf{x} + \mathbf{h})) - \mathbf{x} - T^{-1}(\mathbf{y} - \mathbf{f}(\mathbf{x})) \\ &= \mathbf{h} + T^{-1}(\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{x} + \mathbf{h})) = \mathbf{h} - T^{-1}(\mathbf{k})\end{aligned}$$

By the contraction property of $\varphi_{\mathbf{y}}$, $\|\mathbf{h} - T^{-1}(\mathbf{k})\| \leq \frac{1}{2}\|\mathbf{h}\|$ and so $\|T^{-1}(\mathbf{k})\| \geq \frac{1}{2}\|\mathbf{h}\|$. This is simply because the neighbourhoods of radius $\frac{1}{2}\|\mathbf{h}\|$ around $\mathbf{0}$ and around $\mathbf{h}$ have no points in common. This implies that

$$\|\mathbf{h}\| \;\leq\; 2\|T^{-1}(\mathbf{k})\| \;\leq\; 2\|T^{-1}\|\,\|\mathbf{k}\| \;=\; \tfrac{1}{\lambda}\|\mathbf{k}\|.$$

(In particular, $\mathbf{k} \to \mathbf{0}$ then $\mathbf{h} \to \mathbf{0}$) Since $\mathbf{f}$ is continuously differentiable and $\mathbf{x} \in U$, $S = \mathrm{D}\mathbf{f}(\mathbf{x})$ is invertible by Proposition 8.2 (because T is and $\lambda = \frac{1}{2\|T^{-1}\|}$). Now

$$\mathbf{g}(\mathbf{y} + \mathbf{k}) - \mathbf{g}(\mathbf{y}) - S^{-1}(\mathbf{k}) = \mathbf{h} - S^{-1}(\mathbf{k}) = -S^{-1}(\mathbf{f}(\mathbf{x} + \mathbf{h}) - \mathbf{f}(\mathbf{x}) - S(\mathbf{h})).$$

Therefore,

$$\frac{\|\mathbf{g}(\mathbf{y}+\mathbf{k}) - \mathbf{g}(\mathbf{y}) - S^{-1}(\mathbf{k})\|}{\|\mathbf{k}\|} \leq \frac{\|S^{-1}\|}{\lambda} \frac{\|\mathbf{f}(\mathbf{x}+\mathbf{h}) - \mathbf{f}(\mathbf{x}) - S(\mathbf{h})\|}{\|\mathbf{h}\|}.$$

As $\mathbf{k} \to \mathbf{0}$ also $\mathbf{h} \to \mathbf{0}$ and the right hand side above tends to 0. Thus $\mathbf{g}$ is differentiable at $\mathbf{y}$ and $\mathrm{D}\mathbf{g}(\mathbf{y}) = \mathrm{S}^{-1}$ and for $\mathbf{y} \in V$ equation (8.3) holds. Claim 4: $\mathbf{g}$ is continuously differentiable on V. Note that $\mathbf{g}$ is continuous on V (since differentiable) and $\mathrm{D}\mathbf{f}$ is continuous on U. By Theorem 8.5 the inversion in (8.3) is a continuous function on $L(\mathbb{R}^n)$, so indeed $\mathrm{D}\mathbf{g}$ is continuous (as a composition of continuous functions). □

Although finding the inverse function $\mathbf{g}$ is typically very hard, its derivative at $\mathbf{b} = \mathbf{f}(\mathbf{a})$ is very easy to compute using the chain rule. The equation $\mathbf{g}(\mathbf{f}(\mathbf{x})) = \mathbf{x}$ gives that $\mathrm{D}\mathbf{g}(\mathbf{f}(\mathbf{x})) \circ \mathrm{D}\mathbf{f}(\mathbf{x})$ is equal to the identity transformation $T(\mathbf{x}) = \mathbf{x}$ represented by the identity matrix $\mathbb{I}_n$. Plugging $\mathbf{y} = \mathbf{f}(\mathbf{x})$ and $\mathbf{x} = \mathbf{g}(\mathbf{y})$ gives equation (8.3). This means that the Jacobian matrix of $\mathbf{g}(\mathbf{y})$ can be obtained by inverting the Jacobian matrix of $\mathbf{f}(\mathbf{x})$ computed at $\mathbf{x} = \mathbf{g}(\mathbf{y})$. This gives

$$\mathrm{J}\mathbf{g}(\mathbf{y}) \;=\; (\mathrm{J}\mathbf{f}(\mathbf{g}(\mathbf{y})))^{-1}.$$

In particular, we can very easily compute the linear approximation to $\mathbf{g}$.

Example 8.2. Let $\mathbf{f} : \mathbb{R}^2 \to \mathbb{R}^2$ be given by $(x, y) \mapsto (x^3 - 2xy^2, x + y)$. Let $\mathbf{a} = (1, -1)$. We have

$$\mathrm{J}\mathbf{f}(\mathrm{x}, \mathrm{y}) = \begin{bmatrix} 3x^2 - 2y^2 & -4xy \\ 1 & 1 \end{bmatrix}, \qquad \mathrm{J}\mathbf{f}(1, -1) = \begin{bmatrix} 1 & 4 \\ 1 & 1 \end{bmatrix}.$$

We have

$$(\mathrm{J}\mathbf{f}(1, -1))^{-1} = \tfrac{1}{3} \begin{bmatrix} -1 & 4 \\ 1 & -1 \end{bmatrix}.$$

By the inverse function theorem, $\mathbf{f}$ is locally invertible at $(1, -1)$ and the affine approximation to $\mathbf{g} = \mathbf{f}^{-1}$ near $\mathbf{f}(1, -1) = (-1, 0)$ is

$$\mathbf{g}(-1 + h, 0 + k) \approx \begin{bmatrix} 1 \\ -1 \end{bmatrix} + \tfrac{1}{3} \begin{bmatrix} -1 & 4 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} h \\ k \end{bmatrix} = \tfrac{1}{3} \begin{bmatrix} 3 - h + 4k \\ -3 + h - k \end{bmatrix}.$$

8.5 Implicit function theorem

The Inverse Function Theorem deals with the case where we have n equations in n unknowns. What if we have more unknowns than equations? There is then no inverse function, but often we can express some unknown in terms of others.

Example 8.3. The equation $x^2 + y^2 + z^2 - 1 = 0$ expresses z as an implicit function of (x, y) near $(0, 0, 1)$. This implicit function can be made explicit: $z = \sqrt{1 - x^2 - y^2}$.

The Implicit Function Theorem tells us under what conditions an implicit function exists. It generalizes its linear version in Theorem 8.3.

Theorem 8.9 (Implicit Function Theorem: short version). *Let* $\mathbf{F} : E \to \mathbb{R}^m$*, for* $E \subset \mathbb{R}^n$ *open, satisfy* $F \in \mathcal{C}^1(E)$*. Suppose that* $\mathbf{c} \in E$ *satisfies* $\mathbf{F}(\mathbf{c}) = \mathbf{0}_m$ *and* $\mathrm{Im}(\mathrm{D}\mathbf{F}(\mathbf{c})) = \mathbb{R}^{\mathrm{m}}$*. Then the system of linear equations* $(\mathrm{D}\mathbf{F}(\mathbf{c}))(\mathbf{x}) = \mathbf{0}$ *has* m *pivotal variables* $\mathbf{y}$ *and* $n - m$ *nonpivotal variables* $\mathbf{z}$*, and there exists a neighborhood of* $\mathbf{c}$ *in which* $\mathbf{F}(\mathbf{x}) = \mathbf{0}$ *implicitly defines* $\mathbf{y}$ *as a function* $\mathbf{g}$ *of* $\mathbf{z}$*.*

The function $\mathbf{g}$ is the implicit function.

Like the Inverse Function Theorem, the Implicit Function Theorem states that locally, the mapping behaves like its derivative (i.e. like its linearization). Since $\mathbf{F}$ goes from a subset of $\mathbb{R}^n$ to $\mathbb{R}^m$, its derivative lies in $\mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$. Theorem 8.3 is then simply the linear version of the implicit function theorem, with $g(\mathbf{z}) = -A_{\mathbf{y}}^{-1} A_{\mathbf{z}} \mathbf{z}$.

As with the inverse function theorem, we will prove a more concrete version of the Implicit Function Theorem.

Theorem 8.10 (Implicit function theorem). *Let $\mathbf{F}(\mathbf{y}, \mathbf{z})$ be a $\mathcal{C}^1$ mapping of an open set $E \subset \mathbb{R}^n = \mathbb{R}^m \times \mathbb{R}^{n-m}$ into $\mathbb{R}^m$ such that $\mathbf{F}(\mathbf{a}, \mathbf{b}) = \mathbf{0}_n$ for some $(\mathbf{a}, \mathbf{b}) \in E$. Let $A = \mathrm{J}\mathbf{F}(\mathbf{a}, \mathbf{b})$, where $A = [A_{\mathbf{y}} | A_{\mathbf{z}}]$. If $A_{\mathbf{y}}$ is invertible, then there exist open sets $U \subset \mathbb{R}^n$, $W \subset \mathbb{R}^{n-m}$ with $(\mathbf{a}, \mathbf{b}) \in U$ and $\mathbf{b} \in W$ such that*

(i) To every $\mathbf{z} \in W$ there corresponds a unique $\mathbf{y}$ such that $(\mathbf{y}, \mathbf{z}) \in U$ and $\mathbf{F}(\mathbf{y}, \mathbf{z}) = \mathbf{0}$.

(ii) If this $\mathbf{y}$ is defined to be $\mathbf{g}(\mathbf{z})$, then $\mathbf{g}$ is a $\mathcal{C}^1$ mapping of W into $\mathbb{R}^m$, $\mathbf{g}(\mathbf{b}) = \mathbf{a}$,

$$\mathbf{F}(\mathbf{g}(\mathbf{z}), \mathbf{z}) = \mathbf{0} \qquad (\mathbf{y} \in W) \tag{8.4}$$

and

$$\mathrm{J}\mathbf{g}(\mathbf{b}) = -\mathrm{A}_{\mathbf{y}}^{-1} \mathrm{A}_{\mathbf{z}}. \tag{8.5}$$

The function $\mathbf{g}$ is "implicitly" defined by (8.4).

Proof. Define $\hat{\mathbf{F}}$ by $\hat{\mathbf{F}}(\mathbf{y}, \mathbf{z}) = (\mathbf{F}(\mathbf{y}, \mathbf{z}), \mathbf{z})$ for $(\mathbf{y}, \mathbf{z}) \in E$. Then $\hat{\mathbf{F}}$ is a $\mathcal{C}^1$ mapping, $\hat{\mathbf{F}} : E \to \mathbb{R}^n$. The linear mapping $\mathrm{D}\hat{\mathbf{F}}(\mathbf{a}, \mathbf{b})$ is invertible as its Jacobian matrix is

$$\mathrm{J}\hat{\mathbf{F}}(\mathbf{y}, \mathbf{z}) = \begin{bmatrix} A_{\mathbf{y}} & A_{\mathbf{z}} \\ \mathbf{0} & \mathbb{I}_{n-m} \end{bmatrix} \qquad \text{and} \qquad \begin{bmatrix} A_{\mathbf{y}} & A_{\mathbf{z}} \\ \mathbf{0} & \mathbb{I}_{n-m} \end{bmatrix}^{-1} = \begin{bmatrix} A_{\mathbf{y}}^{-1} & -A_{\mathbf{y}}^{-1} A_{\mathbf{z}} \\ \mathbf{0} & \mathbb{I}_{n-m} \end{bmatrix}.$$

We can therefore apply the Inverse Function Theorem to $\hat{\mathbf{F}}$: there exists U, V open sets in $\mathbb{R}^n$ with $(\mathbf{a}, \mathbf{b}) \in U$, $\hat{\mathbf{F}}(\mathbf{a}, \mathbf{b}) = (\mathbf{0}, \mathbf{b}) \in V$ such that $\hat{\mathbf{F}}$ is one-to-one mapping of U onto V. Now restrict V to points $(\mathbf{0}, \mathbf{z})$:

$$W = \{\mathbf{z} \in \mathbb{R}^{n-m} : (\mathbf{0}, \mathbf{z}) \in V\} \qquad (\text{we have } \mathbf{b} \in W).$$

This set is open since V is open (each $(\mathbf{0}, \mathbf{z})$ is interior in V and projection of an open ball is an open ball). If $\mathbf{z} \in W$ then $(\mathbf{0}, \mathbf{z}) = \hat{\mathbf{F}}(\mathbf{y}, \mathbf{z})$ for some $(\mathbf{y}, \mathbf{z}) \in U$ and $\mathbf{F}(\mathbf{y}, \mathbf{z}) = \mathbf{0}$ by definition of $\hat{\mathbf{F}}$. To show that to each $\mathbf{z} \in W$ there corresponds a unique $\mathbf{y}$ such that $(\mathbf{y}, \mathbf{z}) \in U$, suppose that $(\mathbf{y}', \mathbf{z}) \in U$ and $\mathbf{F}(\mathbf{y}', \mathbf{z}) = \mathbf{0}$. Then

$$\hat{\mathbf{F}}(\mathbf{y}', \mathbf{z}) = (\mathbf{F}(\mathbf{y}', \mathbf{z}), \mathbf{z}) = (\mathbf{F}(\mathbf{y}, \mathbf{z}), \mathbf{z}) = \hat{\mathbf{F}}(\mathbf{y}, \mathbf{z}).$$

Since $\hat{\mathbf{F}}$ is one-to-one in U it implies that $\mathbf{y} = \mathbf{y}'$ so $\mathbf{y}$ must be unique.

For the second part of the theorem let $\mathbf{g}(\mathbf{z})$, for $\mathbf{z} \in W$, such that $(\mathbf{g}(\mathbf{z}), \mathbf{z}) \in U$ and $\mathbf{F}(\mathbf{g}(\mathbf{z}), \mathbf{z}) = 0$. Then $\hat{\mathbf{F}}(\mathbf{g}(\mathbf{z}), \mathbf{z}) = (\mathbf{0}, \mathbf{z})$ for $\mathbf{z} \in W$. If $\hat{\mathbf{G}}$ is a mapping of V onto U that inverts $\hat{\mathbf{F}}$, then $\hat{\mathbf{G}}$ is a $\mathcal{C}^1$ mapping by the Inverse Function Theorem and $(\mathbf{g}(\mathbf{z}), \mathbf{z}) = \hat{\mathbf{G}}(\mathbf{0}, \mathbf{z})$ (because $\hat{\mathbf{F}}(\mathbf{g}(\mathbf{z}), \mathbf{z}) = (\mathbf{0}, \mathbf{z})$). Therefore, $\mathbf{g}$ is also a $\mathcal{C}^1$ mapping.

Finally, to compute $\mathrm{D}\mathbf{g}(\mathbf{b})$, put $(\mathbf{g}(\mathbf{z}), \mathbf{z}) = \Phi(\mathbf{z})$. Then

$$\mathrm{J}\Phi(\mathbf{z}) = \begin{bmatrix} \mathrm{J}\mathbf{g}(\mathbf{z}) \\ \mathbb{I}_m \end{bmatrix} \qquad \mathbf{z} \in \mathrm{W}.$$

Since $\mathbf{F}(\Phi(\mathbf{z})) = \mathbf{0}$ in W, the chain rule shows that

$$\mathrm{J}\mathbf{F}(\Phi(\mathbf{z})) \cdot \mathrm{J}\Phi(\mathbf{z}) = \mathbf{0}_{\mathrm{m}\times(\mathrm{n}-\mathrm{m})}.$$

When $\mathbf{z} = \mathbf{b}$ then $\Phi(\mathbf{z}) = (\mathbf{a}, \mathbf{b})$ and $\mathrm{J}\mathbf{F}(\Phi(\mathbf{b})) = \mathrm{A}$. Thus $A \cdot \mathrm{J}\Phi(\mathbf{b})$ is equal to the zero matrix $\mathbf{0}_{m\times(n-m)}$ so

$$[A_{\mathbf{y}} \mid A_{\mathbf{z}}] \cdot \begin{bmatrix} \mathrm{J}\mathbf{g}(\mathbf{z}) \\ \mathbb{I}_{n-m} \end{bmatrix} \;=\; A_{\mathbf{y}}\mathrm{J}\mathbf{g}(\mathbf{z}) + \mathrm{A}_{\mathbf{z}}.$$

It then follows that $\mathrm{J}\mathbf{g}(\mathbf{z}) = -\mathrm{A}_{\mathbf{y}}^{-1}\mathrm{A}_{\mathbf{z}}$, which completes the proof. □

Note that $\mathbf{g}$ is typically hard to find but its derivative is explicit.

Remark 8.4. The Implicit Function Theorem says that if $\mathrm{D}\mathbf{F}(\mathbf{c})$ is onto, the set $\mathbf{F}(\mathbf{x}) = \mathbf{0}$ locally admits a parameterization $\Phi(\mathbf{z}) = (\mathbf{g}(\mathbf{z}), \mathbf{z})$ with parameters $\mathbf{z}$. We will use this fact in the constraint optimization, which we discuss in the following chapter.

Constrained optimization is one of the important applications of the Implicit Function Theorem in Economics. Another such example is the analysis of comparative statics, which we now illustrate with an example.

8.6 An application in economics

Consider a firm that produces a good y using a n inputs $\mathbf{x} = (x_1, \ldots, x_n)$. The firm sells the output and acquires the inputs in competitive markets: The market price of y is p, which we assume fixed throughout, and the cost of each unit of $\mathbf{x}$ is $\mathbf{w} = (w_1, \ldots, w_n)$. The firm's technology is given by $f : \mathbb{R}^n_+ \to \mathbb{R}_+$ given by

$$f(\mathbf{x}) = \prod_{i=1} x_i^{a_i} =: \mathbf{x}^{\mathbf{a}}, \qquad (a_1, \ldots, a_n > 0,\ a_1 + \cdots + a_n < 1).$$

Its profits take the form

$$\pi(\mathbf{x};\mathbf{w}) \;=\; pf(\mathbf{x}) - \mathbf{w}^T\mathbf{x}.$$

The firm selects $\mathbf{x}$ in order to maximize profits. Our question is how its choice of x_i is affected by a change in w_i. Notice that w_1 affects the choice of x_1 not only in a direct way but also indirectly through its effect on other entries of $\mathbf{x}$.

Define the gradient of π with respect to $\mathbf{x}$ as

$$F : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n, \qquad F(\mathbf{x};\mathbf{w}) = \nabla\pi(\mathbf{x};\mathbf{w}).$$

The i-th entry of the gradient is

$$F_i(\mathbf{x};\mathbf{w}) \;=\; pa_i\mathbf{x}^{\mathbf{a}-\mathbf{e}_i} - w_i.$$

Given the price vector $\mathbf{w}^*$, the optimal point $\mathbf{x}^*$ (suppose it has strictly positive entries) satisfies the first order conditions $F(\mathbf{x}^*;\mathbf{w}^*) = \mathbf{0}$. Our aim is to first show that locally around $\mathbf{x}^*$ the vector $\mathbf{x}$ can be expressed as a function of $\mathbf{w}$. The Jacobian of $\mathbf{F}$ with respect to $\mathbf{x}$ is

$$\mathrm{J}_{\mathbf{x}}\mathbf{F}(\mathbf{x};\mathbf{w}) \;=\; \left(\left[\frac{\mathrm{a_i a_j}}{\mathrm{x_i x_j}}\right]_{\mathrm{i,j=1,\ldots,n}} - \mathrm{diag}\left(\frac{\mathrm{a_i}}{\mathrm{x_i^2}}\right)\right)\cdot\mathbf{x}^{\mathbf{a}}. \tag{8.6}$$

It is useful to introduce $\mathbf{u}$ to be a vector in $\mathbb{R}^n$ such that $u_i = \frac{a_i}{x_i}$ and D to be a diagonal matrix such that $D_{ii} = \frac{a_i}{x_i^2}$. Then $\mathrm{J}_{\mathbf{x}}\mathbf{F}(\mathbf{x};\mathbf{w}) = (\mathbf{u}\mathbf{u}^{\mathrm{T}} - \mathrm{D})\cdot\mathbf{x}^{\mathbf{a}}$.

Lemma 8.1. *The matrix* $\mathrm{J}_{\mathbf{x}}\mathbf{F}(\mathbf{x};\mathbf{w})$ *is negative definite for all* $\mathbf{x} > 0$.

Proof. It is enough to show that $D - \mathbf{u}\mathbf{u}^T$ is positive definite. We will show this by proving that all principal submatrices of this matrix have a strictly positive determinant. Since every principal submatrix has essentially the same form as the complete matrix we show only the complete case. We have

$$\det(D - \mathbf{u}\mathbf{u}^T) = \det(D)\det(I_n - D^{-1/2}\mathbf{u}\mathbf{u}^T D^{-1/2}).$$

By the Sylvester's determinant identity

$$\det(I_n - D^{-1/2}\mathbf{u}\mathbf{u}^T D^{-1/2}) \;=\; (1 - \mathbf{u}^T D^{-1}\mathbf{u}) \;=\; 1 - \sum_{i=1}^{n} a_i \;>\; 0.$$

□

We also note that

$$\mathrm{J}_{\mathbf{w}}\mathbf{F}(\mathbf{x},\mathbf{w}) \;=\; -\mathbb{I}_{\mathrm{n}}$$

Now applying the Implicit Function Theorem we get that locally around $(\mathbf{x}^*,\mathbf{w}^*)$ the vector $\mathbf{x}$ is a function of $\mathbf{w}$. The derivative of the implicit function $\mathbf{x} = \mathbf{g}(\mathbf{w})$ is

$$\mathrm{J}\mathbf{g}(\mathbf{w}^*) \;=\; -(\mathrm{J}_{\mathbf{x}}\mathbf{F}(\mathbf{x}^*,\mathbf{w}^*))^{-1}\mathrm{J}_{\mathbf{w}}\mathbf{F}(\mathbf{x}^*,\mathbf{w}^*) \;=\; (\mathrm{J}_{\mathbf{x}}\mathbf{F}(\mathbf{x}^*,\mathbf{w}^*))^{-1}.$$

So for example, to get local dependence of x_i with respect to w_i we again use the fact that $\mathrm{J}_{\mathbf{x}}\mathbf{F}(\mathbf{x}^*,\mathbf{w}^*)$ is negative definite to conclude that this dependence is negative. Finally, we note that the inverse of $\mathrm{J}_{\mathbf{x}}\mathbf{F}(\mathbf{x}^*,\mathbf{w}^*)$ can be computed using the Sherman-Morrison formula again exploiting the fact that this Jacobian is a sum of a diagonal matrix and a rank-one matrix.

