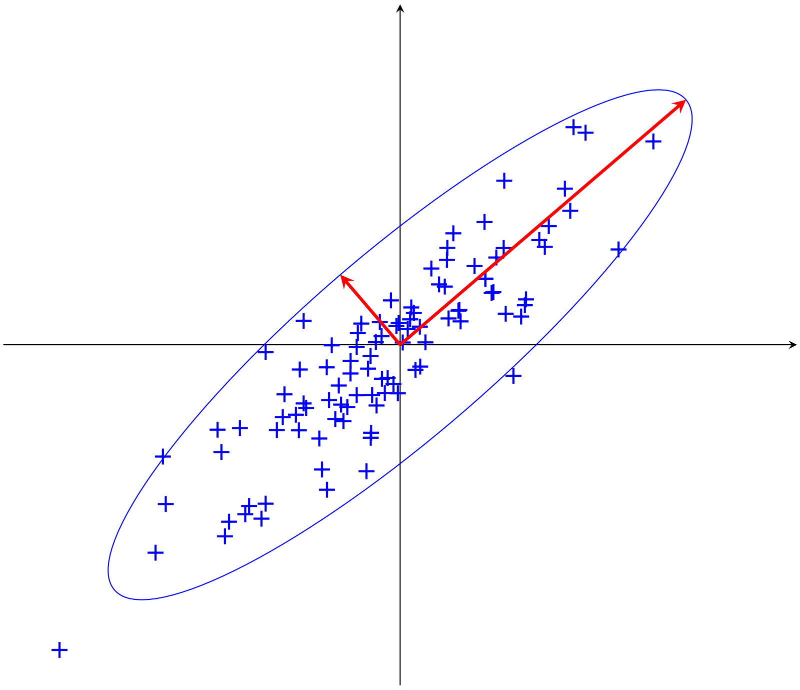
Specialization in Business Mathematics

Analysis and Linear Algebra

Josef Leydold



October 14, 2013

Institute for Statistics and Mathematics · WU Wien

Contents

Bibliography

The following books have been used to prepare this course.

[1] Kevin Houston. *How to Think Like a Mathematician*. Cambridge University Press, 2009.

[2] Knut Sydsæter and Peter Hammond. *Essential Mathematics for Economics Analysis*. Prentice Hall, 3rd edition, 2008.

[3] Knut Sydsæter, Peter Hammond, Atle Seierstad, and Arne Strøm. *Further Mathematics for Economics Analysis*. Prentice Hall, 2005.

[4] Alpha C. Chiang and Kevin Wainwright. *Fundamental Methods of Mathematical Economics*. McGraw-Hill, 4th edition, 2005.

Part I

Propedaetics

1
Introduction

1.1 Learning Outcomes

The learning outcomes of the two parts of this course in *Mathematics* are threefold:

- Mathematical reasoning
- Fundamental concepts in mathematical economics
- Extend mathematical toolbox

Topics

- Linear Algebra:
  - Vector spaces, basis and dimension
  - Matrix algebra and linear transformations
  - Norm and metric
  - Orthogonality and projections
  - Determinants
  - Eigenvalues
- Topology
  - Neighborhood and convergence
  - Open sets and continuous functions
  - Compact sets
- Calculus
  - Limits and continuity
  - Derivative, gradient and Jacobian matrix
  - Mean value theorem and Taylor series

  - Inverse and implicit functions
  - Static optimization
  - Constrained optimization

- Integration
  - Antiderivative
  - Riemann integral
  - Fundamental Theorem of Calculus
  - Leibniz's rule
  - Multiple integral and Fubini's Theorem

- Dynamic analysis
  - Ordinary differential equations (ODE)
  - Initial value problem
  - linear and logistic differential equation
  - Autonomous differential equation
  - Phase diagram and stability of solutions
  - Systems of differential equations
  - Stability of stationary points
  - Saddle path solutions

- Dynamic analysis
  - Control theory
  - Hamilton function
  - Transversality condition
  - Saddle path solutions

1.2 A Science and Language of Patterns

> Mathematics consists of propositions of the form: P implies Q, but you never ask whether P is true. (Bertrand Russell)

The mathematical universe is built-up by a series of definitions, theorems and proofs.

Axiom
A statement that is assumed to be true.
Axioms define basic concepts like sets, natural numbers or real numbers: A family of elements with rules to manipulate these.

⋮

⋮

Definition
Introduce a new notion. (Use known terms.)
Theorem
A statement that describes properties of the new object:
If . . . then . . .
Proof
Use true statements (other theorems!) to show that this statement is true.

⋮

⋮

New Definition
Based on observed interesting properties.
Theorem
A statement that describes properties of the new object.
Proof
Use true statements (including former theorems) to show that the statement is true.

⋮

????

Here is a very simplistic example:

Even number. An **even number** is a natural number n that is divisible by 2. Definition 1.1

If n is an even number, then n^2 is even. Theorem 1.2

PROOF. If n is divisible by 2, then n can be expressed as $n = 2k$ for some $k \in \mathbb{N}$. Hence $n^2 = (2k)^2 = 4k^2 = 2(2k^2)$ which also is divisible by 2. Thus n^2 is an even number as claimed. □

The *if ... then ...* structure of mathematical statements is not always obvious. Theorem 1.2 may also be expressed as: *The square of an even number is even.*

When reading the definition of *even number* we find the terms *divisible* and *natural numbers*. These terms must already be well-defined: We say that a natural number n is divisible by a natural number k if there exists a natural number m such that $n = k \cdot m$.

What are *natural numbers*? These are defined as a set of objects that satisfies a given set of rules, i.e., by *axioms*[1].

Of course the development in mathematics is not straightforward as indicate in the above diagram. It is rather a tree with some additional links between the branches.

1.3 Mathematical Economics

The quote from Bertrand Russell may seem disappointing. However, this exactly is what we are doing in *Mathematical Economics*.

An economic model is a simplistic picture of the real world. In such a model we list all our assumptions and then deduce patterns in our model from these "axioms". E.g., we may try to derive propositions like: "When we increase parameter X in model Y then variable Z declines." It is not the task of mathematics to validate the assumptions of the model, i.e., whether the model describes the real world sufficiently well.

Verification or falsification of the model is the task of economists.

1.4 About This Manuscript

This manuscript is by no means a complete treatment of the material. Rather it is intended as a road map for our course. The reader is invited to consult additional literature if she wants to learn more about particular topics.

As this course is intended as an extension of the course *Mathematische Methoden* the reader is encouraged to look at the given handouts for examples and pictures. It is also assumed that the reader has successfully mastered all the exercises of that course. Moreover, we will not repeat all definitions given there.

1.5 Solving Problems

In this course we will have to solve many problems. For this task the reader may use any theorem that have already been proved up to this point. Missing definitions could be found in the handouts for the course *Mathematische Methoden*. However, one *must not* use any results or theorems from these handouts.

[1] The natural numbers can be defined by the so called Peano axioms.

Roughly spoken there are two kinds of problems:

- Prove theorems and lemmata that are stated in the main text. For this task you may use any result that is presented up to this particular proposition that you have to show.
- Problems where additional statements have to be proven. Then all results up to the current chapter may be applied, unless stated otherwise.

Some of the problems are hard. Here is **Polya's four step plan** for tackling these issues.

(i) Understand the problem

(ii) Devise a plan

(iii) Execute the problem

(iv) Look back

1.6 Symbols and Abstract Notions

Mathematical illiterates often complain that mathematics deals with abstract notions and symbols. However, this is indeed the foundation of the great power of mathematics.

Let us give an example[2]. Suppose we want to solve the quadratic equation

$$x^2 + 10x = 39\,.$$

Muḥammad ibn Mūsā al-Khwārizmī (c. 780–850) presented an algorithm for solving this equation in his text entitled *Al-kitāb al-muḫtaṣar fī ḥisāb al-jabr wa-l-muqābala* (*The Condensed Book on the Calculation of al-Jabr and al Muqabala*). In his text he distinguishes between three kinds of quantities: the *square* [of the unknown], the *root* of the square [the unknown itself], and the *absolute numbers* [the constants in the equation]. Thus he stated our problem as

> "What must be the square which, when increased by ten of its own roots, amounts to thirty-nine?"

and presented the following recipe:

> "The solution is this: you halve the number of roots, which in the present instance yields five. This you multiply by itself; the product is twenty-five. Add this to thirty-nine; the sum us sixty-four. Now take the root of this which is eight, and subtract from it half the number of the roots, which is five; the remainder is three. This is the root of the square which you sought for."

[2] See Sect. 7.2.1 in Victor J. Katz (1993), *A History of Mathematics*, HarperCollins College Publishers.

Using modern mathematical (abstract!) notation we can express this algorithm in a more condensed form as follows:

The solution of the quadratic equation $x^2 + bx = c$ with $b, c > 0$ is obtained by the procedure

1. Halve b.
2. Square the result.
3. Add c.
4. Take the square root of the result.
5. Subtract $b/2$.

It is easy to see that the result can abstractly be written as

$$x = \sqrt{\left(\frac{b}{2}\right)^2 + c} - \frac{b}{2}\,.$$

Obviously this problem is just a special case of the general form of a quadratic equation

$$ax^2 + bx + c = 0, \quad a, b, c \in \mathbb{R}$$

with solution

$$x_{1,2} = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}\,.$$

Al-Khwārizmī provides a purely geometrically proof for his algorithm. Consequently, the constants b and c as well as the unknown x must be positive quantities. Notice that for him $x^2 = bx + c$ is a different type of equations. Thus he has to distinguish between six different types of quadratic equations for which he provides algorithms for finding their solutions (and a couple of types that do not have positive solutions at all). For each of these cases he presents geometric proofs. And Al-Khwārizmī did not use letters nor other symbols for the unknown and the constants.

— Summary

- Mathematics investigates and describes *structures* and *patterns*.
- *Abstraction* is the reason for the great power of mathematics.
- Computations and *procedures* are part of the mathematical toolbox.
- Students of this course have mastered all the exercises from the course *Mathematische Methoden*.
- Ideally students read the corresponding chapters of this manuscript *in advance before each lesson*!

2

Logic

We want to look at the foundation of mathematical reasoning.

2.1 Statements

We use a naïve definition.

Definition 2.1

A **statement** is a sentence that is either true (T) or false (F) – but not both.

Example 2.2

- *"Vienna is the capital of Austria."* is a true statement.
- *"Bill Clinton was president of Austria."* is a false statement.
- *"19 is a prime number"* is a true statement.
- *"This statement is false"* is not a statement.
- *"x is an odd number."* is not a statement. ♢

2.2 Connectives

Statements can be connected to more complex statements by means of words like *"and"*, *"or"*, *"not"*, *"if . . . then . . . "*, or *"if and only if"*. Table 2.3 lists the most important ones.

2.3 Truth Tables

Truth tables are extremely useful when learning logic. Mathematicians do not use them in day-to-day work but they provide clarity for the beginner. Table 2.4 lists truth values for important connectives.

Notice that the negation of "All cats are gray" is not "All cats are not gray" but "Not all cats are gray", that is, "There is at least one cat that is not gray".

Let P and Q be two statements.

Connective	Symbol	Name
not P	$\neg P$	Negation
P and Q	$P \wedge Q$	Conjunction
P or Q	$P \vee Q$	Disjunction
if P then Q	$P \Rightarrow Q$	Implication
Q if and only if P	$P \Leftrightarrow Q$	Equivalence

Table 2.3

Connectives fo
statements

Let P and Q be two statements.

P	Q	$\neg P$	$P \wedge Q$	$P \vee Q$	$P \Rightarrow Q$	$P \Leftrightarrow Q$
T	T	F	T	T	T	T
T	F	F	F	T	F	F
F	T	T	F	T	T	F
F	F	T	F	F	T	T

Table 2.4

Truth table for
important conn

2.4 If ... then ...

In an implication $P \Rightarrow Q$ there are two parts:

- Statement P is called the **hypothesis** or **assumption**, and
- Statement Q is called the **conclusion**.

The truth values of an **implication** seems a bit *mysterious*. Notice that $P \Rightarrow Q$ says nothing about the truth of P or Q.

Example 2.5

Which of the following statements are true?

- "If Barack Obama is Austrian citizen, then he may be elected for Austrian president."
- "If Ben is Austrian citizen, then he may be elected for Austrian president." ◊

2.5 Quantifier

Definition 2.6

The phrase "for all" is the **universal quantifier**.
It is denoted by ∀.

Definition 2.7

The phrase "there exists" is the **existential quantifier**.
It is denoted by ∃.

— Problems

2.1 Verify that the statement $(P \Rightarrow Q) \Leftrightarrow (\neg P \vee Q)$ is always true.

HINT: Compute the truth table for this statement.

2.2 **Contrapositive.** Verify that the statement

$$(P \Rightarrow Q) \Leftrightarrow (\neg Q \Rightarrow \neg P)$$

is always true. Explain this statement and give an example.

2.3 Express $P \vee Q$, $P \Rightarrow Q$, and $P \Leftrightarrow Q$ as compositions of P and Q by means of $\neg$ and $\wedge$. Prove your statement by truth tables.

2.4 Another connective is **exclusive-or** $P \oplus Q$. This statement is true if and only if exactly one of the statements P or Q is true.

(a) Establish the truth table for $P \oplus Q$.

(b) Express this statement by means of “not”, “and”, and “or”. Verify your proposition by means of truth tables.

2.5 Construct the truth table of the following statements:

(a) $\neg\neg P$ (b) $\neg(P \wedge Q)$ (c) $\neg(P \vee Q)$

(d) $\neg P \wedge P$ (e) $\neg P \vee P$ (f) $\neg P \vee \neg Q$

2.6 A **tautology** is a statement that is always true. A **contradiction** is a statement that is always false.

Which of the statements in the above problems is a tautology or a contradiction?

2.7 Assume that the statement $P \Rightarrow Q$ is true. Which of the following statements are true (or false). Give examples.

(a) $Q \Rightarrow P$ (b) $\neg Q \Rightarrow P$

(c) $\neg Q \Rightarrow \neg P$ (d) $\neg P \Rightarrow \neg Q$

3
Definitions, Theorems and Proofs

We have to read mathematical texts *and need to know what that terms mean.*

3.1 Meanings

A mathematical text is build around a skeleton of the form *"definition – theorem – proof"*. Besides that one also finds, e.g., examples, remarks, or illustrations. Here is a very short description of these terms.

- **Definition**: an explanation of the mathematical meaning of a word.
- **Theorem**: a very important true statement.
- **Proposition**: a less important but nonetheless interesting true statement.
- **Lemma**: a true statement used in proving other statements (auxiliary proposition; pl. *lemmata*).
- **Corollary**: a true statement that is a simple deduction from a theorem.
- **Proof**: the explanation of why a statement is true.
- **Conjecture**: a statement believed to be true, but for which we have no proof.
- **Axiom**: a basic assumption about a mathematical situation.

3.2 Reading

When reading definitions:

- Observe precisely the given condition.
- Find examples.
- Find standard examples (which you should memorize).
- Find trivial examples.
- Find extreme examples.
- Find non-examples, i.e., an example that *do not* satisfy the condition of the definition.

When reading theorems:

- Find assumptions and conditions.
- Draw a picture.
- Apply trivial or extreme examples.
- What happens to non-examples?

3.3 Theorems

Mathematical proofs are statements of the form *"if A then B"*. It is always possible to rephrase a theorem in this way. E.g., the statement "$\sqrt{2}$ is an irrational number" can be rewritten as *"If $x = \sqrt{2}$ then x is a irrational number"*.

When talking about mathematical theorems the following two terms are extremely important.

Definition 3.1

A **necessary** condition is one which must hold for a conclusion to be true. It does not guarantee that the result is true.

Definition 3.2

A **sufficient** condition is one which guarantees the conclusion is true. The conclusion may even be true if the condition is not satisfied.

So if we have the statement *"if A then B"*, i.e., $A \Rightarrow B$, then

- A is a sufficient condition for B, and
- B is a necessary condition for A (sometimes also written as $B \Leftarrow A$).

3.4 Proofs

Finding proofs is an art and a skill that needs to be trained. The mathematician's toolbox provide the following main techniques.

Direct Proof

The statement is derived by a straightforward computation.

If n is an odd number, then n^2 is odd. Proposition 3.3

PROOF. If n is odd, then it is not divisible by 2 and thus n can be expressed as $n = 2k+1$ for some $k \in \mathbb{N}$. Hence

$$n^2 = (2k+1)^2 = 4k^2 + 4k + 1$$

which is not divisible by 2, either. Thus n^2 is an odd number as claimed. □

Contrapositive Method

The **contrapositive** of the statement $P \Rightarrow Q$ is

$$\neg Q \Rightarrow \neg P\,.$$

We have already seen in Problem 2.2 that $(P \Rightarrow Q) \Leftrightarrow (\neg Q \Rightarrow \neg P)$. Thus in order to prove statement $P \Rightarrow Q$ we also may prove its contrapositive.

If n^2 is an even number, then n is even. Proposition 3.4

PROOF. This statement is equivalent to the statement:

> "If n is not even (i.e., odd), then n^2 is not even (i.e., odd)."

However, this statements holds by Proposition 3.3 and thus our proposition follows. □

Obviously we also could have used a direct proof to derive Proposition 3.4. However, our approach has an additional advantage: Since we already have shown that Proposition 3.3 holds, we can use it for our proof and avoid unnecessary computations.

Indirect Proof

This technique is a bit similar to the contrapositive method. Yet we assume that both P and $\neg Q$ are true and show that a contradiction results. Thus it is called **proof by contradiction** (or *reductio ad absurdum*). It is based on the equivalence $(P \Rightarrow Q) \Leftrightarrow \neg(P \wedge \neg Q)$. The advantage of this method is that we get the statement $\neg Q$ for free even when Q is difficult to show.

The square root of 2 is irrational, i.e., it cannot be written in form m/n where m and n are integers. Proposition 3.5

PROOF. Suppose the contrary that $\sqrt{2} = m/n$ where m and n are integers. *Without loss of generality* we can assume that this quotient is in its simplest form. (Otherwise cancel common divisors of m and n.) Then we find

$$\frac{m}{n} = \sqrt{2} \qquad \Leftrightarrow \qquad \frac{m^2}{n^2} = 2 \qquad \Leftrightarrow \qquad m^2 = 2n^2$$

Consequently m^2 is even and thus m is even by Proposition 3.4. So $m = 2k$ for some integer k. We then find

$$(2k)^2 = 2n^2 \qquad \Leftrightarrow \qquad 2k^2 = n^2$$

which implies that n is even and there exists an integer j such that $n = 2j$. However, we have assumed that m/n was in its simplest form; but we find

$$\sqrt{2} = \frac{m}{n} = \frac{2k}{2j} = \frac{k}{j}$$

a contradiction. Thus we conclude that $\sqrt{2}$ cannot be written as a quotient of integers. □

The phrase *"without loss of generality"* (often abbreviated as *"w.l.o.g."*) is used in cases when a general situation can be easily reduced to some special case which simplifies our arguments. In this example we just have to cancel out common divisors.

Proof by Induction

Induction is a very powerful technique. It is applied when we have an infinite number of statements $A(n)$ indexed by the natural numbers. It is based on the following theorem.

Theorem 3.6 **Principle of mathematical induction.** Let $A(n)$ be an infinite collection of statements with $n \in \mathbb{N}$. Suppose that

(i) $A(1)$ is true, and

(ii) $A(k) \Rightarrow A(k+1)$ for all $k \in \mathbb{N}$.

Then $A(n)$ is true for all $n \in \mathbb{N}$.

PROOF. Suppose that the statement does not hold for all n. Let j be the smallest natural number such that $A(j)$ is false. By assumption (i) we have $j > 1$ and thus $j - 1 \geq 1$. Note that $A(j-1)$ is true as j is the smallest possible. Hence assumption (ii) implies that $A(j)$ is true, a contradiction. □

When we apply the induction principle the following terms are useful.

- Checking condition (i) is called the **initial step**.
- Checking condition (ii) is called the **induction step**.
- Assuming that $A(k)$ is true for some k is called the **induction hypothesis**.

Let $q \in \mathbb{R}$ and $n \in \mathbb{N}$ Then Proposition 3.7

$$\sum_{j=0}^{n-1} q^j = \frac{1-q^n}{1-q}$$

PROOF. For a fixed $q \in \mathbb{R}$ this statement is indexed by natural numbers. So we prove the statement by induction.

Initial step: Obviously the statement is true for $n = 1$.

Induction step: We assume by the induction hypothesis that the statement is true for $n = k$, i.e.,

$$\sum_{j=0}^{k-1} q^j = \frac{1-q^k}{1-q} .$$

We have to show that the statement also holds for $n = k+1$. We find

$$\sum_{j=0}^{k} q^j = \sum_{j=0}^{k-1} q^j + q^k = \frac{1-q^k}{1-q} + q^k = \frac{1-q^k}{1-q} + \frac{(1-q)q^k}{1-q} = \frac{1-q^{k+1}}{1-q}$$

Thus by the Principle of Mathematical Induction the statement is true for all $n \in \mathbb{N}$. □

Proof by Cases

It is often useful to break a given problem into cases and tackle each of these individually.

Triangle inequality. Let a and b be real numbers. Then Proposition 3.8

$$|a+b| \le |a| + |b|$$

PROOF. We break the problem into four cases where a and b are positive and negative, respectively.

Case 1: $a \ge 0$ and $b \ge 0$. Then $a+b \ge 0$ and we find $|a+b| = a+b = |a|+|b|$.

Case 2: $a < 0$ and $b < 0$. Now we have $a+b < 0$ and $|a+b| = -(a+b) = (-a)+(-b) = |a|+|b|$.

Case 3: Suppose one of a and b is positive and the other negative. W.l.o.g. we assume $a < 0$ and $b \ge 0$. (Otherwise reverse the rôles of a and b.) Notice that $x \le |x|$ for all x. We have the following to subcases:
Subcase (a): $a+b > 0$ and we find $|a+b| = a+b \le |a|+|b|$.
Subcase (b): $a+b < 0$ and we find $|a+b| = -(a+b) = (-a)+(-b) \le |-a| + |-b| = |a|+|b|$.

This completes the proof. □

Counterexample

A **counterexample** is an example where a given statement does not hold. It is sufficient to find *one* counterexample to disprove a conjecture. Of course it is not sufficient to give just one example to prove a conjecture.

Reading Proofs

Proofs are often hard to read. When reading or verifying a proof keep the following in mind:

- Break into pieces.
- Draw pictures.
- Find places where the assumptions are used.
- Try extreme examples.
- Apply to a non-example: Where does the proof fail?

Mathematicians seem to like the word **trivial** which means *self-evident* or *being the simplest possible case*. Make sure that the argument really is evident for you[1].

3.5 Why Should We Deal With Proofs?

The great advantage of mathematics is that one can assess the truth of a statement by studying its proof. Truth is not determined by a higher authority who says “because I say so”. (On the other hand, it is you that has to check the proofs given by your lecturer. Copying a wrong proof from the blackboard is your fault. In mathematics the incantation “But it has been written down by the lecturer” does not work.)

Proofs help us to gain confidence in the truth of our statements.

Another reason is expressed by Ludwig Wittgenstein: *Beweise reinigen die Begriffe.* We learn something about the mathematical objects.

3.6 Finding Proofs

The only way to determine the truth or falsity of a mathematical statement is with a mathematical proof. Unfortunately, finding proofs is not always easy.

M. Sipser[2]. has the following tips for producing a proof:

[1] Nasty people say that *trivial* means: “I am confident that the proof for the statement is easy but I am too lazy to write it down.”

[2] See Sect. 0.3 in Michael Sipser (2006), *Introduction to the Theory of Computation*, 2nd international edition, Course Technology.

- *Find examples.* Pick a few examples and observe the statement in action. Draw pictures. Look at extreme examples and non-examples. See what happens when you try to find counterexamples.
- *Be patient.* Finding proofs takes times. If you do not see how to do it right away, do not worry. Researchers sometimes work for weeks or even years to find a single proof.
- *Come back to it.* Look over the statement you want to prove, think about it a bit, leave it, and then return a few minutes or hours later. Let the unconscious, intuitive part of your mind have a chance to work.
- *Try special cases.* If you are stuck trying to prove your statement, try something easier. Attempt to prove a special case first. For example, if you cannot prove your statement for every $n \geq 1$, first try to prove it for $k = 1$ and $k = 2$.
- *Be neat.* When you are building your intuition for the statement you are trying to prove, use simple, clear pictures and/or text. Sloppiness gets in the way of insight.
- *Be concise.* Brevity helps you express high-level ideas without getting lost in details. Good mathematical notation is useful for expressing ideas concisely.