Exercise 8.6. Reformulate and simplify the above example so that it uses only the Inverse Function Theorem.

Part II
Optional topics

Chapter 9
Optimization (1 lecture)

9.1 Second order derivatives

Let $f : E \to \mathbb{R}$ ($E \subset \mathbb{R}^n$ open) have partial derivatives $\mathrm{D_1f}, \ldots, \mathrm{D_nf}$. If these are differentiable, we define the second-order partial derivative of f by

$$D_{ij}f = \mathrm{D_iD_jf} \qquad \text{for i, j} = 1, \ldots, \mathrm{n}.$$

If these are continuous on E, we write $f \in \mathcal{C}^2(E)$.

The following is a second-order generalization of the Mean Value Theorem. Higher-order version can be also easily formulated.

Theorem 9.1. *Let $f : E \to \mathbb{R}$ where $E \subset \mathbb{R}^2$ is open. Assume $\mathrm{D_1f}, \mathrm{D_2f}, \mathrm{D_{21}f}$ exist in E. Let $Q \subset E$ be a 2-cell $[a, a+h] \times [b, b+k]$. Set*

$$\Delta(f, Q) = f(a+h, b+k) - f(a+h, b) - f(a, b+k) + f(a, b).$$

Then there exists $(x, y) \in Q^\circ$ such that

$$\Delta(f, Q) = h \cdot k \cdot (\mathrm{D_{21}f})(\mathrm{x}, \mathrm{y}).$$

Proof. Let $u(t) = f(t, b+k) - f(t, b)$, then $u'(t) = \mathrm{D_1f(t, b+k) - D_1f(t, b)}$. By the Mean Value Theorem for u and for $\mathrm{D_1f}$, there exists $x \in (a, a+h)$ and $y \in (b, b+k)$ such that

$$\Delta(f, Q) = u(a+h) - u(a) = hu'(x) = h[\mathrm{D_1f(x, b+k) - D_1f(x, b)}] = \mathrm{hk(D_{21}f)(x, y)}.$$

□

Theorem 9.2. *Let $f : E \to \mathbb{R}$ where $E \subset \mathbb{R}^2$ is open. Assume that $\mathrm{D_1f}$, $\mathrm{D_2f}$, $\mathrm{D_{21}f}$ exist in E and $D_{21}f$ is continuous at some $(a, b) \in E$. Then $\mathrm{D_{12}f}$ exists at (a, b) and*

$$D_{12}f(a, b) \;=\; \mathrm{D_{21}f(a, b)}.$$

Proof. Set $T = \mathrm{D}_{21}\mathrm{f}(\mathrm{a},\mathrm{b})$ and $\epsilon > 0$. If h and k are sufficiently small and Q is as in the Theorem 9.1, then for all $(x, y) \in Q$

$$|T - \mathrm{D}_{21}\mathrm{f}(\mathrm{x},\mathrm{y})| < \epsilon$$

by continuity of $\mathrm{D}_{21}\mathrm{f}$. Then, in particular, by Theorem 9.1 $|\frac{\Delta(f,Q)}{hk} - T| < \epsilon$. Fix h and let $k \to 0$, then we get

$$\left|\frac{\mathrm{D}_2\mathrm{f}(\mathrm{a+h},\mathrm{b}) - \mathrm{D}_2\mathrm{f}(\mathrm{a},\mathrm{b})}{h} - T\right| < \epsilon.$$

Since ϵ was arbitrary, it follows that $\mathrm{D}_{12}\mathrm{f}(\mathrm{a},\mathrm{b}) = \mathrm{T}$. □

Corollary 9.1. *If $f \in \mathcal{C}^2(E)$, $E \subset \mathbb{R}^n$, then $\mathrm{D}_{\mathrm{ij}}\mathrm{f} = \mathrm{D}_{\mathrm{ji}}\mathrm{f}$ for all $i, j = 1, \ldots, n$.*

In the next example we will consider a function for which the symmetry of the second derivatives fails to hold.

Example 9.1. Let $f : \mathbb{R}^2 \to \mathbb{R}$ be defined as

$$f(x,y) = \begin{cases} \frac{xy(x^2-y^2)}{x^2+y^2} & \text{if } (x,y) \neq (0,0), \\ 0 & \text{if } (x,y) = (0,0). \end{cases}$$

We compute the partial derivatives

$$D_1 f(x,y) = \begin{cases} \frac{y(x^4+4x^2y^2-y^4)}{(x^2+y^2)^2} & \text{if } (x,y) \neq (0,0), \\ 0 & \text{if } (x,y) = (0,0). \end{cases}$$

and

$$D_2 f(x,y) = \begin{cases} \frac{x(x^4-4x^2y^2-y^4)}{(x^2+y^2)^2} & \text{if } (x,y) \neq (0,0), \\ 0 & \text{if } (x,y) = (0,0). \end{cases}$$

where in both cases, the derivative at zero is computed from the definition. Now we check from the definition that $\mathrm{D}_{21}\mathrm{f}(0,0) = -1$ and $\mathrm{D}_{12}\mathrm{f}(0,0) = 1$.

For completeness of the discussion we define the second derivative.

Definition 9.1. Function $\mathbf{f} : E \to \mathbb{R}^m$, $E \subset \mathbb{R}^m$ open, is twice differentiable at $\mathbf{a} \in E$ if:

1. The derivative $\mathrm{D}\mathbf{f}(\mathbf{x})$ exists for all $\mathbf{x}$ in some neighborhood U of $\mathbf{a}$.
2. For every $\mathbf{h} \in \mathbb{R}^n$ the function $w_{\mathbf{h}} : U \to \mathbb{R}^m$ given by

$$\mathbf{x} \mapsto \mathrm{D}\mathbf{f}(\mathbf{x})(\mathbf{h})$$

is differentiable at $\mathbf{a}$.

In that case the mapping

$$(\mathbf{h}', \mathbf{h}) \mapsto \mathrm{D}w_{\mathbf{h}}(\mathbf{x})(\mathbf{h}')$$

is bilinear. We call this map the second derivative of $\mathbf{f}$ at $\mathbf{a}$ and denote by $\mathrm{D}^2\mathbf{f}(\mathbf{a})$.

Example 9.2. In Example 7.7 we showed that the derivative of $f(\mathbf{x}) = \mathbf{x}^T A\mathbf{x} + \mathbf{x}^T\mathbf{b} + c$ is the linear transformation $\mathrm{Df}(\mathbf{x})$ that maps $\mathbf{h}$ to $(2\mathbf{x}^T A + \mathbf{b}^T)\mathbf{h}$. Thus for a fixed $\mathbf{h} \in \mathbb{R}^n$ we have $w_{\mathbf{h}}(\mathbf{x}) = (2\mathbf{x}^T A + \mathbf{b}^T)\mathbf{h}$ and so $\mathrm{Dw}_{\mathbf{h}}$ is a linear map represented by $2\mathbf{h}^T A$. Therefore $D^2 f(\mathbf{x})$ is the bilinear mapping $(\mathbf{h}, \mathbf{h}') \mapsto 2\mathbf{h}^T A\mathbf{h}'$.

Similarly like for the second derivative, the second derivative (if exists) is completely specified by the second order partial derivatives. The Hessian $\nabla^2\mathbf{f}$ of $\mathbf{f}$ is an $n \times n$ matrix whose (i, j)-th entry is $D_{ij}\mathbf{f}(\mathbf{a})$. We then have

$$\mathrm{D}^2\mathbf{f}(\mathbf{a})(\mathbf{h}, \mathbf{h}') \;=\; \mathbf{h}^{\mathrm{T}} \cdot \nabla^2\mathbf{f}(\mathbf{a}) \cdot \mathbf{h}'.$$

Also here, computing the Hessian matrix is the best approach. In the following exercise it is easier to compute the second derivative directly from the definition.

Exercise 9.1. Find the second derivative of the matrix valued function $X \mapsto X^{-1}$ defined on the subset Ω_n of invertible $n \times n$ matrices.

9.2 Constrained optima and Lagrange multipliers

In applications, constraints in an optimization problem can often be expressed as zeros of some functions. Let E be an open subset of $\mathbb{R}^n$, and let $\mathbf{F} : E \to \mathbb{R}^m$ be a $\mathcal{C}^1$ mapping. Define

$$X \;=\; \{\mathbf{x} \in E :\; \mathbf{F}(\mathbf{x}) = \mathbf{0}_m\}.$$

In this section we study the problem of optimizing a function $f : E \to \mathbb{R}$ over the set X. More precisely, we will derive necessary conditions for a point $\mathbf{c} \in X$ to be a local optimum of f over X.

Exercise 9.2. Recall Definition 7.5 and show that $\mathbf{c} \in X$ is a local maximum of f over X if and only if there exists a neighborhood U of $\mathbf{c}$ in $\mathbb{R}^n$ such that $f(\mathbf{c}) \geq f(\mathbf{x})$ for all $\mathbf{x} \in U \cap X$.

Suppose first that we are in this favourable situation that X admits a global parameterization: there is an invertible $\mathcal{C}^1$ function $\Phi : \mathbb{R}^d \to X$, whose inverse is also $\mathcal{C}^1$. Then optimizing a function $f(\mathbf{x})$ over X is equivalent to optimizing $f(\Phi(\mathbf{z}))$ over $\mathbb{R}^d$.

Exercise 9.3. With assumptions as above, show that a point $\mathbf{c} = \Phi(\mathbf{b})$ is a local optimum of f over X *if and only if* $\mathbf{b}$ is a local optimum of $f \circ \Phi$.

In this sense, the constrained optimization over X can be reduced to unconstrained optimization over $\mathbb{R}^d$. In particular, if f is differentiable everywhere then the necessary condition for an optimum at $\mathbf{z} \in \mathbb{R}^d$ is that $\nabla(f \circ \Phi)(\mathbf{z}) = \mathbf{0}_n$. Using the chain rule (Theorem 7.4) we get that

$$\mathrm{J(f \circ \Phi)(\mathbf{z})} \;=\; \mathrm{Jf}(\Phi(\mathbf{z})) \cdot \mathrm{J}\Phi(\mathbf{z}) \;=\; \mathbf{0}_{1\times \mathrm{d}}.$$

Equivalently, for every $\mathbf{h} \in \mathbb{R}^d$ we have $\mathrm{Jf}(\Phi(\mathbf{z}))\cdot \mathrm{J}\Phi(\mathbf{z})\cdot \mathbf{h} = 0$ and so $\mathrm{Jf}(\Phi(\mathbf{z}))$ transforms each vector in the image of $\mathrm{J}\Phi(\mathbf{z})$ to zero:

$$\mathrm{Im}(\mathrm{J}\Phi(\mathbf{z})) \;\subset\; \ker(\mathrm{Jf}(\Phi(\mathbf{z}))).$$

For example, this condition holds if

$$\mathrm{Jf}(\Phi(\mathbf{z})) = (1,-1,0) \qquad \text{and} \qquad \mathrm{J}\Phi(\mathbf{z}) = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 1 \end{bmatrix}.$$

Example 9.3. Let $f(x,y,z) = e^{x^2+y^2} + z^2$ and minimize f over the set of points satisfying $2x-y-1=0$ and $3x-z-2=0$. This two equations define a line in $\mathbb{R}^3$ which is parametrized by $\Phi(t) = (t, 2t-1, 3t-2)$. Therefore, minimizing f over X is equivalent to minimizing

$$f(t, 2t-1, 3t-2) = e^{5t^2-4t+1} + (3t-2)^2.$$

Example 9.4. Suppose X is given in $\mathbb{R}^3$ by two non-linear equations $y = x^2$, $z = x^3$. Then $\Phi(t) = (t, t^2, t^3)$ for $t \in \mathbb{R}$ parameterizes X. Suppose we want to optimize $f(x,y,z) = xz - 2y + 1$ over X. This amounts to optimizing $f(\Phi(t)) = t^4 - 2t^2 + 1 = (t^2-1)^2$, which has two global minima for $t = \pm 1$ and one local maximum at $t = 0$. It follows that the global minima of f are $(-1,1,-1)$ and $(1,1,1)$, and the local maximum is given by $(0,0,0)$. This function has no global maxima.

In practice often it is hard to find a global parametrization of X or such a parametrization does not exist. Using the Implicit Function Theorem we can still proceed if such a parameterization exists at least locally; c.f. Remark 8.4.

Before making any formal statement the following discussion should be useful. For unconstrained optimization problem we learned from Corollary 7.2 that the necessary condition for $f : U \to \mathbb{R}$ to have a maximum at $\mathbf{c}$ is that $\nabla f(\mathbf{c}) = \mathbf{0}$, or equivalently, $\mathrm{D_u f}(\mathbf{c}) = 0$ for every direction $\mathbf{u}$. In other words, *it is not possible to improve the function moving infinitesimally from* $\mathbf{c}$. The same holds true for constrained optimization problems over X, but now the infinitesimal directions we allow are only the ones that do not take us away from X, that is, $\mathrm{D_u F}(\mathbf{c}) = \mathbf{0}_{\mathrm{m}}$.

Definition 9.2. If $\mathrm{Im}(\mathrm{DF}(\mathbf{c})) = \mathbb{R}^{\mathrm{m}}$, define the tangent space to X at point $\mathbf{c} \in X$ as the kernel of $\mathrm{DF}(\mathbf{c})$

$$\mathrm{Tan}_{\mathbf{c}}X \;:=\; \ker(\mathrm{D}\mathbf{F}(\mathbf{c})) \;\subset\; \mathbb{R}^{\mathbf{n}}.$$

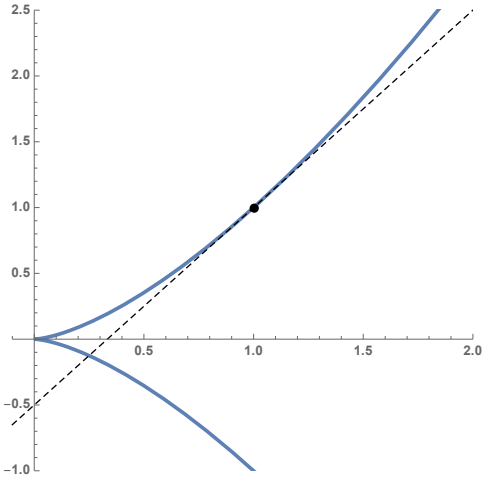

Fig. 9.1 The tangent space to $F(x,y) = y^2 - x^3 = 0$ at $(1,1)$.

Example 9.5. For a function $F(x,y) = y^2 - x^3$ we have $\mathrm{JF(x,y)} = (-3\mathrm{x}^2, 2\mathrm{y})$. If $\mathbf{c} = (1,1)$ then $\mathrm{JF}(\mathbf{c}) = (-3,2)$. The kernel is the set of points (x,y) such that $-3x + 2y = 0$. Shifting this linear space to the point $(1,1)$ we get the line tangent to $F(x,y) = 0$ at this point, c.f. Figure 9.2. On the other hand, if $\mathbf{c} = (0,0)$ then $F(\mathbf{c}) = 0$ but $\mathrm{JF}(\mathbf{c}) = (0,0)$ and so the condition $\mathrm{Im}(\mathrm{DF}(\mathbf{c})) = \mathbb{R}$ does not hold. Here, the kernel of $\mathrm{JF}(\mathbf{c})$ is the whole $\mathbb{R}^2$.

To get some geometric intuition behind this definition recall that for every $t \in \mathbb{R}$ and $\mathbf{u} \in \mathbb{R}^n$ we get

$$\mathbf{F}(\mathbf{c} + t\mathbf{u}) - \mathbf{F}(\mathbf{c}) \;=\; \mathrm{D}\mathbf{F}(\mathrm{c})(\mathrm{t}\mathbf{u}) + \mathbf{r}(\mathrm{t}\mathbf{u}),$$

where $\lim_{t\to 0} \frac{\mathbf{r}(t\mathbf{u})}{t\|\mathbf{u}\|} = \mathbf{0}$. Using the fact that $\mathbf{F}(\mathbf{c}) = \mathbf{0}$ and linearity of the derivative we get that

$$\lim_{t\to 0} \tfrac{1}{t\|\mathbf{u}\|}\mathbf{F}(\mathbf{c} + t\mathbf{u}) \;=\; \tfrac{1}{\|\mathbf{u}\|}\mathrm{D}\mathbf{F}(\mathrm{c})(\mathbf{u}).$$

The left hand side represents the instanteneous rates of change of each component of $\mathbf{F}$ as we move from $\mathbf{c}$ in the direction of $\mathbf{u}$ (c.f. Section 7.1). The right-hand side is zero if and only if $\mathbf{u} \in \ker(\mathrm{D}\mathbf{F}(\mathbf{c}))$. In this case moving infinitesimally in the direction $\mathbf{u}$ the value of $\mathbf{F}(\mathbf{c} + t\mathbf{u})$ remains $\mathbf{0}$ (up to the first order), or, in other words, $\mathbf{c} + t\mathbf{u} \in X$; c.f. Figure 9.2. On the other hand, if $\mathbf{v} \notin \ker(\mathrm{D}\mathbf{F}(\mathbf{c}))$ then moving infinitesimally from $\mathbf{c}$ in the direction $\mathbf{v}$ will take us out of X.

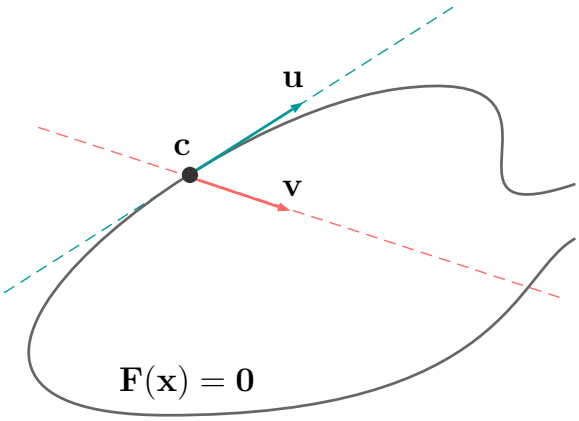

Fig. 9.2 A curve $\mathbf{F}(\mathbf{x}) = \mathbf{0}$ and its tangent line at $\mathbf{c}$ spanned by $\mathbf{u}$.

Theorem 9.3. *Let* $\mathbf{F} : E \to \mathbb{R}^m$ *and* $f : E \to \mathbb{R}$ *be* $\mathcal{C}^1$ *on* E. *Let* X *be as above,* $\mathbf{c} \in X$, *and assume* $\mathrm{Im}(\mathrm{D}\mathbf{F}(\mathbf{c})) = \mathbb{R}^{\mathrm{m}}$. *If* $\mathbf{c}$ *is a local optimum of* f *restricted to* X *then*

$$\ker(\mathrm{D}\mathbf{F}(\mathbf{c})) \quad \subset \quad \ker(\mathrm{D}f(\mathbf{c})). \tag{9.1}$$

On the intuitive level the proof should be clear. If there exists a vector $\mathbf{u}$ in the kernel of $\mathrm{D}\mathbf{F}(\mathbf{c})$ that does not lie in the kernel of $\mathrm{D}f(\mathbf{c})$ then $\mathrm{D}_{\mathbf{u}}\mathbf{F}(\mathbf{c}) = \mathbf{0}$ and $\mathrm{D}_{\mathbf{u}}\mathrm{f}(\mathbf{c}) \neq \mathbf{0}$. Moving infinitesimally in the direction $\mathbf{u}$ or $-\mathbf{u}$ allows us to stay in X up to the first order but increase or decrease the value of f. A formal proof follows.

Proof. By the Implicit Function Theorem (Theorem 8.10) the equation $\mathbf{F}(\mathbf{x}) = \mathbf{0}$ implicitly expresses m passive variables in terms of $n-m$ active variables in some neighbourhood U of $\mathbf{c}$. Write $\mathbf{c} = (\mathbf{a}, \mathbf{b})$, where $\mathbf{a} \in \mathbb{R}^m$ corresponds to the passive variables. There is a function $\mathbf{g}$ such that $\mathbf{c} = (\mathbf{g}(\mathbf{b}), \mathbf{b})$ and $\Phi(\mathbf{z}) = (\mathbf{g}(\mathbf{z}), \mathbf{z})$ locally parameterizes X near $\mathbf{c}$, that is $\mathbf{F}(\Phi(\mathbf{z})) = \mathbf{0}$ in a neigbourhood W of $\mathbf{b}$. The point $\mathbf{c}$ is a local optimum of f over X if and only if $\mathbf{b}$ is a local optimum of $f \circ \Phi$. Indeed, if $f(\mathbf{c})$ has the optimal value in some neighborhood $U' \cap X$ of $\mathbf{c}$ in X (say $U' \subset U$) if and only if $(f \circ \Phi)(\mathbf{b})$ has the optimal value in some neighborhood $W' \subset W$. Now we use the necessary condition for unconstrained optimization. An optimum must necessarily satisfy $\mathrm{D}(\mathrm{f} \circ \Phi)(\mathbf{b}) = \mathbf{0}$. Using the chain rule, we get

$$\mathrm{D}(\mathrm{f} \circ \Phi)(\mathbf{b}) = \mathrm{D}f(\Phi(\mathbf{b})) \circ \mathrm{D}\Phi(\mathbf{b}) = \mathrm{D}f(\mathbf{c}) \circ \mathrm{D}\Phi(\mathbf{b}).$$

For this to be a zero transformation, the image of $\mathrm{D}\Phi(\mathbf{b})$ must be contained in the kernel of $\mathrm{D}f(\mathbf{c})$, $\mathrm{Im}(\mathrm{D}\Phi(\mathbf{c})) \subset \ker(\mathrm{D}f(\mathbf{c}))$. To finish the proof it is enough to show that $\mathrm{Im}(\mathrm{D}\Phi(\mathbf{c})) = \ker(\mathrm{D}\mathbf{F}(\mathbf{c}))$. Writing $A = \mathrm{J}\mathbf{F}(\mathbf{a}, \mathbf{b})$ we get

$$\mathrm{J}\Phi(\mathbf{b}) \;=\; \begin{bmatrix}\mathrm{Jg}(\mathbf{b})\\ \mathbb{I}_{n-m}\end{bmatrix} \overset{(8.5)}{=} \begin{bmatrix}-A_y^{-1}A_z\\ \mathbb{I}_{n-m}\end{bmatrix}$$

and therefore

$$(\mathbf{h},\mathbf{k}) \in \mathrm{Im}(\mathrm{D}\Phi(\mathbf{b})) \quad\Longleftrightarrow\quad \mathbf{h} = -\mathrm{A}_{\mathbf{y}}^{-1}\cdot\mathrm{A}_{\mathbf{z}}\cdot\mathbf{k} \quad\Longleftrightarrow\quad [\mathrm{A}_{\mathbf{y}} \mid \mathrm{A}_{\mathbf{z}}]\begin{bmatrix}\mathbf{h}\\ \mathbf{k}\end{bmatrix} = \mathbf{0}.$$

This last condition is equivalent to $(\mathbf{h},\mathbf{k}) \in \ker(\mathrm{J}\mathbf{F}(\mathbf{c}))$ and so $\mathrm{Im}(\mathrm{D}\Phi(\mathbf{b})) = \ker(\mathrm{D}\mathbf{F}(\mathbf{c}))$. □

This gives the following fundamental theorem of constrained optimization.