3.7 When You Write Down Your Own Proof

When you believe that you have found a proof, you must write it up properly. View a proof as a kind of debate. It is you who has to *convince* your readers that your statement is indeed true. A well-written proof is a sequence of statements, wherein each one follows by simple reasoning from previous statements in the sequence. All your reasons you may use must be axioms, definitions, or theorems that your reader already accepts to be true.

— Summary

- Mathematical papers have the structure *"Definition – Theorem – Proof"*.
- A theorem consists of an *assumption* or hypothesis and a *conclusion*.
- We distinguish between *necessary* and *sufficient* conditions.
- *Examples* illustrate a notion or a statement. A good example shows a typical property; extreme examples and non-examples demonstrate special aspects of a result. An example *does not* replace a proof.

- *Proofs* verify theorems. They only use definitions and statements that have already be shown true.
- There are some techniques for proving a theorem which may (or may not) work: *direct proof, indirect proof, proof by contradiction, proof by induction, proof cases*.
- Wrong conjectures may be disproved by *counterexamples*.
- When reading definitions, theorems or proofs: find examples, draw pictures, find assumptions and conclusions.

— Problems

3.1 Consider the following statement:

> Suppose that a, b, c and d are real numbers. If $ab = cd$ and $a = c$, then $b = d$.
>
> Proof: We have
>
> $$\begin{aligned} & ab = cd \\ \Leftrightarrow\quad & ab = ad, \text{ as } a = c, \\ \Leftrightarrow\quad & b = d, \text{ by cancellation.} \end{aligned}$$

Unfortunately, this statement is false. Where is the mistake? Fix the proposition, i.e., change the statement such that it is true.

3.2 Prove that the square root of 3 is irrational, i.e., it cannot be written in form m/n where m and n are integers.

HINT: Use the same idea as in the proof of Proposition 3.5.

3.3 Suppose one uses the same idea as in the proof of Proposition 3.5 to show that the square root of 4 is irrational. Where does the proof fail?

3.4 Prove by induction that

$$\sum_{j=1}^{n} j = \frac{1}{2}n(n+1)\,.$$

3.5 The binomial coefficient is defined as

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

It also can be computed by $\binom{n}{0} = \binom{n}{n} = 1$ and the following recursion:

$$\binom{n+1}{k+1} = \binom{n}{k} + \binom{n}{k+1} \qquad \text{for } k = 0, \dots, n-1$$

This recursion can be illustrated by Pascal's triangle:

$$\begin{array}{ccccccccccc}
 & & & & & \binom{0}{0} & & & & & \\
 & & & & \binom{1}{0} & & \binom{1}{1} & & & & \\
 & & & \binom{2}{0} & & \binom{2}{1} & & \binom{2}{2} & & & \\
 & & \binom{3}{0} & & \binom{3}{1} & & \binom{3}{2} & & \binom{3}{3} & & \\
 & \binom{4}{0} & & \binom{4}{1} & & \binom{4}{2} & & \binom{4}{3} & & \binom{4}{4} & \\
\binom{5}{0} & & \binom{5}{1} & & \binom{5}{2} & & \binom{5}{3} & & \binom{5}{4} & & \binom{5}{5}
\end{array}
\qquad
\begin{array}{ccccccccccc}
 & & & & & 1 & & & & & \\
 & & & & 1 & & 1 & & & & \\
 & & & 1 & & 2 & & 1 & & & \\
 & & 1 & & 3 & & 3 & & 1 & & \\
 & 1 & & 4 & & 6 & & 4 & & 1 & \\
1 & & 5 & & 10 & & 10 & & 5 & & 1
\end{array}$$

Prove this recursion by a direct proof.

3.6 Prove the binomial theorem by induction:

HINT: Use the recursion from Problem 3.5.

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k\, y^{n-k}$$

Part II

Linear Algebra

4

Matrix Algebra

We want to cope with rows *and* columns.

4.1 Matrix and Vector

An $m \times n$ **matrix** (pl. **matrices**) is a rectangular array of mathematical expressions (e.g., numbers) that consists of m rows and n columns. We write Definition 4.1

$$\mathbf{A} = (a_{ij})_{m\times n} = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{pmatrix} .$$

We use bold upper case letters to denote matrices and corresponding lower case letters for their entries. For example, the entries of matrix $\mathbf{A}$ are denote by a_{ij}. In addition, we also use the symbol $[\mathbf{A}]_{ij}$ to denote the entry of $\mathbf{A}$ in row i and column j.

A **column vector** (or *vector* for short) is a matrix that consists of a single column. We write Definition 4.2

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} .$$

A **row vector** is a matrix that consists of a single row. We write Definition 4.3

$$\mathbf{x}' = (x_1, x_2, \dots, x_m) .$$

We use bold lower case letters to denote vectors. Symbol $\mathbf{e}_k$ denotes a column vector that has zeros everywhere except for a one in the kth position.

The set of all (column) vectors of length n with entries in $\mathbb{R}$ is denoted by $\mathbb{R}^n$. The set of all $m \times n$ matrices with entries in $\mathbb{R}$ is denoted by $\mathbb{R}^{m \times n}$.

It is convenient to write $\mathbf{A} = (\mathbf{a}_1, \dots, \mathbf{a}_n)$ to denote a matrix with *column* vectors $\mathbf{a}_1, \dots, \mathbf{a}_n$. We write $\mathbf{A} = \begin{pmatrix} \mathbf{a}'_1 \\ \vdots \\ \mathbf{a}'_m \end{pmatrix}$ to denote a matrix with *row* vectors $\mathbf{a}'_1, \dots, \mathbf{a}'_m$.

Definition 4.4

An $n \times n$ matrix is called a **square matrix**.

Definition 4.5

An $m \times n$ matrix where all entries are 0 is called a **zero matrix**. It is denoted by $\mathbf{0}_{nm}$.

Definition 4.6

An $n \times n$ square matrix with ones on the main diagonal and zeros elsewhere is called **identity matrix**. It is denoted by $\mathbf{I}_n$.

We simply write $\mathbf{0}$ and $\mathbf{I}$, respectively, if the size of $\mathbf{0}_{nm}$ and $\mathbf{I}_n$ can be determined by the context.

Definition 4.7

A **diagonal matrix** is a square matrix in which all entries outside the main diagonal are all zero. The diagonal entries themselves may or may not be zero. Thus, the $n \times n$ matrix D is diagonal if $d_{ij} = 0$ whenever $i \neq j$. We denote a diagonal matrix with entries $x_1, \dots, x_n$ by $\text{diag}(x_1, \dots, x_n)$.

Definition 4.8

An **upper triangular matrix** is a square matrix in which all entries *below* the main diagonal are all zero. Thus, the $n \times n$ matrix U is an upper triangular matrix if $u_{ij} = 0$ whenever $i > j$.

Notice that identity matrices and square zero matrices are examples for both a diagonal matrix and an upper triangular matrix.

4.2 Matrix Algebra

Definition 4.9

Two matrices $\mathbf{A}$ and $\mathbf{B}$ are **equal**, $\mathbf{A} = \mathbf{B}$, if they have the same number of rows and columns and

$$a_{ij} = b_{ij} .$$

Definition 4.10

Let $\mathbf{A}$ and $\mathbf{B}$ be two $m \times n$ matrices. Then the **sum** $\mathbf{A} + \mathbf{B}$ is the $m \times n$ matrix with elements

$$[\mathbf{A} + \mathbf{B}]_{ij} = a_{ij} + b_{ij} .$$

That is, **matrix addition** is performed element-wise.

Definition 4.11

Let $\mathbf{A}$ be an $m \times n$ matrix and $\alpha \in \mathbb{R}$. Then we define $\alpha\mathbf{A}$ by

$$[\alpha\mathbf{A}]_{ij} = \alpha\, a_{ij} .$$

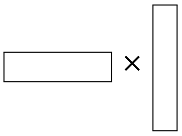
That is, **scalar multiplication** is performed element-wise.

Definition 4.12

Let $\mathbf{A}$ be an $m \times n$ matrix and $\mathbf{B}$ an $n \times k$ matrix. Then the **matrix product** $\mathbf{A} \cdot \mathbf{B}$ is the $m \times k$ matrix with elements defined as

$$[\mathbf{A} \cdot \mathbf{B}]_{ij} = \sum_{s=1}^{n} a_{is} b_{sj} .$$

That is, **matrix multiplication** is performed by multiplying *rows by columns*.



Theorem 4.13

Rules for matrix addition and multiplication.
Let $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, and $\mathbf{D}$, matrices of appropriate size. Then

(1) $\mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A}$

(2) $(\mathbf{A} + \mathbf{B}) + \mathbf{C} = \mathbf{A} + (\mathbf{B} + \mathbf{C})$

(3) $\mathbf{A} + \mathbf{0} = \mathbf{A}$

(4) $(\mathbf{A} \cdot \mathbf{B}) \cdot \mathbf{C} = \mathbf{A} \cdot (\mathbf{B} \cdot \mathbf{C})$

(5) $\mathbf{I}_m \cdot \mathbf{A} = \mathbf{A} \cdot \mathbf{I}_n = \mathbf{A}$

(6) $(\alpha \mathbf{A}) \cdot \mathbf{B} = \mathbf{A} \cdot (\alpha \mathbf{B}) = \alpha(\mathbf{A} \cdot \mathbf{B})$

(7) $\mathbf{C} \cdot (\mathbf{A} + \mathbf{B}) = \mathbf{C} \cdot \mathbf{A} + \mathbf{C} \cdot \mathbf{B}$

(8) $(\mathbf{A} + \mathbf{B}) \cdot \mathbf{D} = \mathbf{A} \cdot \mathbf{D} + \mathbf{B} \cdot \mathbf{D}$

PROOF. See Problem 4.7.

Notice: **In general matrix multiplication is not commutative!** ⚠

$$\mathbf{AB} \neq \mathbf{BA}$$

4.3 Transpose of a Matrix

Definition 4.14

The **transpose** of an $m \times n$ matrix $\mathbf{A}$ is the $n \times m$ matrix $\mathbf{A}'$ (or $\mathbf{A}^t$ or $\mathbf{A}^{\mathrm{T}}$) defined as

$$[\mathbf{A}']_{ij} = (a_{ji}) .$$

Theorem 4.15

Let $\mathbf{A}$ be an $m \times n$ matrix and $\mathbf{B}$ an $n \times k$ matrix. Then

(1) $\mathbf{A}'' = \mathbf{A}$,

(2) $(\mathbf{AB})' = \mathbf{B}'\mathbf{A}'$.

PROOF. See Problem 4.15.

Definition 4.16

A square matrix $\mathbf{A}$ is called **symmetric** if $\mathbf{A}' = \mathbf{A}$.

4.4 Inverse Matrix

Definition 4.17

A square matrix $\mathbf{A}$ is called **invertible** if there exists a matrix $\mathbf{A}^{-1}$ such that

$$\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}\,.$$

Matrix $\mathbf{A}^{-1}$ is then called the **inverse matrix** of $\mathbf{A}$.
$\mathbf{A}$ is called **singular** if such a matrix does not exist.

Theorem 4.18

Let $\mathbf{A}$ be an invertible matrix. Then its inverse $\mathbf{A}^{-1}$ is uniquely defined.

PROOF. See Problem 4.18.

Theorem 4.19

Let $\mathbf{A}$ and $\mathbf{B}$ be two invertible matrices of the same size. Then $\mathbf{AB}$ is invertible and

$$(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}\,.$$

PROOF. See Problem 4.19.

Theorem 4.20

Let $\mathbf{A}$ be an invertible matrix. Then the following holds:

(1) $(\mathbf{A}^{-1})^{-1} = \mathbf{A}$

(2) $(\mathbf{A}')^{-1} = (\mathbf{A}^{-1})'$

PROOF. See Problem 4.20.