Theorem 9.4. *(Lagrange multipliers) With assumptions like in Theorem 9.3. If $\mathbf{c} \in X$ is a local optimum of f restricted to X then there exist numbers $\lambda_1,\ldots,\lambda_m$ such that:*

$$\nabla f(\mathbf{c}) \;=\; \lambda_1\nabla F_1(\mathbf{c}) + \cdots + \lambda_m\nabla F_m(\mathbf{c}).$$

In other words $\mathbf{c}$ is a stationary point of the Lagrangian function

$$L(\mathbf{x};\lambda) = f(\mathbf{x}) - \sum_{i=1}^{m}\lambda_i F_i(\mathbf{x}).$$

Proof. By Theorem 9.3, if $\mathbf{c}$ is a local optimum over X then $\ker(\mathrm{J}\mathbf{F}(\mathbf{c})) \subset \ker(\mathrm{Jf}(\mathbf{c}))$, where $\mathrm{J}\mathbf{F}(\mathbf{c}) \in \mathbb{R}^{\mathrm{m}\times\mathrm{n}}$ is the Jacobian matrix of $\mathbf{F}$ and $\mathrm{Jf}(\mathbf{c}) \in \mathbb{R}^{1\times\mathrm{n}}$ is (up to transposition) the gradient of f. Recall (5.3), which implies

$$\mathrm{Im}(\mathrm{J}\mathbf{F}(\mathbf{c})^{\mathrm{T}})^{\perp} \subset \mathrm{Im}(\mathrm{Jf}(\mathbf{c})^{\mathrm{T}})^{\perp}.$$

This, by Lemma 5.3, is equivalent to

$$\mathrm{Im}(\mathrm{Jf}(\mathbf{c})^{\mathrm{T}}) \subset \mathrm{Im}(\mathrm{J}\mathbf{F}(\mathbf{c})^{\mathrm{T}}),$$

which happens if and only if $\nabla f(\mathbf{c}) = \mathrm{J}\mathbf{F}(\mathbf{c})^{\mathrm{T}}\boldsymbol{\lambda}$ for some $\boldsymbol{\lambda} \in \mathbb{R}^m$. □

Example 9.6. Suppose we want to maximize $f(x,y) = x + y$ over the ellipse $x^2+2y^2 = 1$. We have $F(x,y) = x^2+2y^2-1$, $\mathrm{JF(x,y)} = [2\mathrm{x}, 4\mathrm{y}]$, and $\mathrm{Jf(x,y)} = [1,1]$. At critical points there must exist λ such that $[1,1] = \lambda[2x,4y]$, which implies that the solution must be of the form

$$x = \tfrac{1}{2\lambda}, \qquad y = \tfrac{1}{4\lambda}.$$

Plugging this into the constraint gives $\lambda = \pm\sqrt{\frac{3}{8}}$. It implies that there are two candidates for local optima

$$(x,y) = \pm\left(\sqrt{\tfrac{2}{3}}, \sqrt{\tfrac{1}{6}}\right).$$

It turns out that one of these points corresponds to the maximum, which is $\sqrt{\frac{3}{2}}$ and the minimum $-\sqrt{\frac{3}{2}}$.

Example 9.7. To find the maximum of the function $x_1 x_2 \cdots x_n$ subject to the constraint

$$x_1^2 + 2x_2^2 + \cdots + nx_n^2 = 1,$$

define first $y_i = \sqrt{i}x_i$. Then the problem is equivalent to maximizing $\frac{1}{\sqrt{n!}} \prod_i y_i$ over the sphere $y_1^2 + \cdots + y_n^2 = 1$. First note that the maximum point contains no zeros because then the value of the function is zero, which is clearly not optimal. By the Lagrange theorem (we easily check it can be used), the optimum for every j must satisfy that $\frac{1}{\sqrt{n!}} \prod_{i \neq j} y_i = 2\lambda y_j$, $\lambda \in R$. In other words (because we can assume $y_j \neq 0$), $\frac{1}{\sqrt{n!}} \prod_i y_i = 2\lambda y_j^2$ and, in particular, all y_i^2 must be equal. So all the optima are of the form $y_i = \pm\frac{1}{\sqrt{n}}$. The maxima correspond to the sign pattern such that the product is positive. The maximum value is $\frac{1}{\sqrt{n!n^n}}$

Chapter 10
Elementary measure theory (3 lectures)

This chapter was created to provide a crash course in measure theory. We decided to omit many proofs. Some are kept to give the student a flavor of the underlying mathematics. Our idea was to introduce basics of the general measure theory with focus on the Legesgue and the probability measures. For more in-depth treatment see, for example, "Probability and measure" by Patrick Billingsley.

10.1 Motivation and measure spaces

Let X be a set. We are interested in defining a measure μ on the set of subsets of X. We would like the measure to be a function which generalizes length, area and volume in the familiar situations. So, for example, if $X = \mathbb{R}$ then we would like the measure of the interval (a, b) to be its length $b - a$. More generally, if $X = \mathbb{R}^k$ then we would like the measure of the k-cell $(a_1, b_1) \times \cdots \times (a_k, b_k)$ to be $\prod_{i=1}^{k}(b_i - a_i)$.

The first question to ask is what should be the defining properties of a measure μ on X. An obvious thing is to require that $\mu(A)$ is always nonnegative (but possibly infinite) and so μ takes values in $[0, +\infty] \subset \overline{\mathbb{R}}$, where $\overline{\mathbb{R}}$ denotes the extended real-number system; c.f. Section 1.2. Moreover, $\mu(A \cup B) = \mu(A) + \mu(B)$ whenever $A, B \subset X$ are disjoint. More generally, for any countable family $\{A_n\}$ of disjoint sets we require

$$\mu\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mu(A_n).$$

We say that μ is **countably additive**. For **probability measures** we also require that $\mu(X) = 1$.

Remark 10.1. To understand countable sums note that for any real sequence $(x_n)_{n\in\mathbb{N}}$, we formally define $\sum_{n=1}^{\infty} x_n$ as the limit, if exists, in $\overline{\mathbb{R}}$ of the sequence $s_k = \sum_{n=1}^{k} x_n$. If $x_n \geq 0$ for all $n \in \mathbb{N}$ then this limit always exists.

Example 10.1. If X is finite, it is enough to specify $\mu(\{i\})$ for all $i \in X$ and then extend to other subsets of X by

$$\mu(A) \;=\; \sum_{i\in A} \mu(\{i\}).$$

If $\mu(\{i\}) = 1$ we get so called **counting measure**.

A nice property of the finite case is that measure could be constructed by specifying its value on some basic sets (singletons) and then extended to all other sets by additivity. In general, this is much too much to ask for. Roughly speaking, there are too many too complicated sets in X to have control over what can happen. If $X = \mathbb{R}$ we have the following important impossibility theorem.

Theorem 10.1 (Hausdorff). *There is no measure defined on* all *subsets of* $\mathbb{R}$ *that is invariant under translations.*

It is standard in measure theory to focus on "nice" families of subsets of X and defining measure over those. Importance of the following definition will become clear soon.

Definition 10.1. A collection $\mathcal{F}$ of subsets of X is a **σ-algebra** if

(i) $X \in \mathcal{F}$,
(ii) $A \in \mathcal{F}$ then $A^c \in \mathcal{F}$,
(iii) $A_1, A_2, \ldots \in \mathcal{F}$ then $\bigcup_{n\in\mathbb{N}} A_n \in \mathcal{F}$.

For every X, $\mathcal{F} = \{\emptyset, X\}$ is a σ-algebra called the **trivial σ-algebra**. If X is finite and $\mathcal{F}$ is such that $\{i\} \in \mathcal{F}$ for all $i \in X$ then $\mathcal{F}$ is equal to the set of all subsets of X. In general, for every X, its power set 2^X is a σ-algebra.

Exercise 10.1. Show that every σ-algebra is closed under countable intersections: if $A_1, A_2, \ldots \in \mathcal{F}$ then $\bigcap_{n\in\mathbb{N}} A_n \in \mathcal{F}$.

From our perspective not all σ-algebras are interesting. We typically want the measure to be explicitly defined on particular subsets (like intervals in $\mathbb{R}$) and then extended in a consistent way to as many other subsets as possible. Exercise 10.2 allows us to define the smallest σ-algebra containing any fixed collection $\mathcal{A}$ of subsets of $\mathcal{X}$ by simply taking the intersection of all σ-algebras containing it.

Exercise 10.2. Consider an arbitrary family of σ-algebras $\mathcal{F}_\lambda$, $\lambda \in \Lambda$. Show that $\bigcap_{\lambda\in\Lambda} \mathcal{F}_\lambda$ is also a σ-algebra.

From Exercise 10.2 we conclude that, given any collection $\mathcal{A}$ of sets of X, there exists a smallest σ-algebra, denoted by $\sigma(\mathcal{A})$ that contains it. This σ-algebra is given as the intersection of *all* σ-algebras containing $\mathcal{A}$. Note that there is always at least one such σ-algebra, namely the power set 2^X. We say that $\mathcal{A}$ generates $\sigma(\mathcal{A})$. An important example comes next.

Definition 10.2. Let X be a metric space. The open sets of X generate a σ-algebra $\mathcal{B}(X)$ called the **Borel algebra**.

Note that directly by definition $\mathcal{B}(X)$ must contain all open and all closed subsets of X. However, it must also contain countable unions of closed sets which are generally neither open nor closed. The Borel σ-algebra is actually quite rich and it is hard to come up with set that does not lie in $\mathcal{B}(X)$ (construction for $X = \mathbb{R}$ relies on the axiom of choice).

We finish this section formally defining a measure.

Definition 10.3. A **measure** is any nonnegative and countably additive function defined on a σ-algebra $\mathcal{F}$ of sets of X.

Definition 10.4. A **measure space** is defined to be a triple $(X, \mathcal{F}, \mu)$, where X is a non-empty set, $\mathcal{F}$ is a σ-algebra of subsets of X, and μ is a measure on $\mathcal{F}$. The measure space is **finite** if $\mu(X) < +\infty$; it is **σ-finite** if X is a countable union of sets on which μ is finite.

A **probability space** is defined to be a triple $(\Omega, \mathcal{F}, \mathbb{P})$ that forms a measure space with an additional assumption that $\mathbb{P}(\Omega) = 1$. In this case Ω is called the space of **elementary events** and $\mathcal{F}$ is the **space of events**. For any $A \in \mathcal{F}$ the number $\mathbb{P}(A)$ is the probability that the event A occurs. The normalizing equation $\mathbb{P}(\Omega) = 1$ just says that the event Ω is certain.

10.2 Lebesgue measure and Extension Theorem

The basic idea behind the **Lebesgue measure** is to construct a measure on $\mathbb{R}$ such that for each bounded interval its measure is simply the length. Then to extend this to as large class of subsets of $\mathbb{R}$ as possible. To introduce it formally, we start with a general discussion.

Definition 10.5. A family $\mathcal{R}$ of sets is called a **ring** if $A, B \in \mathcal{R}$ implies that

$$A \cup B \in \mathcal{R} \qquad \text{and} \qquad A \setminus B \in \mathcal{R}.$$

Since $A \cap B = A \setminus (A \setminus B)$, we also have $A \cap B \in \mathcal{R}$ if $\mathcal{R}$ is a ring. Also, $\emptyset \in \mathcal{R}$ but, in general, X does not have to lie in $\mathcal{R}$.

Exercise 10.3. Show that every ring of subsets of X is closed under finite unions, finite intersections, and symmetric difference.

Definition 10.6. We say that ρ is a **set function** defined on a ring $\mathcal{R}$ of subsets of X if ρ assigns to each $A \in \mathcal{R}$ a number $\rho(A)$ of the extended real number system $\overline{\mathbb{R}}$. We say that ρ is **additive** if $A, B \in \mathcal{R}$, $A \cap B = \emptyset$ implies

$$\rho(A \cup B) = \rho(A) + \rho(B).$$

Moreover, ρ is **countably additive** if for any family of sets $A_n \in \mathcal{R}$, $n \in \mathbb{N}$, such that $A_i \cap A_j = \emptyset$ for all $i \neq j$ and $\bigcup_n A_n \in \mathcal{R}$ we have that

$$\rho\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \rho(A_n).$$

We say that ρ is **nonnegative** if its values lie in $[0, +\infty]$.

Note if ρ is additive then $\rho(\emptyset) = \rho(\emptyset) + \rho(\emptyset)$ and so $\rho(\emptyset) = 0$.

Remark 10.2. Our goal is to define a nonnegative countably additive function on $\mathcal{R}$ and then to extend it to a nonnegative countably additive function on the σ-algebra $\sigma(\mathcal{R})$ generated by $\mathcal{R}$.

Exercise 10.4. Show that if ρ is an additive set function on $\mathcal{R}$ and $A_1, \ldots, A_N$ are pairwise disjoint sets in $\mathcal{R}$ then

$$\rho\left(\bigcup_{n=1}^{N} A_n\right) = \sum_{n=1}^{N} \rho(A_n).$$

Proposition 10.1. *An additive and nonnegative set function ρ on a ring $\mathcal{R}$ of sets has the following properties for any two $A, B \in \mathcal{R}$:*

(a) $\rho(A \cup B) + \rho(A \cap B) = \rho(A) + \rho(B)$.
(b) If $\rho(A) < +\infty$ *then* $\rho(A \cap B) < +\infty$.
(c) If $\rho(A \cap B) < +\infty$, *then* $\rho(A \setminus B) = \rho(A) - \rho(A \cap B)$.
(d) If $B \subset A$ *then* $\rho(B) \leq \rho(A)$.

Proof. In the proof we use the fact that for $A, B \in \mathcal{R}$ also $A \setminus B$, $B \setminus A$, $A \cap B$ lie in $\mathcal{R}$.
(a) Decompose $A \cup B$ as the disjoint union of $A \setminus B$, $B \setminus A$, and $A \cap B$. All these sets lie in $\mathcal{R}$ so $\rho(A \cup B) + \rho(A \cap B) = \rho(A \setminus B) + \rho(B \setminus A) + 2\rho(A \cap B)$. The same equality holds for $\rho(A) + \rho(B)$ by using the disjoint decompositions $A = (A \setminus B) \cup (A \cap B)$ and $B = (B \setminus A) \cup (A \cap B)$.
(b) Decompose $A = (A \setminus B) \cup (A \cap B)$ and use the fact that $\rho(A \setminus B) + \rho(A \cap B)$ is a finite number to conclude that both $\rho(A \cap B)$ and $\rho(A \setminus B)$ must be finite.
(c) Again decompose $A = (A \setminus B) \cup (A \cap B)$ to get $\rho(A \setminus B) + \rho(A \cap B) = \rho(A)$. If $\rho(A \cap B) < +\infty$, we can subtract it from both sides of this equation (even if some other quantities are not finite).
(d) Decompose $A = B \cup (A \setminus B)$ to get that $\rho(A) = \rho(B) + \rho(A \setminus B)$. □

Proposition 10.2. *If ρ is nonnegative and countably additive on $\mathcal{R}$ and $A, A_1, A_2, \ldots$ are sets in $\mathcal{R}$ such that $A \subset \bigcup_{n=1}^{\infty} A_n$, then $\rho(A) \leq \sum_{n=1}^{\infty} \rho(A_n)$.*

Proof. Let $B_n = A_n \setminus \left(\bigcup_{k=1}^{n-1} A_k\right)$ and note that $B_n \in \mathcal{R}$. Then $A = \bigcup_{n=1}^{\infty}(A \cap B_n)$ is a disjoint decomposition. Since ρ is countably additive and $A \in \mathcal{R}$, $\rho(A) = \sum_{n=1}^{\infty} \rho(A \cap B_n)$. By Proposition 10.1(d), $\rho(A \cap B_n) \leq \rho(B_n) \leq \rho(A_n)$. It follows that $\rho(A) = \sum_{n=1}^{\infty} \rho(A \cap B_n) \leq \sum_{n=1}^{\infty} \rho(A_n)$. □

We are finally ready to present the construction of the Legesgue measure on $X = \mathbb{R}$. The construction for general $\mathbb{R}^k$ is similar. Let $\mathcal{E}$ be the set of **elementary sets of** $\mathbb{R}$, that is, the sets which are finite disjoint unions of bounded intervals in $\mathbb{R}$ of the form $(a, b]$ where $-\infty < a \leq b < +\infty$. Therefore, if $A \in \mathcal{E}$ then $A = I_1 \cup \ldots \cup I_n$ is a disjoint union where $I_i = (a_i, b_i]$. We will always a convention that $I_1, \ldots, I_n$ are ordered from the left to the right, that is, $a_1 < a_2 < \cdots < a_n$.

Lemma 10.1. *$\mathcal{E}$ forms a ring of sets.*

Proof. Let $A, B \in \mathcal{E}$, that is, $A = I_1 \cup \cdots \cup I_m$, $B = J_1 \cup \cdots \cup J_n$ where $I_i \cap I_j = \emptyset$, $J_i \cap J_j = \emptyset$ for all i, j and I_i, J_j are of the form $(a, b]$. To show that $A \setminus B$ lies in $\mathcal{E}$ note that

$$A \setminus B = A \cap B^c = \bigcup_{i=1}^{m} (I_i \cap B^c)$$

is a disjoint union. It is then enough to show that each $I_i \cap B^c$ lies in $\mathcal{E}$. Write $I_i = (a, b]$, $J_j = (a_j, b_j]$ then

$$I_i \cap B^c \;=\; (a, b] \cap \left((-\infty, a_1] \cup \bigcup_{j=1}^{n-1} (b_j, a_{j+1}] \cup (b_n, +\infty)\right)$$

Using the distributive law we conclude that $I_i \cap B^c$ is a disjoint union of intervals of the given form. This shows that $A \setminus B$ lies in $\mathcal{R}$. To show that $A \cap B$ lies in $\mathcal{R}$ decompose it into the disjoint union $(A \setminus B) \cup (A \cap B) \cup (B \setminus A)$. By the first part of the proof, it is enough to show that $A \cap B$ lies in $\mathcal{R}$. We have $A \cap B = \bigcup_{i,j} (I_i \cap J_j)$ which is a disjoint union. Moreover, each $I_i \cap J_j$ is an interval of the form $(a, b]$. □

There is an obvious way to define a nonnegative and additive set function m on the ring $\mathcal{E}$. If $A \in \mathcal{E}$ then $A = \bigcup_{i=1}^{n} I_i$, where I_i are bounded and disjoint intervals, $I_i = (a_i, b_i]$. Let $m(I_i) = b_i - a_i$ and

$$m(A) = m(I_1) + \ldots + m(I_n).$$

We first show that this set function is well-defined.

Lemma 10.2. *If A has two representations as finite unions of disjoint bounded intervals, that is, $A = \bigcup_{i=1}^m I_i = \bigcup_{j=1}^n J_j$, then $\sum_i m(I_i) = \sum_j m(J_j)$.*

Proof. Let $J_j = (a_j, b_j]$. If $m = 1$, $A = (a, b]$ then $a_1 = a$, $b_n = b$, and $b_i = a_{i+1}$ for all $i = 1, \ldots, n-1$. It follows that

$$\sum_{j=1}^n m(J_j) = (b_1 - a_1) + (b_2 - a_2) + \cdots + (b_n - a_n) = b_n - a_1 = b - a.$$

In the general case, when $m \geq 1$, we use the fact that $J_j \subset \bigcup_{i=1}^m I_i$ and so $J_j = \bigcup_{i=1}^m (I_i \cap J_j)$ is a disjoint union of intervals (each $I_i \cap J_j$ is an interval), which gives that $m(J_j) = \sum_{i=1}^m m(I_i \cap J_j)$ and so

$$\sum_j m(J_j) = \sum_{j=1}^n \sum_{i=1}^m m(I_i \cap J_j).$$

We get the same equality decomposing each I_i as $I_i = \bigcup_{j=1}^n (I_i \cap J_j)$. □

Suppose now that $A = \bigcup_{n=1}^\infty A_n$ where A_n and *disjoint* sets in $\mathcal{E}$. If $A \in \mathcal{E}$, do we also have $m(A) = \sum_n m(A_n)$? The following fundamental result implies that this holds, or in other words that m is countably additive.

Proposition 10.3. *The function m is countably additive on $\mathcal{E}$.*

Before we prove this result, we need a short discussion of regularity.

Definition 10.7. A nonnegative additive set function defined on $\mathcal{E}$ is said to be **regular** if for every $A \in \mathcal{E}$ and $\epsilon > 0$ there exist a closed set F and an open set G in $\mathcal{E}$ such that $F \subset A \subset G$ and

$$m(G) - \epsilon \ \leq \ m(A) \ \leq \ m(F) + \epsilon.$$

Lemma 10.3. *The set function m is regular.*

Proof. Let $A \in \mathcal{E}$. If $A = (a, b]$, we take $G = (a - \epsilon/2, b + \epsilon/2)$ and $F = [a + \epsilon/2, b - \epsilon/2]$ (or $F = \emptyset$ if $b - a \leq \epsilon$). In the general case, A is a union of N intervals I_i. Choose G_i, F_i as above but replacing ϵ with ϵ/n and let $F = \bigcup_i F_i$, $G = \bigcup_i G_i$. Both F and G lie in $\mathcal{E}$. Moreover, $m(F) = \sum_{i=1}^n m(F_i) \geq (m(I_i) - \epsilon/n) = m(A) - \epsilon$, and, by Proposition 10.2, $m(G) \leq \sum_{i=1}^n m(G_i) \leq \sum_{i=1}^n (m(I_i) + \epsilon/n) = m(A) + \epsilon$.

□

Proof of Proposition 10.3. Let $\{A_n\}$ be a sequence of disjoint sets in $\mathcal{E}$ with union A in $\mathcal{E}$. Since m is nonnegative and additive, by Proposition 10.1(d),

$$m(A) \ \geq \ m(\bigcup_{i=1}^n A_i) \ = \ \sum_{i=1}^n m(A_i) \quad \text{for every } n \in \mathbb{N}.$$

Passing to the limit we get $m(A) \geq \sum_{i=1}^{\infty} m(A_i)$. For the reverse inequality, fix $\epsilon > 0$. By regularity of m there exists a closed set $F \in \mathcal{E}$ and open $G_n \in \mathcal{E}$ such that $F \subset A$, $A_n \subset G_n$ for all n, $m(F) \geq m(A) - \epsilon$, and $m(G_n) \leq m(A_n) + \epsilon/2^n$. Then $F \subset \bigcup_{n=1}^{\infty} G_n$, and compactness implies covering compactness (c.f. Theorem 4.9) and so $F \subset \bigcup_{n=1}^{N} G_n$ for some N. Hence,

$$m(A) - \epsilon \leq m(F) \leq \sum_{i=1}^{N} m(G_n) \leq \sum_{n=1}^{N} (m(A_n) + \epsilon/2^n) \leq \sum_{n=1}^{\infty} m(A_n) + \epsilon.$$

Since ϵ is arbitrary, $m(A) \leq \sum_{n=1}^{\infty} m(A_n)$, and the proposition follows. □

Suppose now that $A = \bigcup_{n=1}^{\infty} A_n$ where A_n are disjoint sets in $\mathcal{E}$ but $A \notin \mathcal{E}$. Can we still sensibly define $m(A)$ as $\sum_n m(A_n)$? Slightly more generally, can we extend m to measure on the σ-algebra $\sigma(\mathcal{E})$? Before we show that this can be done we show that $\sigma(\mathcal{E})$ is really the Borel σ-algebra.

Proposition 10.4. $\sigma(\mathcal{E}) = \mathcal{B}(\mathbb{R})$.

Proof. In Exercise 2.16 we have proven that each open set in $\mathbb{R}$ is a countable union of open intervals. Each open interval is a countable union of intervals of the form $(a, b]$. Indeed,

$$(a, b) = \bigcup_{n \in \mathbb{N}} (a, b - \tfrac{1}{n}].$$

This implies that $\mathcal{B}(\mathbb{R}) \subset \sigma(\mathcal{E})$. For the opposite inclusion we show that $\mathcal{E} \subset \mathcal{B}(\mathbb{R})$. For this, it is enough to show that each interval of the form $(a, b]$ lies in $\mathcal{B}(\mathbb{R})$, which follows because $\mathcal{B}(\mathbb{R})$ is closed under countable intersections and

$$[a, b) = \bigcap_{n \in \mathbb{N}} (a - \tfrac{1}{n}, b).$$

. □

We would like to use this construction to define a measure on the σ-algebra $\mathcal{B}(\mathbb{R})$. The extension is carried out by the general Extension Theorem. We formulate this theorem more generally for countably additive set functions ρ that are **σ-finite** on $\mathcal{R}$, that is, each set in $\mathcal{R}$ is contained in a countable union of sets in $\mathcal{R}$ on which ρ is finite. The ring $\mathcal{E}$ obviously satisfies this property.

Theorem 10.2 (Carathéodory's Extension Theorem). *A σ-finite non-negative and countably additive set function on a ring $\mathcal{R}$ has a unique extension to a measure μ on $\sigma(\mathcal{R})$.*

The construction of the measure μ in the extension theorem goes as follows. To every subset $E \subset X$ we define its **outer measure**

$$\mu^*(E) = \inf \sum_{n=1}^{\infty} \mu(A_n),$$

where $A_n \in \mathcal{R}$ and the inf is taken over all countable coverings of E by sets in $\mathcal{R}$. It is clear that $\mu^*(E) \geq 0$ and $\mu^*(A) \leq \mu^*(E)$ if $A \subset E$. It is also clear that $\mu^*(A) = \mu(A)$ for all $A \in \mathcal{R}$. The proof of the Extension Theorem relies on carefully showing that μ^* is a countably additive function on $\sigma(\mathcal{R})$ and that μ^* is a unique such function. We omit the detailed proof.

Example 10.2. In case of the Lebesgue measure, every countable set has measure zero. But there are uncountable sets of measure zero. The **Cantor set** may be taken as example. Let $E_0 = [0, 1]$. Remove the segment $(\frac{1}{3}, \frac{2}{3})$, and let E_1 be the resulting union of intervals $E_1 = [0, \frac{1}{3}] \cup [\frac{2}{3}, 1]$. Remove the middle thirds of these intervals to get

$$E_2 = [0, \tfrac{1}{9}] \cup [\tfrac{2}{9}, \tfrac{3}{9}] \cup [\tfrac{6}{9}, \tfrac{7}{9}] \cup [\tfrac{8}{9}, 1].$$

Continuing in this way, we obtain a sequence of compact sets E_n such that

(a) $E_1 \supset E_2 \supset E_3 \supset \ldots$,
(b) E_n is the union of 2^n intervals, each of length 3^{-n}.

The *nonempty* compact set $C = \bigcap_{n=1}^{\infty} E_n$ is called the Cantor set. It can be shown that

$$C = \left\{ \sum_{n=1}^{\infty} \tfrac{a_n}{3^n} : \ a_n \in \{0, 2\} \right\}.$$

In other words C is in bijection with infinite $0/2$ sequences. By the same argument as in Theorem 2.2 conclude that C is not countable. Nevertheless, C has Lebesque measure zero. It is easily seen that for every $n \in \mathbb{N}$

$$m(E_n) = \left(\tfrac{2}{3}\right)^n$$

and since $C = \bigcap E_n$, $C \subset E_n$ for every n, so that $m(C) = 0$.

10.3 Measurable functions and Lebesgue integral

Consider the measure space $(X, \mathcal{F}, \mu)$. The sets in $\mathcal{F}$ are called the **measurable sets**.

Definition 10.8. A function $f : X \to \mathbb{R}$ is **measurable** if

$$f^{\text{pre}}((-\infty, t]) \ = \ \{x \in X : \ f(x) \leq t\}.$$

is measurable for any $t \in \mathbb{R}$.

Exercise 10.5. Show that if f is measurable then $\{x \in X : f(x) < t\}$, $\{x \in X : f(x) \geq t\}$, $\{x \in X : f(x) > t\}$, and $\{x \in X : f(x) = t\}$ are all measurable for all $t \in \mathbb{R}$.

Exercise 10.6. Show that f is measurable if and only if the preimage of any Lebesgue-measurable set in $\mathbb{R}$ is measurable. Compare this with an alternative definition of continuity in Theorem 3.8. It is now clear how to define measurable functions between any two measure spaces, right?

If $\mathcal{F} = \{\emptyset, X\}$ then only the constant functions are measurable. In fact the constant functions are always measurable for any $\mathcal{F}$. If $\mathcal{F}$ consists of all subsets of X, then every function from X to $\mathbb{R}$ is measurable.

Exercise 10.7. Show that if f is measurable then λf and $\lambda + f$ ($\lambda \in \mathbb{R}$) are also measurable.

Exercise 10.8. Show that if f and g are measurable then $\max\{f, g\}$ and $\min\{f, g\}$ are measurable. Conclude that $f^+(x) = \max\{f(x), 0\}$, $f^-(x) = -\min\{f(x), 0\}$, and $|f|(x) = |f(x)|$ are all measurable too.

Proposition 10.5. *If f and g are measurable functions then the three sets*

$$\{x \in X :\ f(x) > g(x)\},\ \{x \in X :\ f(x) \geq g(x)\},\ \{x \in X :\ f(x) = g(x)\}$$

are all measurable.

Proof. Note that $f(x) > g(x)$ is equivalent to $f(x) > r > g(x)$ for some $r \in \mathbb{Q}$. The set

$$\{x \in X : f(x) > g(x)\} \ = \ \bigcup_{r \in \mathbb{Q}} \big(\{x \in X : f(x) > r\} \cap \{x \in X : g(x) < r\}\big)$$

is then measurable as a countable union of measurable sets. It follows that the set

$$\{x \in X : f(x) \geq g(x)\} \ = \ X \setminus \{x \in X : g(x) > f(x)\}$$

is measurable and so it the set

$$\{x \in X : f(x) = g(x)\} \ = \ \{x \in X : f(x) \geq g(x)\} \cap \{x \in X : g(x) \geq f(x)\}.$$

□

Exercise 10.9. Show that if f and g are measurable then $f + g$, and fg are all measurable.

Exercise 10.10. Show that if $\{f_n\}$ is a sequence of measurable functions then $f(x) = \sup_n f_n(x)$ and $g(x) = \inf_n f_n(x)$ are measurable.

Note that measure has not been mentioned in our discussion of measurable functions. In fact, the class of measurable functions on X depends only on the underlying σ-field.

We start our discussion of the Lebesgue integral with the following definition.

Definition 10.9. Let s be a real-valued function defined on X. If the range of s is finite, we say that s is a **simple function**.
Let $A \subset X$ and put

$$\chi_A(x) = \begin{cases} 1 & x \in A, \\ 0 & x \notin A. \end{cases}$$

χ_A is called the **characteristic function of** A.

Suppose the range of s consists of the distinct numbers $c_1, \ldots, c_n$. Let

$$A_i = \{x : \ s(x) = c_i\}.$$

Then

$$s(x) = \sum_{i=1}^{n} c_i \chi_{A_i}(x),$$

that is, every simple function is a finite linear combination of characteristic functions. Note that by construction $A_1, \ldots, A_n$ are disjoint and their union is X, hence it forms a partition of X.

Exercise 10.11. Show that $s(x) = \sum_{i=1}^{n} c_i \chi_{A_i}(x)$ is measurable if and only if each A_i is measurable.

We will use the fact that every function can be approximated in a suitable sense by simple functions.

Theorem 10.3. *Let $f : X \to \mathbb{R}$. There exists a sequence $\{s_n\}$ of simple functions such that $\forall x \in X \ s_n(x) \to f(x)$ as $n \to \infty$. If f is measurable, $\{s_n\}$ may be chosen to be a sequence of measurable functions. If $f \geq 0$, $\{s_n\}$ may be chosen to be a monotonically increasing sequence.*

Proof. If $f \geq 0$ define for every $n \in \mathbb{N}$, $i = 1, 2, \ldots, n2^n$

$$A_{ni} \ = \ \left\{x : \ \tfrac{i-1}{2^n} \leq f(x) < \tfrac{i}{2^n}\right\}, \qquad F_n \ = \ \{x : \ f(x) \geq n\}.$$

Put

$$s_n(x) \ = \ \sum_{i=1}^{n2^n} \tfrac{i-1}{2^n} \chi_{A_{ni}}(x) + n\chi_{F_n}(x).$$

For a fixed x, if n is sufficiently big, we have $|s_n(x) - f(x)| < \frac{1}{2^n}$, which proves pointwise convergence. In the general case, let $f = f^+ - f^-$ and apply the preceding construction to f^+ and f^-. $\square$

We shall now define integration on a measurable space $(X, \mathcal{F}, \mu)$. The main examples are $(\mathbb{R}, \mathcal{B}(\mathbb{R}), m)$ or a probability space. It is hopefully clear how the integral over $A \in \mathcal{B}(\mathbb{R})$ of a simple function s should be defined. Suppose

$$s(x) \;=\; \sum_{i=1}^{n} c_i \chi_{A_i}(x) \qquad (c_i \geq 0)$$

is measurable and suppose $A \in \mathcal{F}$. We define

$$I_A(s) \;:=\; \sum_{i=1}^{n} c_i \mu(A \cap A_i).$$

Exercise 10.12. Let $s, \tilde{s}$ be two nonnegative measurable simple functions on X. Show that $\tilde{s} \leq s$ implies that $s - \tilde{s}$ is also a nonnegative measurable simple function on X. Further, show that $I_A(\tilde{s}) + I_A(s - \tilde{s}) = I_A(s)$. Conclude that $I_A(s + \tilde{s}) = I_A(s) + I_A(\tilde{s})$ for *any* two nonnegative measurable simple functions $s, \tilde{s}$ on X.

Like in the case of the Lebesgue measure, now the idea is to extend this definition to general measurable functions. If f is measurable and nonnegative, we define the **Lebesgue integral of f, with respect to the measure μ, over the set** A as

$$\int_A f \mathrm{d}\mu \;=\; \sup I_A(s),$$

where the sup is taken over all measurable simple functions s such that $0 \leq s \leq f$. Note that the integral may have value $+\infty$.

The following exercise shows that the Lebesgue integral satisfies the first obvious property. Make sure you understand why.

Exercise 10.13. Show that if $s \geq 0$ is a simple measurable function then $\int_A s \mathrm{d}\mu = I_A(s)$.

If f is not nonnegative, then the above construction gives $\int_E f^+ \mathrm{d}\mu$ and $\int_E f^- \mathrm{d}\mu$ (both f^+ and f^- are measurable by Exercise 10.8). If at least one of these integrals is finite we define

$$\int_A f \mathrm{d}\mu \;=\; \int_A f^+ \mathrm{d}\mu - \int_A f^- \mathrm{d}\mu.$$

If both integrals are finite, then $\int_A f \mathrm{d}\mu$ is finite and we say that f is **integrable** (or summable) on A in the Lebesgue sense, with respect to μ; we write $f \in \mathcal{L}(\mu)$ on A. If $\mu = m$, the usual notation is: $f \in \mathcal{L}$ on A.

Exercise 10.14. Let $X = \{1, 2, 3\}$, $\mathcal{F} = 2^X$, and let μ be the counting measure on X. Show that for any function $f : X \to \mathbb{R}$

$$\int_A f \mathrm{d}\mu \;=\; \sum_{i \in A} f(i).$$

The best strategy for proving many properties of integrals is by first showing that they hold for nonnegative basic functions. Then for general nonnegative measurable functions and finally for any measurable function by writing $f = f^+ - f^-$.

Example 10.3. We will show that if $f \in \mathcal{L}(\mu)$ on A then $\int_A cf\mathrm{d}\mu = c\int_A f\mathrm{d}\mu$ for every $c \in \mathbb{R}$. If $f = \sum_i c_i\chi_{A_i}$ is a simple nonnegative and measurable function then so is cf and

$$I_A(cf) = \sum_i cc_i\mu(A \cap A_i) = c\sum_i c_i\mu(A \cap A_i) = cI_A(f).$$

If f is nonnegative then

$$\int_A cf\mathrm{d}\mu = \sup_{s\le cf} I_A(s) = \sup_{s\le f} I_A(cs) = c\sup_{s\le f} I_A(s) = c\int_A f\mathrm{d}\mu.$$

Finally, if f is any measurable function then $\int cf^+\mathrm{d}\mu = c\int f^+\mathrm{d}\mu$ and $\int cf^-\mathrm{d}\mu = c\int f^-\mathrm{d}\mu$ and so

$$\int cf\mathrm{d}\mu = \int cf^+\mathrm{d}\mu - \int cf^-\mathrm{d}\mu = c\int f^+\mathrm{d}\mu - c\int f^-\mathrm{d}\mu = c\int f\mathrm{d}\mu.$$

The proof of the following result will be left as an exercise.

Proposition 10.6. *The following properties of the Lebesgue integral hold.*

(a) If f is measurable and bounded on A, and if $\mu(A) < +\infty$, then $f \in \mathcal{L}(\mu)$.
(b) If $a \le f(x) \le b$ for all $x \in A$, and $\mu(A) < +\infty$, then

$$a\mu(A) \le \int_A f\mathrm{d}\mu \le b\mu(A).$$

(c) If $f, g \in \mathcal{L}(\mu)$ on A, and if $f(x) \le g(x)$ for $x \in A$, then

$$\int_A f\mathrm{d}\mu \le \int_A g\mathrm{d}\mu.$$

(d) If $\mu(A) = 0$ and f is measurable then $\int_A f\mathrm{d}\mu = 0$.
(e) If $f \in \mathcal{L}(\mu)$ on E, $A \in \mathcal{F}$, and $A \subset E$, then $f \in \mathcal{L}(\mu)$ on A.
(f) Suppose f is measurable on A, $|f| \le g$, and $g \in \mathcal{L}(\mu)$ on A. Then $f \in \mathcal{L}(\mu)$ on A.

Two fundamental tools in the theory of Lebesgue integrals are the Lebesgue's monotone convergence theorem and the Lebesgue's dominated convergence theorem. They both provide a list of conditions under which for a sequence (f_n) of measurable functions it holds that

$$\lim_{n\to\infty}\int_A f_n\mathrm{d}\mu = \int_A f\mathrm{d}\mu, \tag{10.1}$$

where $A \in \mathcal{F}$ and f is the pointwise limit of the sequence (f_n), that is, $f(x) = \lim_{n\to+\infty} f_n(x)$.

Theorem 10.4 (Lebesgue's monotone convergence theorem). *If the sequence (f_n) satisfies $0 \le f_1(x) \le f_2(x) \le \cdots$ for all $x \in A$ then (10.1) holds.*

Theorem 10.5 (Lebesgue's dominated convergence theorem). *If there exists a function $g \in \mathcal{L}(\mu)$ on A, such the sequence (f_n) satisfies $|f_n(x)| \le g(x)$ for all $n \in \mathbb{N}$ and $x \in A$ then (10.1) holds.*

Suppose $A = [a, b]$, μ is the Lebesgue measure, and $\mathcal{F} = \mathcal{B}(\mathbb{R})$. Instead of $\int_A f \mathrm{d}\mu$ it is customary to use the familiar notation

$$\int_a^b f(x)\mathrm{d}x.$$

Recycling the standard notation introduces no ambiguity: In this case when the standard (Riemann) integral exists the Lebesgue integral also exists and both integrals are equal. However, the Lebesgue integral is defined for a much wider class of functions.

Example 10.4. Let f be a function on $\mathbb{R}$ such that $f(x) = 1$ is $x \in \mathbb{Q}$ and $f(x) = 0$ if $x \in \mathbb{R} \setminus \mathbb{Q}$. The Riemann integral does not exists. The fact that $\mathbb{Q}$ is countable implies both that it is measurable and that the Lebegue integral exists and it is equal to zero.

10.4 Probability spaces

Suppose $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space then a measurable function $Y : \Omega \to \mathbb{R}$ is called a **random variable**. By definition, the function $F_Y : \mathbb{R} \to [0, 1]$ given by

$$F_Y(t) \;=\; \mathbb{P}(\{\omega \in \Omega : Y(\omega) \le t\}) \;=\; \mathbb{P}(Y \le t),$$

is well defined. The function F_Y is called the **cumulative distribution function**. The mapping $Y : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ allows to transport probability measure $\mathbb{P}$ from an abstract space Ω to $\mathbb{R}$.

Definition 10.10. A **probability distribution on** $\mathbb{R}$ is any probabilistic measure μ on $\mathcal{B}(\mathbb{R})$. A **probability distribution of a real-valued random variable** Y is the probability distribution μ_Y, defined on $\mathcal{B}(\mathbb{R})$ by

$$\mu_Y(B) \;=\; \mathbb{P}(\{\omega \in \Omega : Y(\omega) \in B\}) \;=\; \mathbb{P}(Y \in B), \qquad B \in \mathcal{B}(\mathbb{R}).$$

One of the most fundamental results of probability relies on the following exercise.

Exercise 10.15. Show that $\mathcal{B}(\mathbb{R})$ is generated by intervals of the form $(-\infty, t]$ for $t \in \mathbb{R}$. Conclude that the cumulative distribution function F_Y uniquely identifies the distribution μ_Y.

Definition 10.11. The σ-field generated by the sets $\{\omega : Y(\omega) \leq t\}$ is called the **σ-field generated by** Y. This is the smallest σ-field with respect to which Y is a measurable function. We write

$$\sigma(Y) \;=\; \{\{\omega \in \Omega : Y(\omega) \in A\} :\; A \in \mathcal{B}(\mathbb{R})\}.$$

Definition 10.12. If μ is a probability distribution on $\mathbb{R}$ and for some function $f : \mathbb{R} \to \mathbb{R}$, $f \in \mathcal{L}$, we have

$$\mu(A) \;=\; \int_A f(x)\mathrm{d}x, \qquad A \in \mathcal{B}(\mathbb{R}),$$

then f is called the **density function** of the distribution μ. A probability distribution that has a density is called **continuous**.

Let X be a nonempty set, and let $\mathcal{F}$ be a σ-algebra of subsets of X. If μ and ν are measures defined on $\mathcal{F}$, we say that ν is **absolutely continuous** with respect to μ, written $\nu \ll \mu$, if $\mu(A) = 0$ implies $\nu(A) = 0$.

Exercise 10.16. Show that a probability distribution of a continuous random variable is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}$.