4.5 Block Matrix

Suppose we are given some vector $\mathbf{x} = (x_1, \dots, x_n)'$. It may happen that we naturally can distinguish between two types of variables (e.g., endogenous and exogenous variables) which we can group into two respective vectors $\mathbf{x}_1 = (x_1, \dots, x_{n_1})'$ and $\mathbf{x}_2 = (x_{n_1+1}, \dots, x_{n_1+n_2})'$ where $n_1 + n_2 = n$. We then can write

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{pmatrix}.$$

Assume further that we are also given some $m \times n$ Matrix $\mathbf{A}$ and that the components of vector $\mathbf{y} = \mathbf{Ax}$ can also be partitioned into two groups

$$\mathbf{y} = \begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{pmatrix}$$

where $\mathbf{y}_1 = (y_1, \dots, x_{m_1})'$ and $\mathbf{y}_2 = (y_{m_1+1}, \dots, y_{m_1+m_2})'$. We then can partition $\mathbf{A}$ into four matrices

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix}$$

where $\mathbf{A}_{ij}$ is a submatrix of dimension $m_i \times n_j$. Hence we immediately find

$$\begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix} \cdot \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{A}_{11}\mathbf{x}_1 + \mathbf{A}_{12}\mathbf{x}_2 \\ \mathbf{A}_{21}\mathbf{x}_1 + \mathbf{A}_{22}\mathbf{x}_2 \end{pmatrix} .$$

Matrix $\begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix}$ is called a **partitioned matrix** or **block matrix**. Definition 4.21

The matrix $\mathbf{A} = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 6 & 7 & 8 & 9 & 10 \\ 11 & 12 & 13 & 14 & 15 \end{pmatrix}$ can be partitioned in numerous ways, e.g., Example 4.22

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix} = \left(\begin{array}{cc|ccc} 1 & 2 & 3 & 4 & 5 \\ 6 & 7 & 8 & 9 & 10 \\ \hline 11 & 12 & 13 & 14 & 15 \end{array}\right) \qquad \diamondsuit$$

Of course a matrix can be partitioned into more than 2 × 2 submatrices. Sometimes there is no natural reason for such a block structure but it might be convenient for further computations.

We can perform operations on block matrices in an obvious ways, that is, we treat the submatrices as of they where ordinary matrix elements. For example, we find for block matrices with appropriate submatrices,

$$\alpha \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix} = \begin{pmatrix} \alpha\mathbf{A}_{11} & \alpha\mathbf{A}_{12} \\ \alpha\mathbf{A}_{21} & \alpha\mathbf{A}_{22} \end{pmatrix}$$

$$\begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix} + \begin{pmatrix} \mathbf{B}_{11} & \mathbf{B}_{12} \\ \mathbf{B}_{21} & \mathbf{B}_{22} \end{pmatrix} = \begin{pmatrix} \mathbf{A}_{11}+\mathbf{B}_{11} & \mathbf{A}_{12}+\mathbf{B}_{12} \\ \mathbf{A}_{21}+\mathbf{B}_{21} & \mathbf{A}_{22}+\mathbf{B}_{22} \end{pmatrix}$$

and

$$\begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix} \cdot \begin{pmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} \\ \mathbf{C}_{21} & \mathbf{C}_{22} \end{pmatrix} = \begin{pmatrix} \mathbf{A}_{11}\mathbf{C}_{11}+\mathbf{A}_{12}\mathbf{C}_{21} & \mathbf{A}_{11}\mathbf{C}_{12}+\mathbf{A}_{12}\mathbf{C}_{22} \\ \mathbf{A}_{21}\mathbf{C}_{11}+\mathbf{A}_{22}\mathbf{C}_{21} & \mathbf{A}_{21}\mathbf{C}_{12}+\mathbf{A}_{22}\mathbf{C}_{22} \end{pmatrix}$$

We also can use the block structure to compute the inverse of a partitioned matrix. Assume that a matrix is partitioned as $(n_1+n_2)\times(n_1+n_2)$ matrix $\mathbf{A} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix}$. Here we only want to look at the special case where $\mathbf{A}_{21} = \mathbf{0}$, i.e.,

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{0} & \mathbf{A}_{22} \end{pmatrix}$$

We then have to find a block matrix $\mathbf{B} = \begin{pmatrix} \mathbf{B}_{11} & \mathbf{B}_{12} \\ \mathbf{B}_{21} & \mathbf{B}_{22} \end{pmatrix}$ such that

$$\mathbf{AB} = \begin{pmatrix} \mathbf{A}_{11}\mathbf{B}_{11}+\mathbf{A}_{12}\mathbf{B}_{21} & \mathbf{A}_{11}\mathbf{B}_{12}+\mathbf{A}_{12}\mathbf{B}_{22} \\ \mathbf{A}_{22}\mathbf{B}_{21} & \mathbf{A}_{22}\mathbf{B}_{22} \end{pmatrix} = \begin{pmatrix} \mathbf{I}_{n_1} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{n_2} \end{pmatrix} = \mathbf{I}_{n_1+n_2} .$$

Thus if $\mathbf{A}_{22}^{-1}$ exists the second row implies that $\mathbf{B}_{21} = \mathbf{0}_{n_2 n_1}$ and $\mathbf{B}_{22} = \mathbf{A}_{22}^{-1}$. Furthermore, $\mathbf{A}_{11}\mathbf{B}_{11} + \mathbf{A}_{12}\mathbf{B}_{21} = \mathbf{I}$ implies $\mathbf{B}_{11} = \mathbf{A}_{11}^{-1}$. At last, $\mathbf{A}_{11}\mathbf{B}_{12} + \mathbf{A}_{12}\mathbf{B}_{22} = \mathbf{0}$ implies $\mathbf{B}_{12} = -\mathbf{A}_{11}^{-1}\mathbf{A}_{12}\mathbf{A}_{22}^{-1}$. Hence we find

$$\begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{0} & \mathbf{A}_{22} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{A}_{11}^{-1} & -\mathbf{A}_{11}^{-1}\mathbf{A}_{12}\mathbf{A}_{22}^{-1} \\ \mathbf{0} & \mathbf{A}_{22}^{-1} \end{pmatrix}$$

— Summary

- A *matrix* is a rectangular array of mathematical expressions.
- Matrices can be added and multiplied by a scalar componentwise.
- Matrices can be multiplied by multiplying rows by columns.
- Matrix addition and multiplication satisfy all rules that we expect for such operations *except* that matrix multiplication is *not commutative*.
- The zero matrix **0** is the neutral element of matrix addition, i.e., **0** plays the same role as 0 for addition of real numbers.
- The identity zero matrix **I** is the neutral element of matrix multiplication, i.e., **I** plays the same role as 1 for multiplication of real numbers.
- There is no such thing as division of matrices. Instead one can use the inverse matrix, which is the matrix analog to the reciprocal of a number.
- A matrix can be partitioned. Thus one obtains a block matrix.

— Exercises

4.1 Let

$$\mathbf{A}=\begin{pmatrix}1 & -6 & 5\\ 2 & 1 & -3\end{pmatrix},\quad \mathbf{B}=\begin{pmatrix}1 & 4 & 3\\ 8 & 0 & 2\end{pmatrix}\quad\text{and}\quad \mathbf{C}=\begin{pmatrix}1 & -1\\ 1 & 2\end{pmatrix}.$$

Compute

(a) $\mathbf{A}+\mathbf{B}$ (b) $\mathbf{A}\cdot\mathbf{B}$ (c) $3\mathbf{A}'$ (d) $\mathbf{A}\cdot\mathbf{B}'$

(e) $\mathbf{B}'\cdot\mathbf{A}$ (f) $\mathbf{C}+\mathbf{A}$ (g) $\mathbf{C}\cdot\mathbf{A}+\mathbf{C}\cdot\mathbf{B}$ (h) $\mathbf{C}^2$

4.2 Demonstrate by means of the two matrices $\mathbf{A}=\begin{pmatrix}1 & -1\\ 1 & 2\end{pmatrix}$ und $\mathbf{B}=\begin{pmatrix}3 & 2\\ -1 & 0\end{pmatrix}$, that matrix multiplication is not commutative in general, i.e., we may find $\mathbf{A}\cdot\mathbf{B}\neq\mathbf{B}\cdot\mathbf{A}$.

4.3 Let $\mathbf{x}=\begin{pmatrix}1\\ -2\\ 4\end{pmatrix}$, $\mathbf{y}=\begin{pmatrix}-3\\ -1\\ 0\end{pmatrix}$.

Compute $\mathbf{x}'\mathbf{x}$, $\mathbf{x}\mathbf{x}'$, $\mathbf{x}'\mathbf{y}$, $\mathbf{y}'\mathbf{x}$, $\mathbf{x}\mathbf{y}'$ und $\mathbf{y}\mathbf{x}'$.

4.4 Let $\mathbf{A}$ be a 3×2 matrix, $\mathbf{C}$ be a 4×3 matrix, and $\mathbf{B}$ a matrix, such that the multiplication $\mathbf{A}\cdot\mathbf{B}\cdot\mathbf{C}$ is possible. How many rows and columns must $\mathbf{B}$ have? How many rows and columns does the product $\mathbf{A}\cdot\mathbf{B}\cdot\mathbf{C}$ have?

4.5 Compute $\mathbf{X}$. Assume that all matrices are quadratic matrices and all required inverse matrices exist.

(a) $\mathbf{AX}+\mathbf{BX}=\mathbf{CX}+\mathbf{I}$ (b) $(\mathbf{A}-\mathbf{B})\mathbf{X}=-\mathbf{BX}+\mathbf{C}$

(c) $\mathbf{AXA}^{-1}=\mathbf{B}$ (d) $\mathbf{XAX}^{-1}=\mathbf{C}(\mathbf{XB})^{-1}$

4.6 Use partitioning and compute the inverses of the following matrices:

(a) $\mathbf{A}=\begin{pmatrix}1 & 0 & 0 & 1\\ 0 & 1 & 0 & 2\\ 0 & 0 & 1 & 3\\ 0 & 0 & 0 & 4\end{pmatrix}$ (b) $\mathbf{B}=\begin{pmatrix}1 & 0 & 5 & 6\\ 0 & 2 & 0 & 7\\ 0 & 0 & 3 & 0\\ 0 & 0 & 0 & 4\end{pmatrix}$

— Problems

4.7 Prove Theorem 4.13.
Which conditions on the size of the respective matrices must be satisfied such that the corresponding computations are defined?

HINT: Show that corresponding entries of the matrices on either side of the equations coincide. Use the formulæ from Definitions 4.10, 4.11 and 4.12.

4.8 Show that the product of two diagonal matrices is again a diagonal matrix.

4.9 Show that the product of two upper triangular matrices is again an upper triangular matrix.

4.10 Show that the product of a diagonal matrix and an upper triangular matrices is an upper triangular matrix.

4.11 Let $\mathbf{A} = (\mathbf{a}_1, \ldots, \mathbf{a}_n)$ be an $m \times n$ matrix.

(a) What is the result of $\mathbf{A}\mathbf{e}_k$?

(b) What is the result of $\mathbf{AD}$ where $\mathbf{D}$ is an $n \times n$ diagonal matrix?

Prove your claims!

4.12 Let $\mathbf{A} = \begin{pmatrix} \mathbf{a}'_1 \\ \vdots \\ \mathbf{a}'_m \end{pmatrix}$ be an $m \times n$ matrix.

(a) What is the result of $\mathbf{e}'_k\mathbf{A}$.

(b) What is the result of $\mathbf{DA}$ where $\mathbf{D}$ is an $m \times m$ diagonal matrix?

Prove your claims!

4.13 Let $\mathbf{A}$ be an $m \times n$ matrix and $\mathbf{B}$ be an $n \times n$ matrix where $b_{kl} = 1$ for fixed $1 \le k, l \le n$ and $b_{ij} = 0$ for $i \neq k$ or $j \neq l$. What is the result of $\mathbf{AB}$? Prove your claims!

4.14 Let $\mathbf{A}$ be an $m \times n$ matrix and $\mathbf{B}$ be an $m \times m$ matrix where $b_{kl} = 1$ for fixed $1 \le k, l \le m$ and $b_{ij} = 0$ for $i \neq k$ or $j \neq l$. What is the result of $\mathbf{BA}$? Prove your claims!

4.15 Prove Theorem 4.15.

HINT: (2) Compute the matrices on either side of the equation and compare their entries.

HINT: Use Theorem 4.15.

4.16 Let $\mathbf{A}$ be an $m \times n$ matrix. Show that both $\mathbf{AA}'$ and $\mathbf{A}'\mathbf{A}$ are symmetric.

4.17 Let $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^n$. Show that matrix $\mathbf{G}$ with entries $g_{ij} = \mathbf{x}'_i\mathbf{x}_j$ is symmetric.

4.18 Prove Theorem 4.18.

HINT: Assume that there exist two inverse matrices $\mathbf{B}$ and $\mathbf{C}$. Show that they are equal.

4.19 Prove Theorem 4.19.

4.20 Prove Theorem 4.20.

HINT: Use Definition 4.17 and apply Theorems 4.19 and 4.15. Notice that $\mathbf{I}^{-1} = \mathbf{I}$ and $\mathbf{I}' = \mathbf{I}$. (Why is this true?)

4.21 Compute the inverse of $\mathbf{A} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{0} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix}$.

(Explain all intermediate steps.)

5

Vector Space

We want to master the concept of linearity.

5.1 Linear Space

Definition 5.1

In Chapter 4 we have introduced addition and scalar multiplication of vectors. Both are performed element-wise. We again obtain a vector of the same length. We thus say that the set of all real vectors of length n,

$$\mathbb{R}^n = \left\{ \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} : x_i \in \mathbb{R}, i = 1, \dots, n \right\}$$

is **closed** under vector addition and scalar multiplication.

In mathematics we find many structures which possess this nice property.

Example 5.2

Let $\mathscr{P} = \left\{\sum_{i=0}^{k} a_i x^i : k \in \mathbb{N},\ a_i \in \mathbb{R}\right\}$ be the set of all polynomials. Then we define a scalar multiplication and an addition on $\mathscr{P}$ by

- $(\alpha p)(x) = \alpha p(x)$ for $p \in \mathscr{P}$ and $\alpha \in \mathbb{R}$,
- $(p_1 + p_2)(x) = p_1(x) + p_2(x)$ for $p_1, p_2 \in \mathscr{P}$.

Obviously, the result is again a polynomial and thus an element of $\mathscr{P}$, i.e., the set $\mathscr{P}$ is closed under scalar multiplication and addition. ♢

Definition 5.3

Vector Space. A **vector space** is any non-empty set of objects that is *closed* under *scalar multiplication* and *addition*.

Of course in mathematics the meanings of the words *scalar multiplication* and *addition* needs a clear and precise definition. So we also give a formal definition:

Definition 5.4

A (real) **vector space** is an object $(\mathcal{V},+,\cdot)$ that consists of a nonempty set $\mathcal{V}$ together with two functions $+ \colon \mathcal{V}\times\mathcal{V}\to\mathcal{V}, (\mathbf{u},\mathbf{v})\mapsto \mathbf{u}+\mathbf{v}$, called *addition*, and $\cdot\colon \mathbb{R}\times\mathcal{V}\to\mathcal{V}, (\alpha,\mathbf{v})\mapsto\alpha\cdot\mathbf{v}$, called *scalar multiplication*, with the following properties:

(i) $\mathbf{v}+\mathbf{u}=\mathbf{u}+\mathbf{v}$, for all $\mathbf{u},\mathbf{v}\in\mathcal{V}$. (Commutativity)

(ii) $\mathbf{v}+(\mathbf{u}+\mathbf{w})=(\mathbf{v}+\mathbf{u})+\mathbf{w}$, for all $\mathbf{u},\mathbf{v},\mathbf{w}\in\mathcal{V}$. (Associativity)

(iii) There exists an element $0\in\mathcal{V}$ such that $0+\mathbf{v}=\mathbf{v}+0=\mathbf{v}$, for all $\mathbf{v}\in\mathcal{V}$. (Identity element of addition)

(iv) For every $\mathbf{v}\in\mathcal{V}$, there exists an $\mathbf{u}\in\mathcal{V}$ such that $\mathbf{v}+\mathbf{u}=\mathbf{u}+\mathbf{v}=0$. (Inverse element of addition)

(v) $\alpha(\mathbf{v}+\mathbf{u})=\alpha\mathbf{v}+\alpha\mathbf{u}$, for all $\mathbf{v},\mathbf{u}\in\mathcal{V}$ and all $\alpha\in\mathbb{R}$. (Distributivity)

(vi) $(\alpha+\beta)\mathbf{v}=\alpha\mathbf{v}+\beta\mathbf{v}$, for all $\mathbf{v}\in\mathcal{V}$ and all $\alpha,\beta\in\mathbb{R}$. (Distributivity)

(vii) $\alpha(\beta\mathbf{v})=(\alpha\beta)\mathbf{v}=\beta(\alpha\mathbf{v})$, for all $\mathbf{v}\in\mathcal{V}$ and all $\alpha,\beta\in\mathbb{R}$.

(viii) $1\mathbf{v}=\mathbf{v}$, for all $\mathbf{v}\in\mathcal{V}$, where $1\in\mathbb{R}$. (Identity element of scalar multiplication)

We write vector space $\mathcal{V}$ for short, if there is no risk of confusion about addition and scalar multiplication.

Example 5.5

It is easy to check that $\mathbb{R}^n$ and the set $\mathcal{P}$ of polynomials in Example 5.2 form vector spaces.

Let $\mathcal{C}^0([0,1])$ and $\mathcal{C}^1([0,1])$ be the set of all continuous and continuously differential functions with domain $[0,1]$, respectively. Then $\mathcal{C}^0([0,1])$ and $\mathcal{C}^1([0,1])$ equipped with pointwise addition and scalar multiplication as in Example 5.2 form vector spaces.

The set $\mathcal{L}^1([0,1])$ of all integrable functions on $[0,1]$ equipped with pointwise addition and scalar multiplication as in Example 5.2 forms a vector space.

A non-example is the first hyperoctant in $\mathbb{R}^n$, i.e., the set

$$H=\{\mathbf{x}\in\mathbb{R}^n : x_i\geq 0\}\,.$$

It is not a vector space as for every $\mathbf{x}\in H\setminus\{0\}$ we find $-\mathbf{x}\notin H$. ◇

Definition 5.6

Subspace. A nonempty subset $\mathcal{S}$ of some vector space $\mathcal{V}$ is called a **subspace** of $\mathcal{V}$ if for every $\mathbf{u},\mathbf{v}\in\mathcal{S}$ and $\alpha,\beta\in\mathbb{R}$ we find $\alpha\mathbf{u}+\beta\mathbf{v}\in\mathcal{S}$.

The fundamental property of vector spaces is that we can take some vectors and create a set of new vectors by means of so called linear combinations.

Definition 5.7

Linear combination. Let $\mathcal{V}$ be a real vector space. Let $\{\mathbf{v}_1,\dots,\mathbf{v}_k\}\subset\mathcal{V}$ be a finite set of vectors and $\alpha_1,\dots,\alpha_k\in\mathbb{R}$. Then $\sum_{i=1}^k\alpha_i\mathbf{v}_i$ is called a **linear combination** of the vectors $\mathbf{v}_1,\dots,\mathbf{v}_k$.

Is is easy to check that the set of all linear combinations of some fixed vectors forms a subspace of the given vector space.

Theorem 5.8

Given a vector space $\mathcal{V}$ and a nonempty subset $S = \{\mathbf{v}_1, \dots, \mathbf{v}_k\} \subset \mathcal{V}$. Then the set of all linear combinations of the elements of S is a subspace of $\mathcal{V}$.

PROOF. Let $\mathbf{x} = \sum_{i=1}^k \alpha_i \mathbf{v}_i$ and $\mathbf{y} = \sum_{i=1}^k \beta_i \mathbf{v}_i$. Then

$$\mathbf{z} = \gamma\mathbf{x} + \delta\mathbf{y} = \gamma \sum_{i=1}^k \alpha_i \mathbf{v}_i + \delta \sum_{i=1}^k \beta_i \mathbf{v}_i = \sum_{i=1}^k (\gamma\alpha_i + \delta\beta_i)\mathbf{v}_i \,.$$

is a linear combination of the elements of S for all $\gamma, \delta \in \mathbb{R}$, as claimed. □

Definition 5.9

Linear span. Let $\mathcal{V}$ be a vector space and $S = \{\mathbf{v}_1, \dots, \mathbf{v}_k\} \subset \mathcal{V}$ be a nonempty subset. Then the subspace

$$\operatorname{span}(S) = \left\{ \sum_{i=1}^k \alpha_i \mathbf{v}_i : \alpha_i \in \mathbb{R} \right\}$$

is referred as the *subspace spanned by* S and called **linear span** of S.

5.2 Basis and Dimension

Definition 5.10

Let $\mathcal{V}$ be a vector space. A subset $S \subset V$ is called a **generating set** of $\mathcal{V}$ if $\operatorname{span}(S) = \mathcal{V}$.

Definition 5.11

A vector space $\mathcal{V}$ is said to be **finitely generated**, if there exists a finite subset S of $\mathcal{V}$ that spans $\mathcal{V}$.

In the following we will restrict our interest to finitely generated real vector spaces. We will show that the notions of a basis and of linear independence are fundamental to vector spaces.

Definition 5.12

Basis. A set S is called a **basis** of some vector space $\mathcal{V}$ if it is a minimal generating set of $\mathcal{V}$. *Minimal* means that every proper subset of S does not span $\mathcal{V}$.

Theorem 5.13

If $\mathcal{V}$ is finitely generated, then it has a basis.

PROOF. Since $\mathcal{V}$ is finitely generated, it is spanned by some finite set S. If S is minimal, we are done. Otherwise, remove an appropriate element and obtain a new smaller set S' that spans $\mathcal{V}$. Repeat this step until the remaining set is minimal. □

Definition 5.14

Linear dependence. Let $S = \{\mathbf{v}_1, \dots, \mathbf{v}_k\}$ be a subset of some vector space $\mathcal{V}$. We say that S is **linearly independent** or the elements of S are linearly independent if for any $\alpha_i \in \mathbb{R}$, $i = 1, \dots, k$, $\sum_{i=1}^k \alpha_i \mathbf{v}_i = \alpha_1\mathbf{v}_1 + \dots + \alpha_k\mathbf{v}_k = 0$ implies $\alpha_1 = \dots = \alpha_k = 0$. The set S is called **linearly dependent**, if it is not linearly independent.

Theorem 5.15

Every nonempty subset of a linearly independent set is linearly independent.

PROOF. Let $\mathcal{V}$ be a vector space and $S = \{\mathbf{v}_1, \dots, \mathbf{v}_k\} \subset \mathcal{V}$ be a linearly independent set. Suppose $S' \subset S$ is linearly dependent. Without loss of generality we assume that $S' = \{\mathbf{v}_1, \dots, \mathbf{v}_m\}$, $m < k$ (otherwise rename the elements of S). Then there exist $\alpha_1, \dots, \alpha_m$ not all 0 such that $\sum_{i=1}^m \alpha_i \mathbf{v}_i = 0$. Set $\alpha_{m+1} = \ldots = \alpha_k = 0$. Then we also have $\sum_{i=1}^k \alpha_i \mathbf{v}_i = 0$, where not all α_i are zero, a contradiction to the linear independence of S.

Theorem 5.16

Every set that contains a linearly dependent set is linearly dependent.

PROOF. See Problem 5.7.

The following theorems gives us a characterization of linearly dependent sets.

Theorem 5.17

Let $S = \{\mathbf{v}_1, \dots, \mathbf{v}_k\}$ be a subset of some vector space $\mathcal{V}$. Then S is linearly dependent if and only if there exists some $\mathbf{v}_j \in S$ such that $\mathbf{v}_j = \sum_{i=1}^k \alpha_i \mathbf{v}_i$ for some $\alpha_1, \dots, \alpha_k \in \mathbb{R}$ with $\alpha_j = 0$.