Theorem 10.6 (Radon-Nikodym Theorem). *Let $(X, \mathcal{F}, \mu)$ be a σ-finite measure space, and let ν be a σ-finite measure on $\mathcal{F}$ with $\nu \ll \mu$. Then there exists a measurable $f \geq 0$ such that $\nu(A) = \int_A f\mathrm{d}\mu$ for all $A \in \mathcal{F}$, and f is unique up to a set of μ-measure zero.*

One of the main applications of the Radon-Nikodym theorem is to assure existence of density functions for continuous random variables. Here μ is the Lebesgue measure on the image of Y and ν is the

Exercise 10.17. Show that if the probability distribution μ has an atom, that is, $\mu(\{y\}) > 0$ for some $y \in \mathbb{R}$, then μ is *not* absolutely continuous with respect to the Lebesgue measure.

Chapter 11
Convex geometry (2 lectures)

Convexity plays an important role in modern statistics and economics. One of the prominent examples is the importance of convex optimization. Although convex optimization is beyond the scope of these lectures, we want to present basics of convex geometry. We will present applications in game theory.

11.1 Convex sets

A set $C \subseteq \mathbb{R}^k$ is *convex* if for any two points $x, y \in C$

$$\mathbf{z}_\lambda \ := \ (1-\lambda)\mathbf{x} + \lambda\mathbf{y} \ \in \ C$$

for all $\lambda \in (0,1)$. The point $\mathbf{z}_\lambda$ can be rewritten as $\mathbf{z}_\lambda = \mathbf{x} + \lambda(\mathbf{y} - \mathbf{x})$ so, as λ varies from 0 to 1, $\mathbf{z}_\lambda$ moves from $\mathbf{x}$ to $\mathbf{y}$ along the segment joining $\mathbf{x}$ and $\mathbf{y}$. This gives a geometric interpretation of convex sets: these are sets such that for any two points in the set, the segment between these two points is contained in the set.

Definition 11.1. A linear combination $\lambda_0\mathbf{x}_0 + \lambda_1\mathbf{x}_1 + \cdots + \lambda_n\mathbf{x}_n$ of vectors $\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_n$ in $\mathbb{R}^k$ is called an **affine combination** if the coefficient satisfy $\lambda_0 + \lambda_1 + \cdots + \lambda_n = 1$.

For example, the set of all affine combinations of any two given vectors $\mathbf{x}_1, \mathbf{x}_2$ is the line crossing $\mathbf{x}_1, \mathbf{x}_2$.

Definition 11.2. A **convex combination** of $\mathbf{x}_0, \ldots, \mathbf{x}_n$ is an affine combination where, in addition, $\lambda_0, \lambda_1, \ldots, \lambda_n \geq 0$. A convex combination is *strictly positive* if all coefficients are strictly positive.

We write Δ_{n+1} for the set of all convex coefficients and Δ°_{n+1} for its subset of strictly positive coefficients. In symbols

$$\Delta_{n+1} = \{(\lambda_0, \ldots, \lambda_n) \in \mathbb{R}^{n+1} : \ \lambda_0, \ldots, \lambda_n \geq 0 \text{ and } \sum_{i=0}^{n} \lambda_i = 1\}, \quad (11.1)$$

and $\Delta_{n+1}^{\circ} = \{(\lambda_0, \ldots, \lambda_n) \in \Delta_{n+1} : \ \lambda_0, \ldots, \lambda_n > 0\}$. The set Δ_{n+1} is called the standard n-simplex.

Given a subset $A \subset \mathbb{R}^k$, its **convex hull** $\mathrm{conv}(A)$ is a subset of $\mathbb{R}^k$ given by all (finite) convex combinations of elements of A. More precisely $\mathrm{conv}(A)$ is the set of points $\mathbf{x}$ such that for some $n \in \mathbb{N}$, points $\mathbf{x}_0, \ldots, \mathbf{x}_n \in A$, and coefficients $(\lambda_0, \ldots, \lambda_n) \in \Delta_{n+1}$

$$\mathbf{x} \ = \ \lambda_0 \mathbf{x}_0 + \lambda_2 \mathbf{x}_2 + \cdots + \lambda_n \mathbf{x}_n.$$

Example 11.1. The standard n-simplex $\Delta_{n+1} \subset \mathbb{R}^{n+1}$ is the convex hull of $\{\mathbf{e}_0, \ldots, \mathbf{e}_n\}$, the canonical basis of $\mathbb{R}^n$

$$\Delta_{n+1} \ = \ \mathrm{conv}\{\mathbf{e}_0, \ldots, \mathbf{e}_n\}. \quad (11.2)$$

Exercise 11.1. Show that the intersection of an arbitrary family of convex sets is convex.

Exercise 11.2. For any two sets $C, D \subset \mathbb{R}^k$, define their Minkowski sum $C + D$ as

$$C + D \ = \ \{\mathbf{x} + \mathbf{y} \in \mathbb{R}^k : \ \mathbf{x} \in C, \mathbf{y} \in D\}.$$

Show that, if C, D are convex then $C + D$ is.

Theorem 11.1 (Caratheodory's Theorem). *Let $A \subset \mathbb{R}^k$. If $\mathbf{x} \in \mathrm{conv}(A)$ then $\mathbf{x}$ can be written as a convex combination of* no more than $k+1$ *points in A.*

Proof. Suppose that $\mathbf{x} \in \mathrm{conv}(A)$ and $\mathbf{x} = \sum_{i=0}^{n} \lambda_i \mathbf{x}_i$ for $n > k$ (otherwise there is nothing to prove), where λ_i are nonnegative and sum to 1. Then $\mathbf{x}_1 - \mathbf{x}_0, \ldots, \mathbf{x}_n - \mathbf{x}_0$ are linearly dependent so there exist $\alpha_1, \ldots, \alpha_n$ not all zero such that

$$\alpha_1(\mathbf{x}_1 - \mathbf{x}_0) + \cdots + \alpha_n(\mathbf{x}_n - \mathbf{x}_0) \ = \ \mathbf{0}_k.$$

If $\alpha_0 := -(\alpha_1 + \cdots + \alpha_n)$ then

$$\alpha_0 \mathbf{x}_0 + \alpha_1 \mathbf{x}_1 + \cdots + \alpha_n \mathbf{x}_n \ = \ \mathbf{0}_k.$$

Because $\alpha_0 + \alpha_1 + \cdots + \alpha_n = 0$, at least one of these numbers is positive. We have

$$\mathbf{x} \ = \ \sum_{i=0}^{n} \lambda_i \mathbf{x}_i - t \sum_{i=0}^{n} \alpha_i \mathbf{x}_i \ = \ \sum_{i=1}^{n} (\lambda_i - t\alpha_i) \mathbf{x}_i.$$

Since it holds for any t, take

$$t^* \;=\; \min_i \left\{ \tfrac{\lambda_i}{\alpha_i} :\; \alpha_i > 0 \right\} \;=\; \tfrac{\lambda_j}{\alpha_j}.$$

Note that $t^* > 0$ and for all $i = 0, \ldots, n$ $(\lambda_i - t^*\alpha_i) \geq 0$ and $\lambda_j - t^*\alpha_j = 0$. This shows that $\mathbf{x} = \sum_i (\lambda_i - t^*\alpha_i)x_i$ is a convex combination that has at most $n-1$ terms with nonzero coefficients. □

Proposition 11.1. *If $A \subset \mathbb{R}^k$ is open, then* $\mathrm{conv}(A)$ *is open.*

Proof. We will show that every $\mathbf{x} \in \mathrm{conv}(A)$ is an interior point. By definition $\mathbf{x} = \lambda_0\mathbf{x}_0 + \cdots + \lambda_n\mathbf{x}_n$ for some n and points $\mathbf{x}_0, \ldots, \mathbf{x}_n \in A$. Since A is open, for each $i = 0, \ldots, n$ there exist open neighbourhoods $N_{r_i}(\mathbf{x}_i) \subset A$. Take $r = \min\{r_0, \ldots, r_n\}$. If $\|\mathbf{y}\| < r$ then $\mathbf{x}_i + \mathbf{y} \in N_{r_i}(\mathbf{x}_i)$ and so they all lie in A. Therefore

$$\mathbf{x} + \mathbf{y} = (\sum_i \lambda_i \mathbf{x}_i) + (\sum_i \lambda_i)\mathbf{y} = \sum_i \lambda_i(\mathbf{x}_i + \mathbf{y}) \;\in\; \mathrm{conv}(A).$$

In other words $N_r(\mathbf{x}) \subset \mathrm{conv}(A)$. □

If A is closed then $\mathrm{conv}(A)$ is not necessarily closed. As an example consider

$$A = \left\{ (x, y) \in \mathbb{R}^2 :\; y \geq \left|\tfrac{1}{x}\right| \right\}.$$

A is closed but $\mathrm{conv}(A) = \mathbb{R} \times (0, \infty)$ is not. Nevertheless, we have the following theorem.

Theorem 11.2. *If $A \subset \mathbb{R}^k$ is compact then* $\mathrm{conv}(A)$ *is compact.*

Proof. By Caratheodory's theorem the map $\Delta_{k+1} \times A^{k+1} \to \mathrm{conv}(A)$ given by

$$(\lambda_0, \ldots, \lambda_k, \mathbf{x}_0, \ldots, \mathbf{x}_k) \;\mapsto\; \lambda_0\mathbf{x}_0 + \cdots + \lambda_k\mathbf{x}_k$$

is onto. It is also continuous as a polynomial (quadratic) map. By Corollary 4.1 a Cartesian product of compact spaces is compact and so $\Delta_{k+1} \times A^{k+1}$ is compact if A is. By Theorem 4.10 this then implies that $\mathrm{conv}(A)$ is compact. □

Exercise 11.3. Show that if C is convex, then the closure $\overline{C}$ and the interior C° are also convex.

Definition 11.3. The points $\mathbf{x}_0, \ldots, \mathbf{x}_n \in \mathbb{R}^k$ are affinely independent if $\lambda_0\mathbf{x}_0 + \lambda_1\mathbf{x}_1 + \ldots + \lambda_n\mathbf{x}_n = \mathbf{0}_k$ and $\lambda_0 + \lambda_1 + \ldots + \lambda_n = 0$ imply that $\lambda_0 = \lambda_1 = \ldots = \lambda_n = 0$.

Exercise 11.4. Show that $\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^k$ are affinely independent if and only if $\mathbf{x}_1 - \mathbf{x}_0, \ldots, \mathbf{x}_n - \mathbf{x}_0$ are linearly independent. In particular, if $\mathbf{x}_0, \ldots, \mathbf{x}_n$ are affinely independent then $n \leq k$.

The convex hull of a collection of $n+1$ affinely independent points is called an n**-simplex**. For example, a 0-simplex is a point, and a 1-simplex is a segment joining two points. An important example is given by the standard n-simplex in $\Delta_{n+1} \subset \mathbb{R}^{n+1}$, see (11.2).

The following important result states that the representation of any $\mathbf{y}$ in an n-simplex as a convex combination of its generators $\mathbf{y} = \sum_i \lambda_i \mathbf{x}_i$ is unique. These λ_i's are called the **barrycentric coordinates** of $\mathbf{y}$.

Theorem 11.3. *Let $\mathbf{x}_0, \ldots, \mathbf{x}_n$ be affinely independent. For any $\mathbf{y}$ in the convex hull $\operatorname{conv}(\{\mathbf{x}_0, \ldots, \mathbf{x}_n\})$ there is a unique set of convex coefficients $\lambda_0, \ldots, \lambda_n$ such that $\mathbf{y} = \sum_{i=0}^n \lambda_i \mathbf{x}_i$.*

Proof. Assume that $\mathbf{y} = \sum_i \lambda_i \mathbf{x}_i = \sum_i \mu_i \mathbf{x}_i$. Then $\sum_i (\lambda_i - \mu_i)\mathbf{x}_i = \mathbf{0}_k$ and $\sum_i (\lambda_i - \mu_i) = 0$ $(1 - 1 = 0)$ so by affine independence $\lambda_i = \mu_i$ for all i. □

It is a simple exercise to show that every n-simplex $S = \operatorname{conv}\{\mathbf{x}_0, \ldots, \mathbf{x}_n\} \subset \mathbb{R}^k$ is homeomorphic to the standard n-simplex. This is such an essential result that we formulate it as a theorem.

Theorem 11.4. *Any n-simplex $T = \operatorname{conv}(\{\mathbf{x}_0, \ldots, \mathbf{x}_n\}) \subset \mathbb{R}^k$ is homeomorphic to Δ_{n+1}.*

Proof. For any $\lambda = (\lambda_0, \ldots, \lambda_n) \in \Delta_{n+1}$ define

$$h(\lambda) \;=\; \lambda_0 \mathbf{x}_0 + \ldots + \lambda_n \mathbf{x}_n.$$

This map is linear and so continuous. It is bijective by Theorem 11.3. The inverse is continuous by Theorem 4.12 because Δ_{n+1} is compact. □

If $S = \operatorname{conv}(\{\mathbf{x}_0, \ldots, \mathbf{x}_n\})$ is an n-simplex and $\{i_0, \ldots, i_m\} \subset \{0, \ldots, n\}$ then $\operatorname{conv}(\{\mathbf{x}_{i_0}, \ldots, \mathbf{x}_{i_m}\})$ is called a (m-dimensional) face of S. Theorem 11.3 implies that for each $\mathbf{y} \in S$ we can read off the the smallest face containing $\mathbf{y}$ uniquely from its barricentric coordinates.

11.2 Minimum distance and separation

Given a subset $C \subset \mathbb{R}^k$ we define the **distance to** C function $d_C : \mathbb{R}^k \to \mathbb{R}$ by

$$d_C(\mathbf{x}) \;:=\; \inf_{\mathbf{y} \in C} \|\mathbf{x} - \mathbf{y}\|.$$

This function is well defined because for every $\mathbf{x} \in \mathbb{R}^k$ the set $\{\|\mathbf{x} - \mathbf{y}\| : \mathbf{y} \in C\} \subset \mathbb{R}$ is bounded from below (by zero) and so its infimum is well-defined. The following result is fundamental for many applications of convex analysis.

Theorem 11.5 (Minimum distance to a set).

(1) Let $E, F \subset \mathbb{R}^k$. Then $d_F : E \to \mathbb{R}$ is a continuous function.
(2) If F is closed, then

$$\forall \mathbf{x} \in E \ \ \exists \mathbf{y} \in F \text{ such that } \ \ d_F(\mathbf{x}) = \|\mathbf{x} - \mathbf{y}\|.$$

(3) If F is also convex, for every $\mathbf{x}$ such $\mathbf{y}$ is unique in F.

Proof. (1) Let $\mathbf{x}_1, \mathbf{x}_2 \in E$. By the triangle inequality, $\|\mathbf{x}_1 - \mathbf{y}\| \leq \|\mathbf{x}_1 - \mathbf{x}_2\| + \|\mathbf{x}_2 - \mathbf{y}\|$. If $\mathbf{y} \in F$ then $d_F(\mathbf{x}_1) \leq \|\mathbf{x}_1 - \mathbf{y}\|$ and so

$$d_F(\mathbf{x}_1) \ \leq \ \|\mathbf{x}_1 - \mathbf{x}_2\| + \|\mathbf{x}_2 - \mathbf{y}\|.$$

Since the inequality holds for every $\mathbf{y} \in F$. Take infimum over all $\mathbf{y} \in F$ to get that $d_F(\mathbf{x}_1) \leq \|\mathbf{x}_1 - \mathbf{x}_2\| + d_F(\mathbf{x}_2)$. In the same way, starting with $\|\mathbf{x}_2 - \mathbf{y}\| \leq \|\mathbf{x}_1 - \mathbf{x}_2\| + \|\mathbf{x}_1 - \mathbf{y}\|$, conclude that $d_F(\mathbf{x}_2) \leq \|\mathbf{x}_1 - \mathbf{x}_2\| + d_F(\mathbf{x}_1)$. It follows that $|d_F(\mathbf{x}_1) - d_F(\mathbf{x}_2)| \leq \|\mathbf{x}_1 - \mathbf{x}_2\|$, which implies continuity of d_F (c.f. Exercise 3.3).

(2) Let $\mathbf{x} \in E$ and fix $\mathbf{y}_0 \in F$. We have $f(\mathbf{x}) \leq \|\mathbf{x} - \mathbf{y}_0\| = r$. Define $\tilde{F} = F \cap \overline{N_r(\mathbf{x})}$. Since F is closed and $\overline{N_r(\mathbf{x})}$ is compact, Theorem 4.5 implies that $\tilde{F}$ is compact. Since $\|\mathbf{x} - \mathbf{y}\|$ is a continuous function of $\mathbf{y}$ (c.f. Exercise **??**), Theorem 4.11 implies that there exists $\mathbf{y}_1 \in \tilde{F}$ such that $\inf_{\mathbf{y} \in F} \|\mathbf{x} - \mathbf{y}\| = \|\mathbf{x} - \mathbf{y}_1\|$.

(3) Let $\mathbf{y}_1$ and $\mathbf{y}_2$ in F be such that $\|\mathbf{x} - \mathbf{y}_1\| = \|\mathbf{x} - \mathbf{y}_2\| = d_F(\mathbf{x})$. Define $\mathbf{p} = \mathbf{y}_2 - \mathbf{y}_1$ and $h : [0, 1] \to \mathbb{R}$ by $h(\lambda) = \|\mathbf{x} - \mathbf{y}_1 - \lambda \mathbf{p}\|^2$. We have $h(0) = h(1)$ and also, because F is convex, $\mathbf{y}_1 + \lambda \mathbf{p} \in F$ for $\lambda \in [0, 1]$ and so h is *minimized* at $\lambda = 0$ and $\lambda = 1$. Since $h(\lambda)$ is a quadratic function with nonnegative coefficient $\|\mathbf{y}_1 - \mathbf{y}_2\|^2$ of λ^2, this is only possible if $\mathbf{y}_1 = \mathbf{y}_2$. □

Proposition 11.2. *Let $E, F \subset \mathbb{R}^k$ with F closed and convex. Let $g : E \to F$ be given by $g(\mathbf{x}) = \arg\inf_{\mathbf{y} \in F} \|\mathbf{x} - \mathbf{y}\|$. Then g is a well-defined and*

$$\|g(\mathbf{x}_2) - g(\mathbf{x}_1)\| \ \leq \ \|\mathbf{x}_2 - \mathbf{x}_1\| \qquad \text{for all } \mathbf{x}_1, \mathbf{x}_2 \in E. \tag{11.3}$$

In particular, g is a continuous function.

Proof. Because F is closed and convex, Theorem 11.5 assures that g is a well-defined function, that is, for each $\mathbf{x} \in E$ there is a unique $\mathbf{y} \in F$ such that $g(\mathbf{x}) = \mathbf{y}$. To show (11.3), take $\mathbf{p} = g(\mathbf{x}_2) - g(\mathbf{x}_1)$; c.f. Figure 11.1. The set of points $g(\mathbf{x}_1) + t\mathbf{p}$ for $t \in [0, 1]$ lies in F, by convexity, and so

$$h(t) = \|\mathbf{x}_1 - g(\mathbf{x}_1) - t\mathbf{p}\|^2$$

has a minimum at $t = 0$. This is a quadratic function in t with a strictly positive coefficient of t^2. The only way for such a function to have a minimum at $t = 0$ is that its derivative at $t = 0$ is nonnegative, or, in other words, the coefficient of t is nonnegative. This coefficient is $-2\langle \mathbf{p}, \mathbf{x}_1 - g(\mathbf{x}_1) \rangle$, which

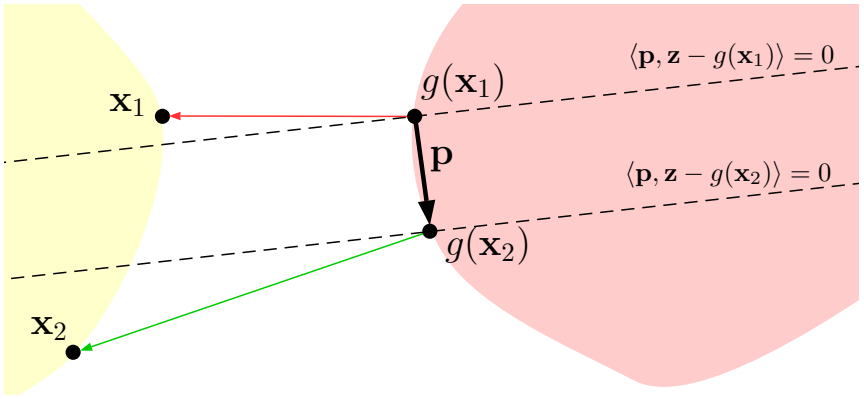

Fig. 11.1 Illustration of the proof of Proposition 11.2

implies that $\langle \mathbf{p}, \mathbf{x}_1 - g(\mathbf{x}_1)\rangle \leq 0$. In a similar way, we show that $\langle \mathbf{p}, \mathbf{x}_2 - g(\mathbf{x}_2)\rangle \geq 0$. But these two inequalities imply that

$$\langle \mathbf{p}, \mathbf{x}_2 - \mathbf{x}_1\rangle \;\geq\; \langle \mathbf{p}, g(\mathbf{x}_2) - g(\mathbf{x}_1)\rangle \;=\; \|\mathbf{p}\|^2.$$

The Cauchy-Schwarz inequality gives that $\langle \mathbf{p}, \mathbf{x}_2 - \mathbf{x}_1\rangle \leq \|\mathbf{p}\|\|\mathbf{x}_2 - \mathbf{x}_1\|$. This implies that $\|\mathbf{p}\| \leq \|\mathbf{x}_2 - \mathbf{x}_1\|$, which is precisely (11.3).