PROOF. Assume that $\mathbf{v}_j = \sum_{i=1}^k \alpha_i \mathbf{v}_i$ for some $\mathbf{v}_j \in S$ such that $\alpha_1, \dots, \alpha_k \in \mathbb{R}$ with $\alpha_j = 0$. Then $0 = (\sum_{i=1}^k \alpha_i \mathbf{v}_i) - \mathbf{v}_j = \sum_{i=1}^k \alpha'_i \mathbf{v}_i$, where $\alpha'_j = \alpha_j - 1 = -1$ and $\alpha'_i = \alpha_i$ for $i \neq j$. Thus we have a solution of $\sum_{i=1}^k \alpha'_i \mathbf{v}_i = 0$ where at least $\alpha'_j \neq 0$. But this implies that S is linearly dependent.

Now suppose that S is linearly dependent. Then we find $\alpha_i \in \mathbb{R}$ not all zero such that $\sum_{i=1}^k \alpha_i \mathbf{v}_i = 0$. Without loss of generality $\alpha_j \neq 0$ for some $j \in \{1, \dots, k\}$. Then we find $\mathbf{v}_j = \frac{\alpha_1}{\alpha_j}\mathbf{v}_1 + \cdots + \frac{\alpha_{j-1}}{\alpha_j}\mathbf{v}_{j-1} + \frac{\alpha_{j+1}}{\alpha_j}\mathbf{v}_{j+1} + \cdots + \frac{\alpha_k}{\alpha_j}\mathbf{v}_k$, as proposed. □

Theorem 5.17 can also be stated as follows: $S = \{\mathbf{v}_1, \dots, \mathbf{v}_k\}$ is a linearly dependent subset of $\mathcal{V}$ if and only if there exists a $\mathbf{v} \in S$ such that $\mathbf{v} \in \operatorname{span}(S \setminus \{\mathbf{v}\})$.

Theorem 5.18

Let $S = \{\mathbf{v}_1, \dots, \mathbf{v}_k\}$ be a linearly independent subset of some vector space $\mathcal{V}$. Let $\mathbf{u} \in \mathcal{V}$. If $\mathbf{u} \notin \operatorname{span}(S)$, then $S \cup \{\mathbf{u}\}$ is linearly independent.

PROOF. See Problem 5.9.

The next theorem provides us an equivalent characterization of a basis by means of linear independent subsets.

Theorem 5.19

Let B be a subset of some vector space $\mathcal{V}$. Then the following are equivalent:

(1) B is a basis of $\mathcal{V}$.

(2) B is linearly independent generating set of $\mathcal{V}$.

PROOF. (1)⇒(2): By Definition 5.12, B is a generating set. Suppose that $B = \{\mathbf{v}_1, \dots, \mathbf{v}_k\}$ is linearly dependent. By Theorem 5.17 there exists some $\mathbf{v} \in B$ such that $\mathbf{v} \in \operatorname{span}(B')$ where $B' = B \setminus \{\mathbf{v}\}$. Without loss of generality assume that $\mathbf{v} = \mathbf{v}_k$. Then there exist $\alpha_i \in \mathbb{R}$ such that $\mathbf{v}_k = \sum_{i=1}^{k-1} \alpha_i \mathbf{v}_i$. Now let $\mathbf{u} \in \mathcal{V}$. Since B is a basis there exist some $\beta_i \in \mathbb{R}$, such that $\mathbf{u} = \sum_{i=1}^{k} \beta_i \mathbf{v}_i = \sum_{i=1}^{k-1} \beta_i \mathbf{v}_i + \beta_k \sum_{i=1}^{k-1} \alpha_i \mathbf{v}_i = \sum_{i=1}^{k-1} (\beta_i + \beta_k \alpha_i) \mathbf{v}_i$. Hence $\mathbf{u} \in \operatorname{span}(B')$. Since $\mathbf{u}$ was arbitrary, we find that B' is a generating set of $\mathcal{V}$. But since B' is a proper subset of B, B cannot be minimal, a contradiction to the minimality of a basis.

(2)⇒(1): Let B be a linearly independent generating set of $\mathcal{V}$. Suppose that B is not minimal. Then there exists a proper subset $B' \subset B$ such that $\operatorname{span}(B') = \mathcal{V}$. But then we find for every $\mathbf{x} \in B \setminus B'$ that $\mathbf{x} \in \operatorname{span}(B')$ and thus B cannot be linearly independent by Theorem 5.17, a contradiction. Hence B must be minimal as claimed. □

Theorem 5.20

A subset B of some vector space $\mathcal{V}$ is a basis if and only if it is a maximal linearly independent subset of $\mathcal{V}$. *Maximal* means that every proper superset of B (i.e., a set that contains B as a proper subset) is linearly dependent.

PROOF. See Problem 5.10.

Theorem 5.21

Steinitz exchange theorem (Austauschsatz). Let B_1 and B_2 be two bases of some vector space $\mathcal{V}$. If there is an $\mathbf{x} \in B_1 \setminus B_2$ then there exists a $\mathbf{y} \in B_2 \setminus B_1$ such that $(B_1 \cup \{\mathbf{y}\}) \setminus \{\mathbf{x}\}$ is a basis of $\mathcal{V}$.

This theorem tells us that we can replace vectors in B_1 by some vectors in B_2.

PROOF. Let $B_1 = \{\mathbf{v}_1, \dots, \mathbf{v}_k\}$ and assume without loss of generality that $\mathbf{x} = \mathbf{v}_1$. (Otherwise rename the elements of B_1.) As B_1 is a basis it is linearly independent by Theorem 5.19. By Theorem 5.15, $B_1 \setminus \{\mathbf{v}_1\}$ is also linearly independent and thus it cannot be a basis of $\mathcal{V}$ by Theorem 5.20. Hence it cannot be a generating set by Theorem 5.19. This implies that there exists a $\mathbf{y} \in B_2$ with $\mathbf{y} \notin \operatorname{span}(B_1 \setminus \{\mathbf{v}_1\})$, since otherwise we had $\operatorname{span}(B_2) \subseteq \operatorname{span}(B_1 \setminus \{\mathbf{v}_1\}) \neq \mathcal{V}$, a contradiction as B_2 is a basis of $\mathcal{V}$.

Now there exist $\alpha_i \in \mathbb{R}$ not all equal to zero such that $\mathbf{y} = \sum_{i=1}^{k} \alpha_i \mathbf{v}_i$. In particular $\alpha_1 \neq 0$, since otherwise $\mathbf{y} \in \operatorname{span}(B_1 \setminus \{\mathbf{v}_1\})$, a contradiction to the choice of $\mathbf{y}$. We then find

$$\mathbf{v}_1 = \frac{1}{\alpha_1}\mathbf{y} - \sum_{i=2}^{k} \frac{\alpha_i}{\alpha_1}\mathbf{v}_i \,.$$

Similarly for every $\mathbf{z} \in \mathcal{V}$ there exist $\beta_j \in \mathbb{R}$ such that $\mathbf{z} = \sum_{j=1}^{k} \beta_j \mathbf{v}_j$. Consequently,

$$\begin{aligned} \mathbf{z} = \sum_{j=1}^{k} \beta_j \mathbf{v}_j = \beta_1 \mathbf{v}_1 + \sum_{j=2}^{k} \beta_j \mathbf{v}_j = \beta_1 \left(\frac{1}{\alpha_1}\mathbf{y} - \sum_{i=2}^{k} \frac{\alpha_i}{\alpha_1}\mathbf{v}_i \right) + \sum_{j=2}^{k} \beta_j \mathbf{v}_j \\ = \frac{\beta_1}{\alpha_1}\mathbf{y} + \sum_{j=2}^{k} \left(\beta_j - \frac{\beta_1}{\alpha_1}\alpha_j \right) \mathbf{v}_j \end{aligned}$$

that is, $(B_1 \cup \{\mathbf{y}\}) \setminus \{\mathbf{x}\} = \{\mathbf{y}, \mathbf{v}_2, \dots, \mathbf{v}_k\}$ is a generating set of $\mathcal{V}$. By our choice of $\mathbf{y}$ and Theorem 5.18 this set is linearly independent. Thus it is a basis by Theorem 5.19. This completes the proof. □

Theorem 5.22 Any two bases B_1 and B_2 of some finitely generated vector space $\mathcal{V}$ have the same size.

PROOF. See Problem 5.11.

⚠ We want to emphasis here that in opposition to the dimension the basis of a vector space is not unique! Indeed there is infinite number of bases.

Definition 5.23 **Dimension.** Let $\mathcal{V}$ be a finitely generated vector space. Let n be the number of elements in a basis. Then n is called the **dimension** of $\mathcal{V}$ and we write $\dim(\mathcal{V}) = n$. $\mathcal{V}$ is called an n-dimensional vector space.

Theorem 5.24 Any linearly independent subset S of some finitely generated vector space $\mathcal{V}$ can be extended into a basis of B with $S \subseteq B$.

PROOF. See Problem 5.12.

Theorem 5.25 Let $B = \{\mathbf{v}_1, \dots, \mathbf{v}_n\}$ be some basis of vector space $\mathcal{V}$. Assume that $\mathbf{x} = \sum_{i=1}^{n} \alpha_i \mathbf{v}_i$ where $\alpha_i \in \mathbb{R}$ and $\alpha_1 \neq 0$. Then $B' = \{\mathbf{x}, \mathbf{v}_2, \dots, \mathbf{v}_n\}$ is a basis of $\mathcal{V}$.

PROOF. See Problem 5.15.

5.3 Coordinate Vector

Let $\mathcal{V}$ be an n-dimensional vector space with basis $B = \{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n\}$. Then we can express a given vector $\mathbf{x} \in \mathcal{V}$ as a linear combination of the basis vectors, i.e.,

$$\mathbf{x} = \sum_{i=1}^{n} \alpha_i \mathbf{v_i}$$

where the $\alpha_i \in \mathbb{R}$.

Theorem 5.26 Let $\mathcal{V}$ be a vector space with some basis $B = \{\mathbf{v}_1, \dots, \mathbf{v}_n\}$. Let $\mathbf{x} \in \mathcal{V}$ and $\alpha_i \in \mathbb{R}$ such that $\mathbf{x} = \sum_{i=1}^{n} \alpha_i \mathbf{v}_i$. Then the coefficients $\alpha_1, \dots, \alpha_n$ are uniquely defined.

PROOF. See Problem 5.16.

This theorem allows us to define the coefficient vector of $\mathbf{x}$.

Definition 5.27

Let $\mathcal{V}$ be a vector space with some basis $B = \{\mathbf{v}_1, \dots, \mathbf{v}_n\}$. For some vector $\mathbf{x} \in \mathcal{V}$ we call the uniquely defined numbers $\alpha_i \in \mathbb{R}$ with $\mathbf{x} = \sum_{i=1}^n \alpha_i \mathbf{v}_i$ the **coefficients** of $\mathbf{x}$ with respect to basis B. The vector $\mathbf{c}(\mathbf{x}) = (\alpha_1, \dots, \alpha_n)'$ is then called the **coefficient vector** of $\mathbf{x}$.

We then have

$$\mathbf{x} = \sum_{i=1}^n c_i(\mathbf{x}) \mathbf{v}_i \,.$$

Notice that $\mathbf{c}(\mathbf{x}) \in \mathbb{R}^n$.

Example 5.28

Canonical basis. It is easy to verify that $B = \{\mathbf{e}_1, \dots, \mathbf{e}_n\}$ forms a basis of the vector space $\mathbb{R}^n$. It is called the **canonical basis** of $\mathbb{R}^n$ and we immediately find that for each $\mathbf{x} = (x_1, \dots, x_n)' \in \mathbb{R}^n$,

$$\mathbf{x} = \sum_{i=1}^n x_i \mathbf{e}_i \,. \qquad \diamondsuit$$

Example 5.29

The set $\mathcal{P}_2 = \{a_0 + a_1 x + a_2 x^2 \colon a_i \in \mathbb{R}\}$ of polynomials of order less than or equal to 2 equipped with the addition and scalar multiplication of Example 5.2 is a vector space with basis $B = \{\mathbf{v}_0, \mathbf{v}_1, \mathbf{v}_2\} = \{1, x, x^2\}$. Then any polynomial $p \in \mathcal{P}_2$ has the form

$$p(x) = \sum_{i=0}^2 a_i x^i = \sum_{i=0}^2 a_i \mathbf{v}_i$$

that is, $\mathbf{c}(p) = (a_0, a_1, a_2)'$. $\diamondsuit$

The last example demonstrates an important consequence of Theorem 5.26: there is a one-to-one correspondence between a vector $\mathbf{x} \in \mathcal{V}$ and its coefficient vector $\mathbf{c}(\mathbf{x}) \in \mathbb{R}^n$. The map $\mathcal{V} \to \mathbb{R}^n, \mathbf{x} \mapsto \mathbf{c}(\mathbf{x})$ preserves the linear structure of the vector space, that is, for vectors $\mathbf{x}, \mathbf{y} \in \mathcal{V}$ and $\alpha, \beta \in \mathbb{R}$ we find (see Problem 5.17)

$$\mathbf{c}(\alpha\mathbf{x} + \beta\mathbf{y}) = \alpha\mathbf{c}(\mathbf{x}) + \beta\mathbf{c}(\mathbf{y}) \,.$$

In other words, the coefficient vector of a linear combination of two vectors is the corresponding linear combination of the coefficient vectors of the two vectors.