□

Definition 11.4. Let $\mathbf{p} \in \mathbb{R}^k$ be such that $\mathbf{p} \neq \mathbf{0}_k$, and $c \in \mathbb{R}$. The corresponding **hyperplane** in $\mathbb{R}^k$ is the set

$$\{\mathbf{x} \in \mathbb{R}^k : \langle \mathbf{p}, \mathbf{x}\rangle = c\}.$$

The **open half-space** is the set

$$\{\mathbf{x} \in \mathbb{R}^k : \langle \mathbf{p}, \mathbf{x}\rangle < c\},$$

and the **closed half-space** is the set

$$\{\mathbf{x} \in \mathbb{R}^k : \langle \mathbf{p}, \mathbf{x}\rangle \leq c\}.$$

We say that two sets $A, B \subset \mathbb{R}^k$ are strictly separated by a hyperplane if there exists $\mathbf{p} \neq \mathbf{0}_k$ and $c \in \mathbb{R}$ such that for every $\mathbf{x} \in A$, $\mathbf{y} \in B$ we have $\langle \mathbf{p}, \mathbf{x}\rangle < c < \langle \mathbf{p}, \mathbf{y}\rangle$. The following theorem is one of the most important results of convex geometry.

Theorem 11.6. *Let C and K be disjoint non-empty convex sets in $\mathbb{R}^k$. Let C be closed and K compact. Then C and K are strictly separated by a hyperplane.*

Note that compactness of K is necessary. For example $A = \{\mathbf{x} \in \mathbb{R}^2 : x_1 \leq 0\}$, $B = \{\mathbf{x} \in \mathbb{R}^2 : x_1 > 0, x_2 \geq 1/x_1\}$ are closed but not strictly separated.

Proof. By Theorem 11.5, $d_C(\mathbf{x})$ is a continuous real-valued function on K and so, by Theorem 4.11, it achieves its minimum. Call it $\mathbf{x}_0 \in K$. By Theorem 11.5(c) there is exactly one $\mathbf{y}_0 \in C$ such that $d_C(\mathbf{x}_0) = \|\mathbf{x}_0 - \mathbf{y}_0\|$. Set $\mathbf{p} = \mathbf{x}_0 - \mathbf{y}_0$. Then $\mathbf{p} \neq \mathbf{0}_k$ and $0 < \|\mathbf{p}\|^2 = \langle \mathbf{p}, \mathbf{x}_0 - \mathbf{y}_0 \rangle$ so

$$\langle \mathbf{p}, \mathbf{x}_0 \rangle \ > \ \langle \mathbf{p}, \mathbf{y}_0 \rangle$$

so it suffices to show that $\langle \mathbf{p}, \mathbf{x} \rangle \geq \langle \mathbf{p}, \mathbf{x}_0 \rangle$ for every $\mathbf{x} \in K$ and $\langle \mathbf{p}, \mathbf{y} \rangle \leq \langle \mathbf{p}, \mathbf{y}_0 \rangle$ for all $\mathbf{y} \in C$. We show the second, the first is similar and follows from convexity. Let $\mathbf{y} \in C$ and set $\mathbf{y}_\lambda = (1-\lambda)\mathbf{y}_0 + \lambda \mathbf{y}$. Then

$$\mathbf{x}_0 - \mathbf{y}_\lambda \ = \ \mathbf{x}_0 - \mathbf{y}_0 - \lambda(\mathbf{y} - \mathbf{y}_0) \ = \ \mathbf{p} - \lambda(\mathbf{y} - \mathbf{y}_0)$$

and so

$$\begin{aligned} \|\mathbf{x}_0 - \mathbf{y}_\lambda\|^2 = \|\mathbf{p} - \lambda(\mathbf{y} - \mathbf{y}_0)\|^2 = \\ = \lambda^2 \|\mathbf{y} - \mathbf{y}_0\|^2 - 2\lambda \langle \mathbf{p}, \mathbf{y} - \mathbf{y}_0 \rangle + \|\mathbf{p}\|^2. \end{aligned}$$

This is a quadratic function of λ that achieves its minimum at $\lambda = 0$ (by construction!). This implies that the derivative of this function at zero must be nonnegative. This derivative is equal to the coefficient of λ which is $2\langle \mathbf{p}, \mathbf{y}_0 - \mathbf{y} \rangle$. This implies that $\langle \mathbf{p}, \mathbf{y} \rangle \leq \langle \mathbf{p}, \mathbf{y}_0 \rangle$ as claimed. □

Exercise 11.5. Show that each closed convex set is an intersection of closed half-spaces.

11.3 Application: Von Neumann's theorem

Formulation of von Neumann's theorem marks the beginning of the game theory. We will formulate the simplest possible version introducing first some basic notation of the game theory. Let $n \geq 2$ will be the number of players in a game. Let X_i be the set of actions available to player i. The outcome is

$$(x_1, \ldots, x_n) \in X_1 \times \cdots \times X_n =: X.$$

We focus on strategic games, where payoffs depend on actions of other players. Payoff of the i-th player is $\pi_i : X \to \mathbb{R}$.

Suppose $n = 2$. We call the two players 'player I' and 'player II'. A zero sum game is when π_1, π_2 satisfy

$$\pi_1(x_1, x_2) \ = \ -\pi_2(x_1, x_2).$$

A mixed strategy involves some kind of randomness. From now on assume X_1, X_2 are finite with m and n elements respecively. Without loss of generality let $X_1 = \{1, \dots, m\}$, $X_2 = \{1, \dots, n\}$. Then a pair of mixed strategies for the two players is $\mathbf{p} \in \Delta_m$ $\mathbf{q} \in \Delta_n$. The payoff is given by the $m \times n$ matrix $A = [a_{ij}]$, where $a_{ij} = \pi_1(i, j)$. We focus on the zero-sum games and so the payoffs of player I are given by $-A$. The expected payoff of player I is

$$f(\mathbf{p}, \mathbf{q}) \;=\; \sum_{i=1}^{m} \sum_{j=1}^{n} p_i q_j a_{ij} \;=\; \mathbf{p}^T A \mathbf{q}.$$

Player I has control over $\mathbf{p}$ but not over $\mathbf{q}$. She wants to maximize $f(\mathbf{p}, \mathbf{q})$.

Theorem 11.7 (Minimax theorem for two-person zero-sum game). *For any matrix $A \in \mathbb{R}^{m \times n}$*

$$V \;=\; \max_{\mathbf{p} \in \Delta_m} \big(\min_{\mathbf{q} \in \Delta_m} f(\mathbf{p}, \mathbf{q}) \big) \;=\; \min_{\mathbf{q} \in \Delta_n} \big(\max_{\mathbf{p} \in \Delta_m} f(\mathbf{p}, \mathbf{q}) \big).$$

Remark 11.1. Note that the minima and maxima are achieved by Theorem 4.11.

Remark 11.2. You can read this theorem as follows: There is a mixed strategy for player I such that her average gain is at least V no matter what player II does.

To prove this theorem we will need a lemma.

Lemma 11.1. *For any function*

$$\sup_{\mathbf{p}} \inf_{\mathbf{q}} f(\mathbf{p}, \mathbf{q}) \le \inf_{\mathbf{q}} \sup_{\mathbf{p}} f(\mathbf{p}, \mathbf{q}).$$

Proof. For all $\mathbf{p}, \mathbf{q}'$ we have $f(\mathbf{p}, \mathbf{q}') \ge \inf_{\mathbf{q}} f(\mathbf{p}, \mathbf{q})$ so for all $\mathbf{q}'$

$$\sup_{\mathbf{p}} f(\mathbf{p}, \mathbf{q}') \ge \sup_{\mathbf{p}} \inf_{\mathbf{q}} f(\mathbf{p}, \mathbf{q}),$$

which implies the statement. □

Proof of the Theorem. By the previous lemma, it remains to prove that

$$V_{\min} := \max_{\mathbf{p} \in \Delta_m} \big(\min_{\mathbf{q} \in \Delta_m} f(\mathbf{p}, \mathbf{q}) \big) \;\;\ge\;\; \min_{\mathbf{q} \in \Delta_n} \big(\max_{\mathbf{p} \in \Delta_m} f(\mathbf{p}, \mathbf{q}) \big) =: V_{\max}$$

Note that

$$V_{\max} = \min_{\mathbf{q}} \max_{i=1,\dots,m} (A\mathbf{q})_i \qquad V_{\min} = \max_{\mathbf{p}} \min_{j=1,\dots,n} (\mathbf{p}^T A)_j.$$

To prove $V_{\min} \ge V_{\max}$ we show $V_{\min} < t < V_{\max}$ is impossible for any $t \in R$. Let $A_0 = [A - t\mathbf{1}\mathbf{1}^T | \mathbb{I}_m]$ be an $m \times (n+m)$ matrix. Let B be the convex hull

of its columns. Then either: (i) $\mathbf{0} \in B$, or (ii) $\mathbf{0} \notin B$.
If $\mathbf{0} \in B$ then there exists $(\boldsymbol{\lambda}, \boldsymbol{\mu}) \in \Delta_{m+n}$ with $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_n)$, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n))$ such that

$$A_0 \begin{bmatrix} \boldsymbol{\lambda} \\ \boldsymbol{\mu} \end{bmatrix} \;=\; A\boldsymbol{\lambda} - t\mathbf{1}\mathbf{1}^T\boldsymbol{\lambda} + \boldsymbol{\mu} \;=\; \mathbf{0}_m.$$

Since $\boldsymbol{\mu} \geq \mathbf{0}_m$, it follows that $A\boldsymbol{\lambda} \leq t\mathbf{1}\mathbf{1}^T\boldsymbol{\lambda}$. The above equation also implies that $\boldsymbol{\lambda} \neq \mathbf{0}_n$ and so $\mathbf{1}^T\boldsymbol{\lambda} > 0$. Defining $\mathbf{q} = \frac{1}{\mathbf{1}^T\boldsymbol{\lambda}}\boldsymbol{\lambda}$ we get $A\mathbf{q} \leq t\mathbf{1}$. So there exists $\mathbf{q}$ such that $\max_{i=1,\ldots,m}(A\mathbf{q})_i \leq t$. This implies that $V_{\max} \leq t$ and so in particular $V_{\min} < t < V_{\max}$ is impossible.
Now suppose $\mathbf{0} \notin B$ then we may use the separating hyperplane theorem with $K = \{\mathbf{0}\}$, $C = B$. According to this theorem there exists $\mathbf{r} \in \mathbb{R}^m$ such that for all columns $\mathbf{a}$ of A_0: $\langle \mathbf{a}, \mathbf{r} \rangle > 0$. Equivalently the vector

$$\mathbf{r}^T A_0 \;=\; [\mathbf{r}^T A - t\mathbf{r}^T\mathbf{1}\mathbf{1}^T \mid \mathbf{r}^T]$$

has only positive entries. In particular, $\mathbf{r}$ has only positive entries and

$$\mathbf{r}^T A \;>\; t\mathbf{r}^T\mathbf{1}\mathbf{1}^T.$$

Let $\mathbf{p} = \frac{\mathbf{r}}{\mathbf{r}^T\mathbf{1}}$. Then $\mathbf{p}^T A > t\mathbf{1}^T$ and so there exists $\mathbf{p} \in \Delta_m$ such that $\min_j(\mathbf{p}^T A)_j > t$, which implies $V_{\min} > t$. In particular $V_{\min} < t < V_{\max}$ is impossible. Then it must be that $V_{\min} \geq V_{\max}$. □

Chapter 12
Brouwer's fixed point theorem (2 lectures)

In Section 8.3 we formulated the Banach fixed point theorem for contractions on a compact set. In this section we provide other useful fixed point theorems and we will illustrate their possible application in economics. We aim in this chapter at proving the following result.

Theorem 12.1 (Brouwer's fixed point theorem). *Let $K \subset \mathbb{R}^k$ be a convex and compact set and let $f : K \to K$ be a continuous function. Then f has a fixed point.*

In general, the proof of this result is long. However, if $k = 1$ the result is elementary: Let $f : [a, b] \to [a, b]$ and define $g(x) = f(x) - x$. Then $g(a) \geq 0$, $g(b) \leq 0$ and g is continuous, so by the intermediate value theorem $g(a) = 0$ for some a.

The proof's idea: We will reduce the proof of the Brouwer's fixed point theorem to the proof of Theorem 12.3, where K is replaced with the standard n-simplex $\Delta_{n+1} = \text{conv}\{\mathbf{e}_0, \ldots, \mathbf{e}_n\}$. For this we will need a sequence of finer and finer subdivisions of the standard simplex into smaller subsimplices such that the diameters of the subsimplices converge to zero. Sperner's lemma will make sure that there is a descending sequence of smaller and smaller subsimplices $T^{(m)} = \text{conv}\{\mathbf{p}_0^{(m)}, \ldots, \mathbf{p}_n^{(m)}\}$ with the property that the i-th coordinate of $\mathbf{p}_i^{(m)}$ satisfies $(\mathbf{p}_i^{(m)})_i \geq (f(\mathbf{p}_i^{(m)}))_i$ ($f(\mathbf{p}_i^{(m)}$ is "further" from $\mathbf{e}_i$ then $\mathbf{p}_i^{(m)}$). Taking $m \to \infty$ we argue this is possible only if $f(\mathbf{z}) = \mathbf{z}$ where $\mathbf{z}$ is the point satisfying $\bigcap_m T^{(m)} = \{\mathbf{z}\}$.

12.1 Simplicial subdivisions of a simplex

Although there are many ways to do define a simplicial subdivision we will stick to one very simple construction.

Definition 12.1 (Very rough). Consider a subdivision of Δ_{n+1} obtained by subdividing each edge into two equal parts. This gives 2^n simplices, each of which looks like $\frac{1}{2}\Delta_{n+1}$. In particular, the diameter of each of subsimplices is $\frac{\sqrt{2}}{2}$. We can iterate this process m times obtaining a subdivision into 2^{mn} subsimplices each of diameter $\frac{\sqrt{2}}{2^m}$.

We will now present a formal way to construct this subdivision. Let $T = \mathrm{conv}(\{\mathbf{e}_0, \ldots, \mathbf{e}_n\})$, where $\mathbf{e}_0, \ldots, \mathbf{e}_n$ are affinely independent; they will be the canonical vectors but our construction does not rely on this assumption. In the barricentric coordinates every $\mathbf{y} \in T$ can be written as

$$\mathbf{y} \;=\; \lambda_0\mathbf{e}_0 + \lambda_1\mathbf{e}_1 + \ldots + \lambda_n\mathbf{e}_n \qquad \lambda_0, \ldots, \lambda_n \geq 0,\ \sum_{i=0}^{n} \lambda_i = 1.$$

The same point can be expressed in terms of

$$\mathbf{x}_0 = \mathbf{e}_0 \qquad \mathbf{x}_1 = \mathbf{e}_1 - \mathbf{e}_0 \qquad \mathbf{x}_2 = \mathbf{e}_2 - \mathbf{e}_1 \qquad \ldots \qquad \mathbf{x}_n = \mathbf{e}_n - \mathbf{e}_{n-1},$$

which leads to

$$\mathbf{y} \;=\; (\sum_{i=0}^{n} \lambda_i)\mathbf{x}_0 + (\sum_{i=1}^{n} \lambda_i)\mathbf{x}_1 + (\sum_{i=2}^{n} \lambda_i)\mathbf{x}_2 + \ldots + \lambda_n\mathbf{x}_n. \tag{12.1}$$

In other words

$$\mathbf{y} \;=\; \mathbf{x}_0 + \alpha_1\mathbf{x}_1 + \cdots + \alpha_n\mathbf{x}_n \qquad 1 \geq \alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_n \geq 0. \tag{12.2}$$

This means means that there is a bijective linear map between T and a subset of $[0,1]^n$ given by $1 \geq \alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_n \geq 0$. In the barrycentric coordinates, the faces of T were described by some of the coordinates being zero. Equation (12.1) implies that in the new coordinates the faces are described by some α_i being equal to each other.

For any permutation $\pi = (p_1, \ldots, p_n)$ of $\{1, \ldots, n\}$ consider a linear map $x_0 \mapsto x_0$, $\mathbf{x}_i \mapsto \mathbf{x}_{p_i}$. Under this map T is mapped to T^π given by all elements of the form

$$T^\pi \;=\; \{\mathbf{y} :\ \mathbf{y} = \mathbf{x}_0 + \alpha_1\mathbf{x}_{p_1} + \cdots + \alpha_n\mathbf{x}_{p_n} \text{ and } 1 \geq \alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_n \geq 0\}.$$

In particular, $T = T^{1\cdots n}$, where $1\cdots n$ denotes the trivial permutation. The union $\bigcup T^\pi$ over all n partitions is

$$P = \{\mathbf{y} :\ \mathbf{y} = \mathbf{x}_0 + \alpha_1\mathbf{x}_1 + \cdots + \alpha_n\mathbf{x}_n \text{ and } 0 \leq \alpha_i \leq 1\}.$$

Therefore P is a parallelepiped, an affine image of the unit cube $C = [0,1]^n$.

If $n = 2$ the situation is depicted in Figure 12.1. If $\mathbf{e}_0 = (1,0,0)$, $\mathbf{e}_1 = (0,1,0)$, $\mathbf{e}_2 = (0,0,1)$ then in the α coordinates these point are $(0,0)$, $(1,0)$ and $(1,1)$ respectively. Their convex hull is the standard simplex $T = T^{12}$; it

represents all points (α_1, α_2) such that $1 \geq \alpha_1 \geq \alpha_2 \geq 0$. The points satisfying $1 \geq \alpha_2 \geq \alpha_1 \geq 0$ lie in the copy T^{21} of T and their are spanned by $(0,0)$, $(1,1)$, and an additional point $(0,1)$, which in the barrycentric coordinates is equal to $(1,-1,1)$. The union of T and T_{21} is the parallelepiped P.

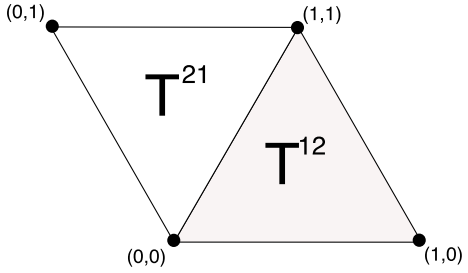

Fig. 12.1 The construction of the parallelepiped covered by $n!$ copies of T, $n = 2$.

We can now subdivide P into 2^n equal parts by subdividing the unit cube $C = [0,1]^n$. For a sequence $\sigma = (s_1, \ldots, s_n) \in \{0,1\}^n$ we can consider the part of C of the form $\frac{1}{2}(\sigma + C)$. This is simply a translation of a scaled version of C. Its image is the parallelepiped P_σ, whose elements are of the form

$$P_\sigma = \{\mathbf{y} : \ \mathbf{y} = \mathbf{x}_0 + \tfrac{1}{2}\sum_{i=1}^{n}(s_i + \alpha_i)\mathbf{x}_i \text{ and } 0 \leq \alpha_i \leq 1\}.$$

For $n = 2$ this is illustrated in Figure 12.2.

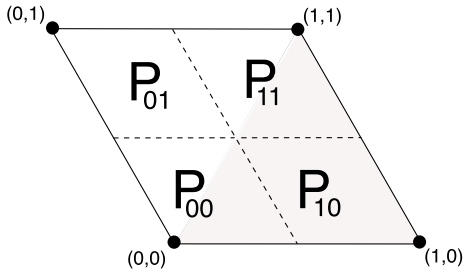

Fig. 12.2 Subdivision of P that comes from a natural subdivision of the unit n-cube into 2^n equal pieces, $n = 2$.

Since P_σ is a scaled version of P, it is also covered by scaled versions of simplices T^π, call them T^π_σ:

$$T^\pi_\sigma \ = \ \{\mathbf{y} : \ \mathbf{y} = \mathbf{x}_0 + \tfrac{1}{2}\sum_{i=1}^{n} s_i\mathbf{x}_i + \tfrac{1}{2}\sum_{i=1}^{n}\alpha_i\mathbf{x}_{p_i} \text{ and } 1 \geq \alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_n \geq 0\}.$$

For $n = 2$ the proposition is illustrated in Figure 12.3. Here the standard 2-simplex is covered by four smaller simplices $T_{10}^{12}, T_{10}^{21}, T_{00}^{12}, T_{11}^{12}$ each of which is a obtained from T by scaling, translation, and, also T_{10}^{21} by the permutation map.

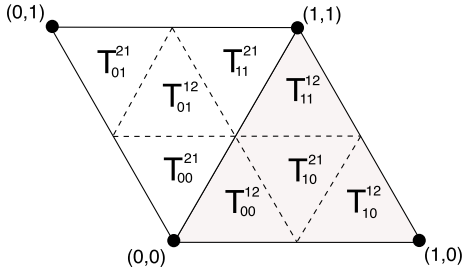

Fig. 12.3 Illustration of Proposition 12.1, $n = 2$.

Proposition 12.1. *Let A be the set of all pairs (π, σ) such that:*

(i) σ satisfies: $i < j$ then $s_i \geq s_j$,
(ii) π satisfies: $s_{p_i} = s_{p_j}$ and $p_i < p_j$ implies $i < j$.

Then

$$T = \bigcup_{(\pi,\sigma)\in A} T_\sigma^\pi.$$

Proof. We first show that $T_\sigma^\pi \subset T$ for each $(\pi, \sigma) \in A$. The points in T_σ^π are of the form

$$\mathbf{y} \;=\; \mathbf{x}_0 + \tfrac{1}{2}\sum_{i=1}^{n} s_i \mathbf{x}_i + \tfrac{1}{2}\sum_{i=1}^{n} \alpha_i \mathbf{x}_{p_i} = \mathbf{x}_0 + \tfrac{1}{2}\sum_{i=1}^{n} (s_{p_i} + \alpha_i)\mathbf{x}_{p_i}.$$

We want to show that if $p_i < p_j$ then $s_{p_i} + \alpha_i \geq s_{p_j} + \alpha_j$. If $s_{p_i} = s_{p_j}$ then this follows by (ii) because $\alpha_i \geq \alpha_j$ if $i < j$. If $s_{p_i} \neq s_{p_j}$ then necessarily $s_{p_i} = 1$ and $s_{p_j} = 0$. Then $s_{p_i} + \alpha_i \geq s_{p_j} + \alpha_j$ holds irrespective of the values of α_i, α_j.

For the opposite inclusion let $\mathbf{y} \in T$ then $\mathbf{y}$ is of the form (12.2). Let t be an integer such that

$$1 \geq \alpha_1 \geq \cdots \geq \alpha_t \geq \tfrac{1}{2} > \alpha_{t+1} \geq \cdots \geq \alpha_n \geq 0.$$

We can now write

$$\mathbf{y} \;=\; \mathbf{x}_0 + \tfrac{1}{2}\sum_{i=1}^{t} \mathbf{x}_i + \tfrac{1}{2}\sum_{i=1}^{n} \alpha_i' \mathbf{x}_i,$$

where $1 \geq \alpha'_1 \geq \cdots \alpha'_t \geq 0$ and $1 \geq \alpha'_{t+1} \geq \cdots \alpha'_n \geq 0$; for $i \leq t$ $\alpha'_i = 2(\alpha_i - \frac{1}{2})$ and for $i > t$ $\alpha'_i = 2\alpha_i$. Let $\sigma = (s_1, \ldots, s_n) \in \{0,1\}^n$ be such that $s_1 = \cdots = s_t = 1$ and $s_{t+1} = \cdots = s_n = 0$. Let $\pi = (p_1, \ldots, p_n)$ be a permutation such that

$$1 \geq \alpha'_{p_1} \geq \alpha'_{p_2} \geq \cdots \geq \alpha'_{p_n} \geq 0.$$

Since σ satisfies (i) and π satisfies (ii), it follows that $\mathbf{y} \in T^\pi_\sigma$ for some $(\pi, \sigma) \in A$. □

The union $\bigcup_{(\pi,\sigma)\in A} T^\pi_\sigma$ forms a **simplicial subdivision** of T, in the sense that:

1. $T = \bigcup_{(\pi,\sigma)\in A} T^\pi_\sigma$, and
2. for any two permutations subsimplices T^π_σ, $T^{\pi'}_{\sigma'}$ their intersection is either empty or forms a face of both.

For each of the smaller subsimplices we can now repeat the process getting a finer subdivision of T. In each step, the diameter of the subsimplices is halved and so, in particular, it very quickly converges to zero. For example, if $n = 2$ then after two steps the standard simplex get subdivided into $2^4 = 16$ subsimpleces, each with diameter $\frac{\sqrt{2}}{4}$ as illustrated in Figure 12.4.

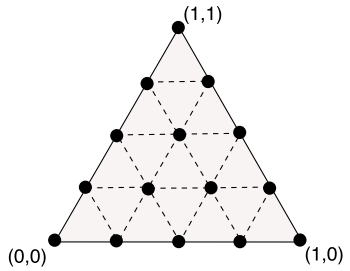

Fig. 12.4 Subdivision of the standard 2-simplex into 16 subsimpleces.

This construction will be essential for our proof of the Brouwer's fixed point theorem main part of which is a combinatorial theorem proved in the next section.

12.2 Sperner's lemma

Consider an n-simplex $T = \text{conv}(\{\mathbf{e}_0, \ldots, \mathbf{e}_n\})$ with a simplicial subdivision (e.g. described in the previous section). Let V denote the set of all vertices of all the subsimplices. For instance, in Figure 12.4, V is given by the set of all 15 solid dots. More generally we have the following.

Example 12.1. If we subdivide the standard simplex using m times the procedure described in the previous section, we get $(2^{m-1}+1)(2^m+1)$ vertices, which in the α-coordinate system are given by all points

$$(\alpha_1, \ldots, \alpha_n) \;=\; \tfrac{1}{2^m}(k_1, \ldots, k_n)$$

such that $k_1, \ldots, k_n \in \{0, 1, \ldots, 2^m\}$ and $k_1 \geq \cdots \geq k_n$. For example, the three points inside the triangle in Figure 12.4 have coordinates $(\frac{1}{2}, \frac{1}{4})$, $(\frac{3}{4}, \frac{1}{4})$, and $(\frac{3}{4}, \frac{1}{2})$, which in the barrycentric coorindates becomes $(\frac{1}{2}, \frac{1}{4}, \frac{1}{4})$, $(\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$, and $(\frac{1}{4}, \frac{1}{4}, \frac{1}{2})$ respectively.

We label the vertices of T with the labeling set $\{0, 1 \ldots, n\}$ so that the vertex $\mathbf{e}_i$ gets label i. Recall by Theorem 11.3 that for any $\mathbf{y} \in T$ the set of coefficients λ_i such that $\mathbf{y} = \sum_i \lambda_i \mathbf{e}_i$ is given uniquely. Let

$$\chi(\mathbf{y}) \;=\; \{i : \lambda_i \neq 0\} \;\subseteq \{0, 1, \ldots, n\}.$$

If $\mathbf{y}$ lies in the interior of T then, $\chi(\mathbf{y}) = \{0, \ldots, n\}$, and if $\mathbf{y}$ lies in a proper face of T, $\chi(\mathbf{y})$ is the set of vertices of T that generate this face. Use the same labeling set to arbitrary label the remaining vertices in V. Formally, this means defining a labeling function $L : V \to \{0, \ldots, n\}$. We say that labelling is proper is a **proper labelling** if $L(v) \in \chi(v)$ for all $v \in V$. A subsimplex is called **completely labeled** if L takes all values $0, \ldots, n$ on its vertices.

Example 12.2. Consider again the subdivision of the 2-simplex into 16 subsimpleces with 15 vertices as given in Figure 12.4. We color vertices with $0 =$ •, $1 =$ •, and $2 =$ •. An example of proper labelling is given in Figure 12.5, the shadowed triangles correspond to completely labeled subsimplices. Note that there are exactly five of them. Although different labellings may give different number of shadowed triangles, by the next theorem we can predict their parity.

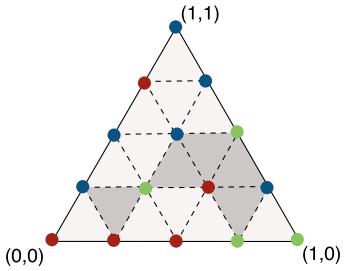

Fig. 12.5 An example of a proper vertex labeling. Completely labelled subsimplices are shaded.

Theorem 12.2 (Sperner's lemma). *Let T be a simplicially subdivided n-simplex. Suppose that the labeling function L is proper. Then there is an **odd** number of completely labeled subsimplices. In particular, there exists at least one such subsimplex.*

Proof. We prove this result by induction with respect to n.

($n = 0$) A 0-simplex is a single point $\mathbf{e}_0$ which has label 0 so it is completely labeled.

($n = 1$) A 1-simplex is simply a line segment. A simplicially subdivided 1-simplex looks like • – • – ⋯ – • – •, where the m inner nodes can be colored as • or • in an arbitrary way. If there is only one internal node, that is $m = 1$, we have either • – • – • or • – • – • and it is clear that there is always precisely one completely labeled subsimplex. Subdividing it further (increasing m) increases the number of completely labeled subsimplices either by zero or two, which establishes the general m case.

(inductive step) Suppose that the statement is true for all simplices of dimension smaller than n, we will show it holds for the n-simplex T. Let $\mathbf{C}$ be the set of all completely labeled subsimplices of T (we want to show that $|\mathbf{C}|$ is odd). Consider, in addition, all subsimplices in the face $\text{conv}\{\mathbf{x}_0, \ldots, \mathbf{x}_n\}$ labeled by $\{0, \ldots, n-1\}$ (by induction there is an odd number of them!). To every such simplex we add a vertex labelled with n and associate it with the resulting n-simples with labels $\{0, \ldots, n\}$. In Figure 12.2 there is one 2-simplex added which we identify with the segment marked with the yellow dot. We add all these n-simplices to $\mathbf{C}$ to form the set $\overline{\mathbf{C}}$. By induction, $|\mathbf{C}|$ is odd if and only if $|\overline{\mathbf{C}}|$ is even. Let $\mathbf{A}$ be the set of all n-simplices in the subdivision of T with labels $\{0, \ldots, n-1\}$ (we call them *almost completely labeled*). Let $\mathbf{E}$ be the set of all $(n-1)$-simplices with labels *exactly* $\{0, 1 \ldots, n-1\}$.

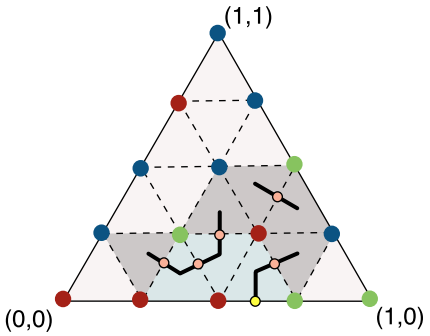

Fig. 12.6 An example of of the proposed edge set. Gray triangles are in $\mathbf{C}$, blue triangles are in $\mathbf{A}$. The yellow circle represents the additional simplex, which added to $\mathbf{C}$ forms $\overline{\mathbf{C}}$.

Note that each simplex in $\mathbf{E}$ is a common face of two simplices in $\overline{\mathbf{C}} \cup A$, in other words, it is equal to $S \cap S'$ for $S, S' \in \overline{\mathbf{C}} \cup A$. We then define a graph

with nodes $\mathbf{A} \cup \overline{\mathbf{C}}$ such that (S, S') forms an edge if $S \cap S' \in \mathbf{E}$. In a graph, the degree of a node is the number of edges incident to it. Since each edge joins two nodes we have the hand-shaking lemma, which says that

$$\sum_{S \in \overline{\mathbf{C}} \cup \mathbf{A}} \text{degree}(S) \;=\; 2|\text{edges}| \;=\; 2|\mathbf{E}|.$$

If $S \in \mathbf{A}$ then one label is repeated and so two faces of S belong to $\mathbf{E}$, so $\text{degree}(S) = 2$. If $S \in \overline{\mathbf{C}}$ then S is adjacent to precisely one other simplex in $\overline{\mathbf{C}} \cap \mathbf{A}$, and so, $\text{degree}(S) = 1$. Thus

$$\sum_{S \in \overline{\mathbf{C}} \cup \mathbf{A}} \text{degree}(S) \;\;=\;\; 2|\mathbf{A}| + |\overline{\mathbf{C}}| \;\;=\;\; 2|\mathbf{E}|.$$

This shows that $|\overline{\mathbf{C}}|$ must be even. □

12.3 Proof of the Brouwer's fixed point theorem

We are now ready to proof the Brouwer's fixed point theorem (Theorem 12.1). Our proof will consist of a sequence of reductions. We will first show that the result holds if $K = \Delta_{n+1}$. Then we will conclude that it holds for any simplex. Finally, we will show that this is already implies the Brouwer's theorem. The crucial step is the following result.

Theorem 12.3. *Let* $\mathbf{f} : \Delta_{n+1} \to \Delta_{n+1}$ *be continuous. Then* $\mathbf{f}$ *has a fixed point.*

Proof. Consider the simplicial subdivision of Δ_{n+1} into 2^n subsimplices described in Section 12.1. Define a labeling of its vertices such that

$$L(\mathbf{v}) \;\in\; \chi(\mathbf{v}) \cap \{i :\; f_i(\mathbf{v}) \le v_i\}.$$

The set on the right is always nonempty because otherwise $(\mathbf{f}(\mathbf{v}))_i > v_i$ for all i, which contradicts that $\mathbf{f}(\mathbf{v}) \in \Delta_{n+1}$. By construction, every such L is a proper labelling, and so by Sperner's lemma there exists a completely labeled subsimplex

$$T^{(1)} \;=\; \text{conv}(\{\mathbf{p}_0^{(1)}, \ldots, \mathbf{p}_n^{(1)}\}) \quad \text{such that } f_i(\mathbf{p}_i^{(1)}) \le (\mathbf{p}_i^{(1)})_i \quad \text{for all } i.$$

Consider now a subsequent subdivision of $T^{(1)}$. By an iterative argument for every $m \ge 1$ there exists a simplex

$$T^{(m)} \;=\; \text{conv}(\{\mathbf{p}_0^{(m)}, \ldots, \mathbf{p}_n^{(m)}\}) \subset T^{(m-1)}$$

such that $f_i(\mathbf{p}_i^{(m)}) \leq (\mathbf{p}_i^{(m)})_i$ for all m and all i. The diameter of $T^{(m)}$ is $\frac{\sqrt{2}}{2^m}$ and so, by Corollary 4.3,

$$\bigcap_{m\geq 1} T^{(m)} \;=\; \{\mathbf{z}\} \qquad \text{for some } \mathbf{z} \in \Delta_{n+1}.$$

Moreover, $\mathbf{p}_i^{(m)} \to \mathbf{z}$ for all i as $m \to \infty$. Since $\mathbf{f}$ is continuous

$$f_i(\mathbf{z}) \;=\; \lim_{m\to\infty} f_i(\mathbf{p}_i^{(m)}) \;\leq\; \lim_{m\to\infty} (\mathbf{p}_i^{(m)})_i \;=\; z_i.$$

This implies that $\mathbf{f}(\mathbf{z}) \leq \mathbf{z}$, which is only possible if $\mathbf{f}(\mathbf{z}) = \mathbf{z}$ because both sides sum to one. □

We will now generalize Theorem 12.3 to any set homeomorphic to the standard n-simplex.

Theorem 12.4. *Let $f : A \to A$ be a continuous function on a set homeomorphic to Δ_{n+1} for some n. Then f has a fixed point.*

Proof. Let $h : \Delta_{n+1} \to A$ be a homeomorphism. Then $h^{-1} \circ f \circ h : \Delta_{n+1} \to \Delta_{n+1}$ is continuous, and so, by Theorem 12.3, there exists $\mathbf{z} \in \Delta_{n+1}$ with $h^{-1}(f(h(\mathbf{z}))) = \mathbf{z}$. In other words $f(h(\mathbf{z})) = h(\mathbf{z})$. Since $\mathbf{x} = h(\mathbf{z}) \in A$ we get that $f(\mathbf{x}) = \mathbf{x}$ as claimed. □

Remark 12.1. This theorem shows that convexity is not the essential assumption about K in Theorem 12.1 but rather lack of holes (see Example 12.3). The two properties are equivalent only in dimension one.

Theorem 12.4 gives in particular a generalization of Theorem 12.3 to any n-simplex (see Theorem 11.4). Now we are ready to prove the Brouwer's fixed point theorem.

Proof of Theorem 12.1. Since K is compact, it is contained in a large simplex T. Define $g : T \to K$ such that

$$g(\mathbf{x}) \;=\; \arg\inf_{\mathbf{y}\in K} \|\mathbf{x} - \mathbf{y}\|.$$

By Proposition 11.2 this is a well-defined and continuous function. By construction $g(\mathbf{x}) = \mathbf{x}$ for $\mathbf{x} \in K$. Consider the sequence of maps

$$T \xrightarrow{g} K \xrightarrow{f} K \xhookrightarrow{i} T,$$

where the last inclusion is a map that simply maps each $\mathbf{x} \in K$ to itself, $i : K \to T$, $i(\mathbf{x}) = \mathbf{x}$. The composition $i \circ f \circ g$ of these maps defines a continuous function $T \to T$. As a consequence of Theorem 12.4 and Theorem 11.4, this composition has a fixed point, that is, there is $\mathbf{z} \in T$ such that $(i \circ f \circ g)(\mathbf{z}) = \mathbf{z}$.

Since the composition $f \circ g$ maps T to K, $(f \circ g)(\mathbf{z}) \in K$ and so $(i \circ f \circ g)(\mathbf{z}) \in K$, which implies that $\mathbf{z} \in K$. But then

$$(i \circ f \circ g)(\mathbf{z}) = i(f(g(\mathbf{z}))) = i(f(\mathbf{z})) = f(\mathbf{z}),$$

which gives that $f(\mathbf{z}) = \mathbf{z}$. □

The following two basic examples show that without convexity of compactness simple counterexamples are possible.

Example 12.3. Consider an annulus $K = \{\mathbf{x} \in \mathbb{R}^2 : \ 1 \leq \|\mathbf{x}\| \leq 2\}$. This is a compact set but a rotation around the origin will have no fixed points.

Example 12.4. The function $f : (0, 1] \to (0, 1]$ given by $f(x) = \frac{x}{2}$ has no fixed points.

12.4 Application: A price equilibrium theorem

Consider a market of n commodities labeled with $\{1, \ldots, n\}$. Each commodity j has a price p_j, which we normalize so that $\sum_{i=1}^{n} p_j = 1$; $\mathbf{p} \in \Delta_n$. There are m consumers in this market labeled with $\{1, \ldots, m\}$. The i-th consumer comes to the market with vector $\mathbf{w}_i \in \mathbb{R}^{n+1}$ of commodities and leaves with a vector $\mathbf{x}_i$ (she exchanges goods with other agents to improve the utility). A consumer cannot spend more than what she earns so

$$\langle \mathbf{p}, \mathbf{w}_i \rangle \ \geq \ \langle \mathbf{p}, \mathbf{x}_i \rangle \qquad \text{for all } i = 1, \ldots, m. \tag{12.3}$$

Summing this over all agents, we get

$$\langle \mathbf{p}, \sum_{i=1}^{m} (\mathbf{x}_i - \mathbf{w}_i) \rangle \ \leq \ 0 \qquad \text{for all } \mathbf{p} \in \Delta_n, \tag{12.4}$$

where $\sum_{i=1}^{m} \mathbf{x}_i$ is the total demand vector and $\sum_{i=1}^{m} \mathbf{w}_i$ is the total supply. The excess demand $\mathbf{f}(\mathbf{p}) = \sum_{i=1}^{m} (\mathbf{x}_i - \mathbf{w}_i)$ depends on the price. We assume that $\mathbf{x}_i$ is chosen to maximize the consumers utility subject to constraints (12.3). However, the utility function appears only implicitly in this simplified scenario; we assume that the maximizer $\mathbf{x}_i$ is unique and such that $\mathbf{f}(\mathbf{p})$ is continuous.

The inequality (12.4) is the weak form of Walras' law (the strong form requires equality). A price vector $\mathbf{p}$ is a free disposal equilibrium price vector if $\mathbf{f}(\mathbf{p}) \leq \mathbf{0}$, which means that no commodity has a positive excess demand. The following equilibrium theorem guarantees existence of such a price vector.

Theorem 12.5 (Arrow-Debreu[1]). *Let* $\mathbf{f} : \Delta_n \to \mathbb{R}^n$ *be continuous and such that for all* $\mathbf{p} \in \Delta_n$, $\langle \mathbf{p}, \mathbf{f}(\mathbf{p}) \rangle \leq \mathbf{0}$ *(weak Walras' law). Then the set*

$$K \;=\; \{\mathbf{p} \in \Delta_n :\; \mathbf{f}(\mathbf{p}) \leq \mathbf{0}\}$$

is compact and non-empty.

Proof. The function $\mathbf{f}$ is a continuous function and so K must be closed as an inverse image of a closed set. By Theorem 4.5, every closed subset of a compact set is compact and $K \subset \Delta_n$ so the only thing to show is that K is nonempty. Define the price adjustment function $\mathbf{h} : \Delta_n \to \Delta_n$ by

$$\mathbf{h}(\mathbf{p}) \;=\; \tfrac{\mathbf{p}+(\mathbf{f}(\mathbf{p}))^+}{1+\langle \mathbf{1}, (\mathbf{f}(\mathbf{p}))^+ \rangle},$$

where $x^+ = \max\{0, x\}$ and $\mathbf{x}^+ = (x_1^+, \ldots, x_n^+)$. We easily check that $\mathbf{h}(\mathbf{p})$ indeed lies in Δ_n:

$$\langle \mathbf{h}(\mathbf{p}), \mathbf{1} \rangle = \tfrac{1}{1+\langle \mathbf{1}, (\mathbf{f}(\mathbf{p}))^+ \rangle} \langle \mathbf{1}, p + (\mathbf{f}(\mathbf{p}))^+ \rangle = 1$$

and nonnegativity is immediate. Since $\mathbf{h}$ is continuous, by the Brouwer's fixed point theorem there exists $\mathbf{q} \in \Delta_n$ such that $\mathbf{h}(\mathbf{q}) = \mathbf{q}$, that is

$$\mathbf{q} \;=\; \tfrac{\mathbf{q}+(\mathbf{f}(\mathbf{q}))^+}{1+\langle \mathbf{1}, (\mathbf{f}(\mathbf{q}))^+ \rangle}. \tag{12.5}$$

If $\mathbf{f}(\mathbf{q})$ has positive entries, then (12.5) implies that $\mathbf{q}$ is simply the normalized version of $(\mathbf{f}(\mathbf{q}))^+$. But in this case, the Walras' law $\langle \mathbf{q}, \mathbf{f}(\mathbf{q}) \rangle \leq 0$ cannot hold! So the only option is that $\mathbf{f}(\mathbf{q}) \leq \mathbf{0}$. □

[1] Kenneth Arrow received Nobel Prize in Economics in 1972 and Gérard Debreu in 1983.

Chapter 13
Set-valued mappings (1 lecture)

13.1 Correspondences and continuity

Let X and Y be metric spaces. By $\mathcal{P}(Y)$ denote the set of all subsets of Y. A **correspondence** $\Phi : X \rightrightarrows Y$ is a function from X to $\mathcal{P}(Y)$, so $\Phi : X \to \mathcal{P}(Y)$ (set-valued function). In this chapter we review the basic theory of set-valued functions. An important aspect of this theory is developing a suitable concept of continuity with the usual intuitive meaning: small changes to the argument induce "small" changes to the resulting sets. Since, in general, we do not have a metric structure on $\mathcal{P}(Y)$ suitable notion of small changes needs to be defined. Here we give a streamlined version of the theory providing certain justification of the definitions. In Section 13.2 we will see that some aspects of this theory can be translated to the standard metric setting.