In this sense $\mathbb{R}^n$ is the prototype of any n-dimensional vector space $\mathcal{V}$. We say that $\mathcal{V}$ and $\mathbb{R}^n$ are **isomorphic**, $\mathcal{V} \cong \mathbb{R}^n$, that is, they have the same structure.

Now let $B_1 = \{\mathbf{v}_1, \dots, \mathbf{v}_n\}$ and $B_2 = \{\mathbf{w}_1, \dots, \mathbf{w}_n\}$ be two bases of vector space $\mathcal{V}$. Let $\mathbf{c}_1(\mathbf{x})$ and $\mathbf{c}_2(\mathbf{x})$ be the respective coefficient vectors of $\mathbf{x} \in \mathcal{V}$. Then we have

$$\mathbf{w}_j = \sum_{i=1}^n c_{1i}(\mathbf{w}_j) \mathbf{v}_i \,, \qquad j = 1, \dots, n$$

and

$$\sum_{i=1}^{n} c_{1i}(\mathbf{x})\mathbf{v}_i = \mathbf{x} = \sum_{j=1}^{n} c_{2j}(\mathbf{x})\mathbf{w}_j = \sum_{j=1}^{n} c_{2j}(\mathbf{x}) \sum_{i=1}^{n} c_{1i}(\mathbf{w}_j)\mathbf{v}_i$$
$$= \sum_{i=1}^{n} \left(\sum_{j=1}^{n} c_{2j}(\mathbf{x}) c_{1i}(\mathbf{w}_j) \right) \mathbf{v}_i \,.$$

Consequently, we find

$$c_{1i}(\mathbf{x}) = \sum_{j=1}^{n} c_{1i}(\mathbf{w}_j) c_{2j}(\mathbf{x})\,.$$

Thus let $\mathbf{U}_{12}$ contain the coefficient vectors of the basis vectors of B_2 with respect to basis B_1 as its columns, i.e.,

$$[\mathbf{U}_{12}]_{ij} = c_{1i}(\mathbf{w}_j)\,.$$

Then we find

$$\mathbf{c}_1(\mathbf{x}) = \mathbf{U}_{12}\mathbf{c}_2(\mathbf{x})\,.$$

Definition 5.30 Matrix $\mathbf{U}_{12}$ is called the **transformation matrix** that transforms the coefficient vector $\mathbf{c}_2$ with respect to basis B_2 into the coefficient vector $\mathbf{c}_1$ with respect to basis B_1.

— Summary

- A *vector space* is a set of elements that can be added and multiplied by a scalar (number).
- A vector space is *closed* under forming *linear combinations*, i.e.,

$$\mathbf{x}_1, \dots, \mathbf{x}_k \in \mathcal{V} \text{ and } \alpha_1, \dots, \alpha_k \in \mathbb{R} \quad \text{implies} \quad \sum_{i=1}^{k} \alpha_i \mathbf{x}_i \in \mathcal{V}\,.$$

- A set of vectors is called *linear independent* if it is not possible to express one of these as a linear combination of the remaining vectors.
- A *basis* is a minimal generating set, or equivalently, a maximal set of linear independent vectors.
- The basis of a given vector space is *not unique*. However, all bases of a given vector space have the same size which is called the *dimension* of the vector space.
- For a given basis every vector has a uniquely defined *coordinate vector*.
- The *transformation matrix* allows to transform a coordinate vector w.r.t. one basis into the coordinate vector w.r.t. another one.
- Every vector space of dimension n “looks like” the $\mathbb{R}^n$.

— Exercises

5.1 Give linear combinations of the two vectors $\mathbf{x}_1$ and $\mathbf{x}_2$.

(a) $\mathbf{x}_1 = \begin{pmatrix} 1 \\ 2 \end{pmatrix}, \quad \mathbf{x}_2 = \begin{pmatrix} 3 \\ 1 \end{pmatrix}$ (b) $\mathbf{x}_1 = \begin{pmatrix} 2 \\ 0 \\ 1 \end{pmatrix}, \quad \mathbf{x}_2 = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}$

— Problems

5.2 Let $\mathcal{S}$ be some vector space. Show that $0 \in \mathcal{S}$.

5.3 Give arguments why the following sets are or are not vector spaces:

(a) The empty set, $\emptyset$.

(b) The set $\{0\} \subset \mathbb{R}^n$.

(c) The set of all $m \times n$ matrices, $\mathbb{R}^{m \times n}$, for fixed values of m and n.

(d) The set of all square matrices.

(e) The set of all $n \times n$ diagonal matrices, for some fixed values of n.

(f) The set of all polynomials in one variable x.

(g) The set of all polynomials of degree less than or equal to some fixed value n.

(h) The set of all polynomials of degree equal to some fixed value n.

(i) The set of points $\mathbf{x}$ in $\mathbb{R}^n$ that satisfy the equation $\mathbf{A}\mathbf{x} = \mathbf{b}$ for some fixed matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ and some fixed vector $\mathbf{b} \in \mathbb{R}^m$.

(j) The set $\{\mathbf{y} = \mathbf{A}\mathbf{x} \colon \mathbf{x} \in \mathbb{R}^n\}$, for some fixed matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$.

(k) The set $\{\mathbf{y} = \mathbf{b}_0 + \alpha\mathbf{b}_1 \colon \alpha \in \mathbb{R}\}$, for fixed vectors $\mathbf{b}_0, \mathbf{b}_1 \in \mathbb{R}^n$.

(l) The set of all functions on $[0,1]$ that are both continuously differentiable and integrable.

(m) The set of all functions on $[0,1]$ that are not continuous.

(n) The set of all random variables X on some given probability space $(\Omega, \mathcal{F}, P)$.

Which of these vector spaces are finitely generated?
Find generating sets for these. If possible give a basis.

5.4 Prove the following proposition: Let $\mathcal{S}_1$ and $\mathcal{S}_2$ be two subspaces of vector space $\mathcal{V}$. Then $\mathcal{S}_1 \cap \mathcal{S}_2$ is a subspace of $\mathcal{V}$.

5.5 Proof or disprove the following statement:

Let $\mathcal{S}_1$ and $\mathcal{S}_2$ be two subspaces of vector space $\mathcal{V}$. Then their *union* $\mathcal{S}_1 \cup \mathcal{S}_2$ is a subspace of $\mathcal{V}$.

5.6 Let $\mathcal{U}_1$ and $\mathcal{U}_2$ be two subspaces of a vector space $\mathcal{V}$. Then the sum of $\mathcal{U}_1$ and $\mathcal{U}_2$ is the set

$$\mathcal{U}_1 + \mathcal{U}_2 = \{\mathbf{u}_1 + \mathbf{u}_2 : \mathbf{u}_1 \in \mathcal{U}_1, \mathbf{u}_2 \in \mathcal{U}_2\}\,.$$

Show that $\mathcal{U}_1 + \mathcal{U}_2$ is a subspace of $\mathcal{V}$.

5.7 Prove Theorem 5.16.

5.8 Does Theorem 5.17 still hold if we allow $\alpha_j \neq 0$?

5.9 Prove Theorem 5.18.

5.10 Prove Theorem 5.20.

HINT: Use Theorem 5.19(2). First assume that B is a basis but not maximal. Then a larger linearly independent subset exists which implies that B cannot be a generating set. For the converse statement assume that B is maximal but not a generating set. Again derive a contradiction.

5.11 Prove Theorem 5.22.

HINT: Look at $B_1 \setminus B_2$. If it is nonempty construct a new basis B'_1 by means of the Austauschsatz where an element in $B_1 \setminus B_2$ is replaced by a vector in $B_2 \setminus B_1$. What is the size of $B'_1 \cap B_2$ compared to that of $B_1 \cap B_2$. When $B'_1 \setminus B_2 \neq \emptyset$ repeat this procedure and check again. Does this procedure continue ad infinitum or does it stop? Why or why not? When it stops at a basis, say, B^*_1, what do we know about the relation of B^*_1 and B_2? Is one included in the other? What does it mean for their cardinalities? What happens if we exchange the roles of B_1 and B_2?

5.12 Prove Theorem 5.24.

HINT: Start with S and add linearly independent vectors (Why is this possible?) until we obtain a maximal linearly independent set. This is then a basis that contains S. (Why?)

5.13 Prove Theorem 5.24 by means of the Austauschsatz.

5.14 Let $\mathcal{U}_1$ and $\mathcal{U}_2$ be two subspaces of a vector space $\mathcal{V}$. Show that

$$\dim(\mathcal{U}_1) + \dim(\mathcal{U}_2) = \dim(\mathcal{U}_1 + \mathcal{U}_2) + \dim(\mathcal{U}_1 \cap \mathcal{U}_2)\,.$$

HINT: Use Theorem 5.24.

5.15 Prove Theorem 5.25.

HINT: Express $\mathbf{v}_1$ as linear combination of elements in B' and show that B' is a generating set by replacing $\mathbf{v}_1$ by this expression. It remains to show that the set is a minimal generating set. (Why is any strict subset not a generating set?)

5.16 Prove Theorem 5.26.

HINT: Assume that there are two sets of numbers $\alpha_i \in \mathbb{R}$ and $\beta_i \in \mathbb{R}$ such that $\mathbf{x} = \sum_{i=1}^n \alpha_i \mathbf{v}_i = \sum_{i=1}^n \beta_i \mathbf{v}_i$. Show that $\alpha_i = \beta_i$ by means of the fact that $\mathbf{x} - \mathbf{x} = 0$.

5.17 Let $\mathcal{V}$ be an n-dimensional vector space. Show that for two vectors $\mathbf{x}, \mathbf{y} \in \mathcal{V}$ and $\alpha, \beta \in \mathbb{R}$,

$$\mathbf{c}(\alpha\mathbf{x} + \beta\mathbf{y}) = \alpha\mathbf{c}(\mathbf{x}) + \beta\mathbf{c}(\mathbf{y})\,.$$

5.18 Show that a coefficient vector $\mathbf{c}(\mathbf{x}) = 0$ if and only if $\mathbf{x} = 0$.

5.19 Let $\mathcal{V}$ be a vector space with bases $B_1 = \{\mathbf{v}_1, \dots, \mathbf{v}_n\}$ and $B_2 = \{\mathbf{w}_1, \dots, \mathbf{w}_n\}$.

(a) How does the transformation matrix $\mathbf{U}_{21}$ look like?

(b) What is the relation between transformation matrices $\mathbf{U}_{12}$ and $\mathbf{U}_{21}$.

6

Linear Transformations

We want to preserve linear structures.

6.1 Linear Maps

In Section 5.3 we have seen that the transformation that maps a vector $\mathbf{x} \in \mathcal{V}$ to its coefficient vector $\mathbf{c}(\mathbf{x}) \in \mathbb{R}^{\dim \mathcal{V}}$ preserves the linear structure of vector space $\mathcal{V}$.