Let $\Phi : X \rightrightarrows Y$ be a correspondence. If $U \subset X$ then then the *image* of U under Φ is

$$\Phi(U) \;:=\; \bigcup_{x\in U} \Phi(x) \;\subset\; Y.$$

It is convenient to think about a function $f : X \to Y$ as a singleton-valued correspondence. The preimage of f was defined in (3.1) as follows

$$f^{\mathrm{pre}}(V) \;:=\; \{p \in X :\; f(p) \in V\}.$$

Thinking about f as a singleton-valued correspondence with values $\{f(x)\} \subset Y$, the condition $p \in f^{\mathrm{pre}}(V)$ can be therefore written in two equivalent forms

$$(a) \quad \{f(p)\} \cap V \neq \emptyset, \qquad\qquad (b) \quad \{f(p)\} \subset V.$$

What looks like an unnecessary formalism gives us two convenient ways to define preimage for correspondences. In analogy to the singleton-valued case, for every $V \subset Y$, we define the **lower preimage** of V as

$$\Phi^{\mathrm{lpre}}(V) \;:=\; \{x \in X :\; \Phi(x) \cap V \neq \emptyset\}.$$

The **upper preimage** is defined as

$$\Phi^{\mathrm{upre}}(V) := \{x \in X : \Phi(x) \subset V\}.$$

For singleton-valued correspondences both notions of preimage coincide. In general, we have $\Phi^{\mathrm{upre}}(V) \subset \Phi^{\mathrm{lpre}}(V)$ and typically the inclusion is strict. The following stronger result partially explains the relevance of having two notions of preimage.

Theorem 13.1. *For every* $V \subseteq \Phi(X)$

$$\Phi(\Phi^{\mathrm{upre}}(V)) \subseteq V \subseteq \Phi(\Phi^{\mathrm{lpre}}(V)). \tag{13.1}$$

Proof. The first inclusion follows directly from the definition of the image and the upper preimage of Φ

$$\Phi(\Phi^{\mathrm{upre}}(V)) = \bigcup_{x:\, \Phi(x)\subseteq V} \Phi(x) \subseteq V.$$

For the other inclusion it is enough to show that $\Phi(\Phi^{\mathrm{lpre}}(V)) \cap V = V$. Indeed,

$$\Phi(\Phi^{\mathrm{lpre}}(V)) \cap V = \bigcup_{x:\, \Phi(x)\cap V\neq\emptyset} (\Phi(x) \cap V) = \bigcup_{x\in X} (\Phi(x) \cap V) = \\ = \Phi(X) \cap V = V.$$

□

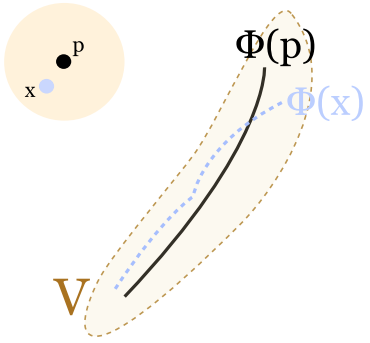

Fig. 13.1 Illustration of upper hemicontinuity in Definition 13.1.

These two notions of preimage are useful in defining a suitable notion of continuity. Fix $p \in X$. For continuity we require that small perturbations

x to p cannot result in explosion of $\Phi(x)$. Consider an open subset V that contains $\Phi(p)$ and all the points of Y that are in less than ϵ distance from $\Phi(p)$. If Φ is continuous in any suitable sense, $\Phi(x)$ should still be contained in V; see Figure 13.1. Equivalently, $x \in \Phi^{\text{upre}}(V)$.

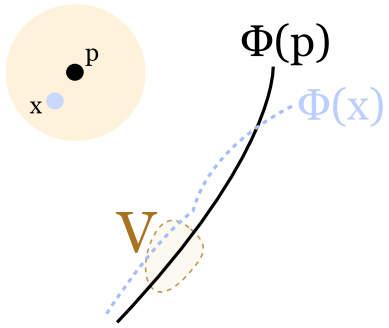

Fig. 13.2 Illustration of lower hemicontinuity in Definition 13.1.

If Φ is continuous we also require that small perturbation x to p cannot result in implosion of $\Phi(x)$. If we could fit an open subset V into $\Phi(p)$ then the condition could be the same as above: that $\Phi(x)$ contains V. However, we cannot expect that $\Phi(p)$ contains an open subset. Instead we take V to be an ϵ neighborhood of a point $y \in \Phi(p)$ and we require that $\Phi(x) \cap V$ is non-empty for a sufficiently small perturbation x of p; see Figure 13.2. Equivalently, $x \in \Phi^{\text{lpre}}(V)$.

Based on this discussion, a standard way of defining continuity of a correspondence at a point $p \in X$ is as follows.

Definition 13.1. Suppose that X, Y are metric spaces and let $\Phi : X \rightrightarrows Y$ be a correspondence. Then we say the following:

(i) Φ is **upper hemicontinuous (uhc) at** p if for every open $V \subset Y$ with $p \in \Phi^{\text{upre}}(V)$, p is an interior point of $\Phi^{\text{upre}}(V)$.
(ii) Φ is **lower hemicontinuous (uhc) at** p if for every open $V \subset Y$ with $p \in \Phi^{\text{lpre}}(V)$, p is an interior point of $\Phi^{\text{lpre}}(V)$.
(iii) Φ is **continuous** at p if it is both upper and lower hemicontinuous at p.

Moreover, Φ is continuous (uhc,lhc) if it is continuous (uhc,lhc) at every p.

We will now repeat the discussion preceding the above definition with a more detail. Let V be the open set of all points with distance less than ϵ from $\Phi(p)$. If Φ is upper hemicontinuous at p then V is always an upper bound on how $\Phi(p)$ may change when we move from p to any other point in a sufficiently small neighborhood of p. If p is an interior point of $\Phi^{\text{upre}}(V)$ then

in a small neighborhood of p all x in this neighborhood will satisfy $\Phi(x) \subset V$. Therefore, upper hemicontinuity assures that the image of Φ cannot explode when we move around p. It does not however rule out the possibility of this image to become suddenly much smaller. Taking now V to be a neighborhood of a given point $y \in \Phi(p)$, we can assure that the image $\Phi(x)$ intersects V if x is sufficiently close to p. Doing this for various $y \in \Phi(p)$ we can assure that $\Phi(x)$ does not implode.

Example 13.1. Consider a compact valued correspondence $\phi : \mathbb{R} \rightrightarrows \mathbb{R}$ of the form

$$\Phi(x) \;=\; \begin{cases} \{0\} & \text{if } x = 0, \\ \{x, \frac{1}{x}\} & \text{otherwise.} \end{cases}$$

Take $p = 0$. We have $0 \in \Phi^{\text{upre}}(V)$ if and only if $\{0\} \subset V$. Take $V = (-1, 1)$ and note that 0 is the only point in V that lies in $\Phi^{\text{upre}}(V)$. It follows that Φ is *not* upper hemicontinuous. On the other for every open $V \subset Y$ containing $p = 0$ it follows that for points q sufficiently close to p the set $\Phi(q) = \{q, \frac{1}{q}\}$ will have a non-empty intersection with V. It follows that Φ is lower hemicontinuous.

Example 13.2. Consider $\phi : \mathbb{R} \rightrightarrows \mathbb{R}^2$ given by $\Phi(x) = \{(x, 0)\}$ if $x \neq 0$ and $\Phi(0) = \{(0, y) : y \in \mathbb{R}\}$. Here $\Phi(x)$ where we move from $x = 0$ to any close point. For example, if V is a small ϵ-neighborhood of a point $(0, 1)$ then $\Phi^{\text{lpre}}(V) = \{0\}$ and so 0 is not an interior point of this set.

The following result together with the exercise below shows parallels with the definition of continuity for functions; c.f. Section 3.3 and Exercise 3.7.

Theorem 13.2. *Suppose that X, Y are metric spaces and let $\Phi : X \rightrightarrows Y$ be a correspondence. Then Φ is upper hemicontinuous if and only if for every open $V \subset Y$ the set $\Phi^{\text{upre}}(V)$ is open in X.*

Proof. Suppose Φ is upper hemicontinuous and take any open $V \subset Y$. If $\Phi^{\text{upre}}(V)$ is empty then it is open so suppose $\Phi^{\text{upre}}(V) \neq \emptyset$ and take any p such that $\Phi(p) \subset V$. Since Φ is upper hemicontinuous at p it follows that p is an interior point of $\Phi^{\text{upre}}(V)$. Since p was arbitrary, it follows that $\Phi^{\text{upre}}(V)$ is open. Now we prove the other direction. Suppose that for every open $V \subset Y$ the set $\Phi^{\text{upre}}(V)$ is open in X. Take any $p \in X$ and any $V \subset Y$ such that $\Phi(p) \subset V$. Since $\Phi^{\text{upre}}(V)$ is open and $p \in \Phi^{\text{upre}}(V)$, it follows that p is an interior point of $\Phi^{\text{upre}}(V)$ and so Φ is upper hemicontinuous at p. Since p was arbitrary the result follows. □

Similarly we have the following.

Theorem 13.3. *Suppose that X, Y are metric spaces and let $\Phi : X \rightrightarrows Y$ be a correspondence. Then Φ is lower hemicontinuous if and only if for every open $V \subset Y$ the set $\Phi^{\text{lpre}}(V)$ is open in X.*

Exercise 13.1. Prove Theorem 13.3.

Exercise 13.2. Check if the correspondence in Example 13.1 is globally lower hemicontinuous.

In practice proving continuity directly from the definition may be hard. The following sequential characterizations may be useful.

Theorem 13.4. *A correspondence $\Phi : X \rightrightarrows Y$ is lower hemicontinuous at $\mathbf{x}$ if and only if for every sequence $\mathbf{x}_n \to \mathbf{x}$ we have that*

$$\forall \mathbf{y} \in \Phi(\mathbf{x}) \quad \exists \text{ sequence } \mathbf{y}_n \in \Phi(\mathbf{x}_n) \quad \text{such that } \mathbf{y}_n \to \mathbf{y}. \tag{13.2}$$

A correspondence Φ is said to be closed-valued (compact-valued) if for every $x \in X$ the set $\Phi(x)$ is closed (compact). The sequential characterization of upper hemicontinuity that we provide works for compact-valued correspondences.

Theorem 13.5. *A compact-valued correspondence $\Phi : X \rightrightarrows Y$ is upper hemicontinuous at $\mathbf{x}$ if and only if for every sequence $\mathbf{x}_n \to \mathbf{x}$ we have that*

$$\forall (\mathbf{y}_n) \text{ with } \mathbf{y}_n \in \Phi(\mathbf{x}_n) \quad \exists \text{ subsequence } \mathbf{y}_{n_k} \to \mathbf{y} \in \Phi(\mathbf{x}). \tag{13.3}$$

13.2 Compact-valued correspondences and metric spaces*

In Chapter 3 we defined continuity of a function between metric spaces. In the case of correspondences this approach does not apply as there is no metric on $\mathcal{P}(Y)$. In this section we focus on a favorable case where $\Phi : X \rightrightarrows Y$ is compact-valued and here, with a suitable choice of metric, continuity of correspondences can be defined in the standard way.

Definition 13.2. If $C, D \subseteq \mathbb{R}^k$ are two closed nonempty sets, the **Hausdorff distance** between C and D is the quantity

$$\mathbb{D}_\infty(C, D) := \sup_{\mathbf{x} \in \mathbb{R}^k} |d_C(\mathbf{x}) - d_D(\mathbf{x})|.$$

The next exercise gives a direct interpretation of this distance.

Exercise 13.3. Show that the supremum in the definition of the Hausdorff distance could be equally taken over $C \cup D$, yielding the alternative formula

$$\mathbb{D}_\infty(C, D) := \max \left\{ \sup_{\mathbf{x} \in C} d_D(\mathbf{x}), \sup_{\mathbf{y} \in D} d_C(\mathbf{y}) \right\},$$

In general, $\mathbb{D}_\infty$ does not define a metric on $\mathcal{P}(Y)$. For example, if $Y = \mathbb{R}$ then the distance between $C = (0, 1)$ and $D = [0, 1]$ is zero despite the fact

that C and D are not equal. Denote by $\mathcal{C}(Y)$ the class of compact subsets of Y.

Theorem 13.6. *For every metric space $Y \subset \mathbb{R}^k$ the Hausdorff distance defines a metric on $\mathcal{C}(Y)$.*

Proof. It is easy to see that $\mathbb{D}_\infty(C,D) \geq 0$. By Theorem 11.5, $d_D(\mathbf{x})$ and $d_C(\mathbf{y})$ are continuous functions and so, by Theorem 4.11, there exist $\mathbf{x}^* \in C$ and $\mathbf{y}^* \in D$ such that $\mathbb{D}_\infty(C,D) = \max\{d_A(\mathbf{y}^*), d_B(\mathbf{x}^*)\}$. This shows that $\mathbb{D}_\infty(C,D) < \infty$. Moreover, this maximum is zero only if $\mathbf{x}^* \in D$ and $\mathbf{y}^* \in C$, or, equivalently, only if $C \subseteq D$ and $D \subseteq C$. To show the triangle inequality note that for all $\mathbf{y} \in D$

$$d_E(\mathbf{x}) \overset{\triangle}{\leq} \|\mathbf{x}-\mathbf{y}\| + d_E(\mathbf{y}) \leq \|\mathbf{x}-\mathbf{y}\| + \mathbb{D}_\infty(D,E).$$

Taking $\inf_{\mathbf{y}\in D}$ we get $d_E(\mathbf{x}) \leq d_D(\mathbf{x}) + \mathbb{D}_\infty(D,E) \leq \mathbb{D}_\infty(C,D) + \mathbb{D}_\infty(D,E)$. In a similar way show that $d_C(\mathbf{z}) \leq \mathbb{D}_\infty(C,D) + \mathbb{D}_\infty(D,E)$. This implies that $\sup_{\mathbf{x}\in C} d_E(\mathbf{x}) \leq \mathbb{D}_\infty(C,D) + \mathbb{D}_\infty(D,E)$ and $\sup_{\mathbf{z}\in E} d_C(\mathbf{z}) \leq \mathbb{D}_\infty(C,D) + \mathbb{D}_\infty(D,E)$. □

With this metric in hand continuity of a compact-valued correspondence can be defined in a straightforward way.

Definition 13.3. A compact-valued correspondence $\Phi : X \rightrightarrows Y \subset \mathbb{R}^k$ is **Hausdorff continuous at** p if and only for every sequence $p_n \to p$ in X, the sequence $\Phi(p_n)$ converges to $\Phi(p)$ in the Hausdorff metric.

Introducing two different concepts of continuity would not be that helpful. Luckily, it turns out that for compact-valued correspondences continuity and Hausdorff continuity are equivalent under relatively mild conditions. Note that Hausdorff continuity means that for every $\mathbf{x}_n \to \mathbf{x}$ also $\Phi(\mathbf{x}_n) \to \Phi(\mathbf{x})$ in the Hausdorff metric, or, in other words,

$$\max_{\mathbf{y}\in\Phi(\mathbf{x}_0)} d_{\Phi(\mathbf{x}_n)}(\mathbf{y}) \to 0 \quad \text{and} \quad \max_{\mathbf{y}_n\in\Phi(\mathbf{x}_n)} d_{\Phi(\mathbf{x}_0)}(\mathbf{y}_n) \to 0.$$

It is fairly easy to see that the first condition translates into (13.2) and so it is equivalent to lower hemicontinuity. The second condition implies that for every $\mathbf{x}_n \to \mathbf{x}_0$ we have that

$$\forall(\mathbf{y}_n) \text{ with } \mathbf{y}_n \in \Phi(\mathbf{x}_n) \text{ and } \mathbf{y}_n \to \mathbf{y} \quad \text{also } \mathbf{y} \in \Phi(\mathbf{x}_0). \tag{13.4}$$

This condition on Φ is called in the literature the **closed graph property**. This property is very closely related to upper hemicontinuity.

Theorem 13.7. *Let $\Phi : X \rightrightarrows Y$ be a nonempty valued correspondence with closed graph property. If for any bounded set B in X the image $\Phi(B)$ is bounded then Φ is upper hemicontinuous.*

We get the following result; see also Proposition 5 in Section E.2.5 of Ok's book.

Theorem 13.8. *Let $\Phi : X \rightrightarrows Y$ be a compact-valued correspondence with Y compact. Then Φ is continuous if and only if it is Hausdorff continuous.*

Proof. If Φ is Hausdorff continuous then by the discussion preceding Theorem 13.7 it is lower semicontinuous and has closed graph property. If $\Phi(X)$ is bounded then Theorem 13.7 implies that Φ is also upper hemicontinuous and hence continuous. For the other direction assume that Φ is lower and upper hemicontinuous. Again by the discussion preceding Theorem 13.7 lower hemicontinuity at $\mathbf{x}$ implies that for every $\mathbf{x}_n \to \mathbf{x}$ and $y_n \in \Phi(\mathbf{x}_n)$ we have $\max_{\mathbf{y}\in\Phi(\mathbf{x}_0)} d_{\Phi(\mathbf{x}_n)}(\mathbf{y}) \to 0$. Let now $(\mathbf{y}_n)$ be an arbitrary sequence such that $\mathbf{y}_n \in \Phi(\mathbf{x}_n)$. Let $\mathbf{y}_{n_k}$ be a converging subsequence and let $\mathbf{y}$ be its limit. Applying Theorem 13.5 to $\mathbf{x}_{n_k} \to \mathbf{x}$ and $\mathbf{y}_{n_k} \to \mathbf{y}$ we get that $\mathbf{y} \in \Phi(\mathbf{x})$. Since this is true for sequence $(\mathbf{y}_n)$ and any convergent subsequence, we get that $\max_{\mathbf{y}_n\in\Phi(\mathbf{x}_n)} d_{\Phi(\mathbf{x}_0)}(\mathbf{y}_n) \to 0$. □

13.3 Kakutani's fixed point theorem

Consider a compact- and convex-valued correspondence $\Phi : S \rightrightarrows S$, where $S \subset \mathbb{R}^k$ is compact and convex. We generalize the notion of a fixed point to such correspondences.

Definition 13.4. A correspondence $\Phi : S \rightrightarrows S$ has a fixed point if there exists $\mathbf{x} \in S$ such that $\mathbf{x} \in \Phi(\mathbf{x})$.

Theorem 13.9 (Kakutani's fixed point theorem). *Consider a convex- and compact-valued correspondence $\Phi : S \rightrightarrows S$, where $S \subset \mathbb{R}^k$ is compact and convex. Suppose that Φ is upper hemicontinuous then Φ has a fixed point.*

Proof. Let first $S = \Delta_{n+1}$ and consider a sequence of simplicial subdivisions given in Section 12.1. By $T^{(m)}$ denote the subdivision into 2^{nm} subsimplices obtained in the m-th step. For any vertex $\mathbf{v} \in T^{(m)}$ let $\varphi_m(\mathbf{v}) = \mathbf{y}$ for some $\mathbf{y} \in \Phi(\mathbf{v})$. If $\mathbf{x} \in S$ is not a vertex of $T^{(m)}$ define $\varphi_m(\mathbf{x})$ as a linear interpolation given by the values over the $n+1$ vertices of the simplex containing $\mathbf{x}$. If $\mathbf{x} = \sum_i \lambda_i \mathbf{v}^i$ then

$$\varphi_m(\mathbf{x}) \;=\; \sum_{i=0}^{n} \lambda_i \varphi_m(\mathbf{v}^i).$$

For every m the function $\varphi_m : S \to S$ is a continuous function and so it has a fixed point $\mathbf{x}^*_m$ by the Brower's fixed point theorem. The sequence $\mathbf{x}^*_m$ for $m \geq 1$ has a convergent subsequence, $\mathbf{x}^*_{m_k} \to \mathbf{x}^* \in S$. In the rest of the proof we show that $\mathbf{x}^* \in \Phi(\mathbf{x}^*)$.

Let $\mathbf{v}_k^0, \ldots, \mathbf{v}_k^n$ be a sequence of vertices of subsimplices in finer and finer subdivisions such that the corresponding simplex contains $x_{n_k}^*$. We have $\mathbf{v}_k^i \to \mathbf{x}^*$. We also know that $\varphi_{n_k}(\mathbf{x}_{n_k}^*) = \mathbf{x}_{n_k}^* \to \mathbf{x}^*$. Denote $\mathbf{y}_k = \varphi_{n_k}(\mathbf{x}_{n_k}^*)$ and $\mathbf{x}_k = \mathbf{v}_k^i$. Then by upper hemicontinuity and Theorem 13.5 we get that $\mathbf{x}^* \in \Phi(\mathbf{x}^*)$.

□

13.4 Application: Existence of Nash equilibria

Let S_1, S_2 be two finite sets representing the (pure) strategies of two players. Without loss of generality let $S_1 = \{1, \ldots, m\}$ and $S_2 = \{1, \ldots, n\}$. Let $\pi_1, \pi_2 : S_1 \times S_2 \to \mathbb{R}$ be the payoffs. Note that any probability distribution over S_1 is a point $\mathbf{p} \in \Delta_{m-1}$ and any probability distribution over S_2 is a point $\mathbf{q} \in \Delta_{n-1}$. We define the expected payoff of player $k = 1, 2$ as

$$\pi_k(\mathbf{p}, \mathbf{q}) \;=\; \sum_{i=1}^{m} \sum_{j=1}^{n} p_i q_j \pi_k(i, j).$$

A pair $(\mathbf{p}, \mathbf{q})$ of probability distributions over $S_1 \times S_2$ is a Nash equilibrium if

$$\pi_1(\mathbf{p}, \mathbf{q}) \;\geq\; \pi_1(\mathbf{p}', \mathbf{q}) \qquad \text{for all } \mathbf{p}' \in \Delta_{m-1}$$

and

$$\pi_2(\mathbf{p}, \mathbf{q}) \;\geq\; \pi_2(\mathbf{p}, \mathbf{q}') \qquad \text{for all } \mathbf{q}' \in \Delta_{n-1}.$$

Theorem 13.10 (Nash's theorem). *Every game has a mixed Nash equilibrium.*

Proof. Define $\Phi_1(\mathbf{q})$ be the set of all distributions $\mathbf{p}'$ that maximize $\pi_1(\mathbf{p}', \mathbf{q})$. Define $\Phi_2(\mathbf{p})$ similarly. Set $\Phi(\mathbf{p}, \mathbf{q}) = \Phi_1(\mathbf{q}) \times \Phi_2(\mathbf{p})$. With this definition, $(\mathbf{p}^*, \mathbf{q}^*)$ is a Nash equilibrium if and only if it is a fixed point of Φ. Its existence follows from Kakutani's fixed point theorem since Φ is convex-valued and is upper hemicontinuous (exercise). □