Definition 6.1

Let $\mathcal{V}$ and $\mathcal{W}$ be two vector spaces. A transformation $\phi\colon \mathcal{V} \to \mathcal{W}$ is called a **linear map** if for all $\mathbf{x}, \mathbf{y} \in \mathcal{V}$ and all $\alpha, \beta \in \mathbb{R}$ the following holds:

(i) $\phi(\mathbf{x}+\mathbf{y}) = \phi(\mathbf{x}) + \phi(\mathbf{y})$

(ii) $\phi(\alpha\mathbf{x}) = \alpha\phi(\mathbf{x})$

We thus have

$$\phi(\alpha\mathbf{x}+\beta\mathbf{y}) = \alpha\phi(\mathbf{x}) + \beta\phi(\mathbf{y})\,.$$

Example 6.2

Every $m \times n$ matrix $\mathbf{A}$ defines a linear map (see Problem 6.2)

$$\phi_{\mathbf{A}}\colon \mathbb{R}^n \to \mathbb{R}^m, \mathbf{x} \mapsto \phi_{\mathbf{A}}(\mathbf{x}) = \mathbf{A}\mathbf{x}\,.$$

◇

Example 6.3

Let $\mathcal{P} = \left\{\sum_{i=0}^{k} a_i x^i : k \in \mathbb{N},\, a_i \in \mathbb{R}\right\}$ be the vector space of all polynomials (see Example 5.2). Then the map $\frac{d}{dx}\colon \mathcal{P} \to \mathcal{P}, p \mapsto \frac{d}{dx}p$ is linear. It is called the **differential operator**[1]. ◇

Example 6.4

Let $\mathscr{C}^0([0,1])$ and $\mathscr{C}^1([0,1])$ be the vector spaces of all continuous and continuously differential functions with domain $[0,1]$, respectively (see Example 5.5). Then the differential operator

$$\frac{d}{dx}\colon \mathscr{C}^1([0,1]) \to \mathscr{C}^0([0,1]), f \mapsto f' = \frac{d}{dx}f$$

is a linear map. ◇

[1] A transformation that maps a function into another function is usually called an **operator**.

Example 6.5 Let $\mathscr{L}$ be the vector space of all random variables X on some given probability space that have an expectation $\mathbb{E}(X)$. Then the map

$$\mathbb{E}\colon \mathscr{L} \to \mathbb{R}, X \mapsto \mathbb{E}(X)$$

is a linear map. ♢

Linear maps can be described by their range and their preimage of 0.

Definition 6.6 **Kernel and image.** Let $\phi\colon \mathcal{V} \to \mathcal{W}$ be a linear map.

(i) The **kernel** (or **nullspace**) of ϕ is the preimage of 0, i.e.,

$$\ker(\phi) = \{\mathbf{x} \in \mathcal{V} : \phi(\mathbf{x}) = 0\}\,.$$

(ii) The **image** (or **range**) of ϕ is the set

$$\operatorname{Im}(\phi) = \phi(\mathcal{V}) = \{\mathbf{y} \in \mathcal{W} : \exists \mathbf{x} \in \mathcal{V},\ \text{s.t. } \phi(\mathbf{x}) = \mathbf{y}\}\,.$$

Theorem 6.7 Let $\phi\colon \mathcal{V} \to \mathcal{W}$ be a linear map. Then $\ker(\phi) \subseteq \mathcal{V}$ and $\operatorname{Im}(\phi) \subseteq \mathcal{W}$ are vector spaces.

PROOF. By Definition 5.6 we have to show that an arbitrary linear combination of two elements of the subset is also an element of the set.

Let $\mathbf{x}, \mathbf{y} \in \ker(\phi)$ and $\alpha, \beta \in \mathbb{R}$. Then by definition of $\ker(\phi)$, $\phi(\mathbf{x}) = \phi(\mathbf{y}) = 0$ and thus $\phi(\alpha\mathbf{x} + \beta\mathbf{y}) = \alpha\phi(\mathbf{x}) + \beta\phi(\mathbf{y}) = 0$. Consequently $\alpha\mathbf{x} + \beta\mathbf{y} \in \ker(\phi)$ and thus $\ker(\phi)$ is a subspace of $\mathcal{V}$.

For the second statement assume that $\mathbf{x}, \mathbf{y} \in \operatorname{Im}(\phi)$. Then there exist two vectors $\mathbf{u}, \mathbf{v} \in \mathcal{V}$ such that $\mathbf{x} = \phi(\mathbf{u})$ and $\mathbf{y} = \phi(\mathbf{v})$. Hence for any $\alpha, \beta \in \mathbb{R}$, $\alpha\mathbf{x} + \beta\mathbf{y} = \alpha\phi(\mathbf{u}) + \beta\phi(\mathbf{v}) = \phi(\alpha\mathbf{u} + \beta\mathbf{v}) \in \operatorname{Im}(\phi)$. □

Theorem 6.8 Let $\phi\colon \mathcal{V} \to \mathcal{W}$ be a linear map and let $B = \{\mathbf{v}_1, \dots, \mathbf{v}_n\}$ be a basis of $\mathcal{V}$. Then $\operatorname{Im}(\phi)$ is spanned by the vectors $\phi(\mathbf{v}_1), \dots, \phi(\mathbf{v}_n)$.

PROOF. For every $\mathbf{x} \in \mathcal{V}$ we have $\mathbf{x} = \sum_{i=1}^{n} c_i(\mathbf{x})\mathbf{v}_i$, where $\mathbf{c}(\mathbf{x})$ is the coefficient vector of $\mathbf{x}$ with respect to B. Then by the linearity of ϕ we find $\phi(\mathbf{x}) = \phi\left(\sum_{i=1}^{n} c_i(\mathbf{x})\mathbf{v}_i\right) = \sum_{i=1}^{n} c_i(\mathbf{x})\phi(\mathbf{v}_i)$. Thus $\phi(\mathbf{x})$ can be represented as a linear combination of the vectors $\phi(\mathbf{v}_1), \dots, \phi(\mathbf{v}_n)$. □

We will see below that the dimensions of these vector spaces determine whether a linear map is invertible. First we show that there is a strong relation between their dimensions.

Theorem 6.9 **Dimension theorem for linear maps.** Let $\phi\colon \mathcal{V} \to \mathcal{W}$ be a linear map. Then

$$\dim \ker(\phi) + \dim \operatorname{Im}(\phi) = \dim \mathcal{V}\,.$$

Proof. Let $\{\mathbf{v}_1,\dots,\mathbf{v}_k\}$ form a basis of $\ker(\phi)\subseteq\mathcal{V}$. Then it can be extended into a basis $\{\mathbf{v}_1,\dots,\mathbf{v}_k,\mathbf{w}_1,\dots,\mathbf{w}_n\}$ of $\mathcal{V}$, where $k+n=\dim\mathcal{V}$. For any $\mathbf{x}\in\mathcal{V}$ there exist unique coefficients a_i and b_j such that $\mathbf{x}=\sum_{i=1}^k a_i\mathbf{v}_i+\sum_{j=1}^n b_i\mathbf{w}_j$. By the linearity of ϕ we then have

$$\phi(\mathbf{x})=\sum_{i=1}^k a_i\phi(\mathbf{v}_i)+\sum_{j=1}^n b_i\phi(\mathbf{w}_j)=\sum_{j=1}^n b_i\phi(\mathbf{w}_j)$$

i.e., $\{\phi(\mathbf{w}_1),\dots,\phi(\mathbf{w}_n)\}$ spans $\operatorname{Im}(\phi)$. It remains to show that this set is linearly independent. In fact, if $\sum_{j=1}^n b_i\phi(\mathbf{w}_j)=0$ then $\phi(\sum_{j=1}^n b_i\mathbf{w}_j)=0$ and hence $\sum_{j=1}^n b_i\mathbf{w}_j\in\ker(\phi)$. Thus there exist coefficients c_i such that $\sum_{j=1}^n b_i\mathbf{w}_j=\sum_{i=1}^k c_i\mathbf{v}_i$, or equivalently, $\sum_{j=1}^n b_i\mathbf{w}_j+\sum_{i=1}^k(-c_i)\mathbf{v}_i=0$. However, as $\{\mathbf{v}_1,\dots,\mathbf{v}_k,\mathbf{w}_1,\dots,\mathbf{w}_n\}$ forms a basis of $\mathcal{V}$ all coefficients b_j and c_i must be zero and consequently the vectors $\{\phi(\mathbf{w}_1),\dots,\phi(\mathbf{w}_n)\}$ are linearly independent and form a basis for $\operatorname{Im}(\phi)$. Thus the statement follows. □

Lemma 6.10

Let $\phi\colon\mathcal{V}\to\mathcal{W}$ be a linear map and let $B=\{\mathbf{v}_1,\dots,\mathbf{v}_n\}$ be a basis of $\mathcal{V}$. Then the vectors $\phi(\mathbf{v}_1),\dots,\phi(\mathbf{v}_n)$ are linearly independent if and only if $\ker(\phi)=\{0\}$.

Proof. By Theorem 6.8, $\phi(\mathbf{v}_1),\dots,\phi(\mathbf{v}_n)$ spans $\operatorname{Im}(\phi)$, that is, for every $\mathbf{x}\in\mathcal{V}$ we have $\phi(\mathbf{x})=\sum_{i=1}^n c_i(\mathbf{x})\phi(\mathbf{v}_i)$ where $\mathbf{c}(\mathbf{x})$ denotes the coefficient vector of $\mathbf{x}$ with respect to B. Thus if $\phi(\mathbf{v}_1),\dots,\phi(\mathbf{v}_n)$ are linearly independent, then $\sum_{i=1}^n c_i(\mathbf{x})\phi(\mathbf{v}_i)=0$ implies $\mathbf{c}(\mathbf{x})=0$ and hence $\mathbf{x}=0$. But then $\mathbf{x}\in\ker(\phi)$. Conversely, if $\ker(\phi)=\{0\}$, then $\phi(\mathbf{x})=\sum_{i=1}^n c_i(\mathbf{x})\phi(\mathbf{v}_i)=0$ implies $\mathbf{x}=0$ and hence $\mathbf{c}(\mathbf{x})=0$. But then vectors $\phi(\mathbf{v}_1),\dots,\phi(\mathbf{v}_n)$ are linearly independent, as claimed. □

Recall that a function $\phi\colon\mathcal{V}\to\mathcal{W}$ is **invertible**, if there exists a function $\phi^{-1}\colon\mathcal{W}\to\mathcal{V}$ such that $\left(\phi^{-1}\circ\phi\right)(\mathbf{x})=\mathbf{x}$ for all $\mathbf{x}\in\mathcal{V}$ and $\left(\phi\circ\phi^{-1}\right)(\mathbf{y})=\mathbf{y}$ for all $\mathbf{y}\in\mathcal{W}$. Such a function exists if ϕ is one-to-one and onto.

A linear map ϕ is **onto** if $\operatorname{Im}(\phi)=\mathcal{W}$. It is **one-to-one** if for each $\mathbf{y}\in\mathcal{W}$ there exists at most one $\mathbf{x}\in\mathcal{V}$ such that $\mathbf{y}=\phi(\mathbf{x})$.

Lemma 6.11

A linear map $\phi\colon\mathcal{V}\to\mathcal{W}$ is one-to-one if and only if $\ker(\phi)=\{0\}$.

Proof. See Problem 6.5.

Lemma 6.12

Let $\phi\colon\mathcal{V}\to\mathcal{W}$ be a linear map. Then $\dim\mathcal{V}=\dim\operatorname{Im}(\phi)$ if and only if $\ker(\phi)=\{0\}$.

Proof. By Theorem 6.9, $\dim\operatorname{Im}(\phi)=\dim\mathcal{V}-\dim\ker(\phi)$. Notice that $\dim\{0\}=0$. Thus the result follows from Lemma 6.11. □

Theorem 6.13

A linear map $\phi\colon\mathcal{V}\to\mathcal{W}$ is invertible if and only if $\dim\mathcal{V}=\dim\mathcal{W}$ and $\ker(\phi)=\{0\}$.

PROOF. Notice that ϕ is onto if and only if $\dim \operatorname{Im}(\phi) = \dim \mathcal{W}$. By Lemmata 6.11 and 6.12, ϕ is invertible if and only if $\dim \mathcal{V} = \dim \operatorname{Im}(\phi)$ and $\ker(\phi) = \{0\}$. □

Theorem 6.14 Let $\phi\colon \mathcal{V} \to \mathcal{W}$ be a linear map with $\dim \mathcal{V} = \dim \mathcal{W}$.

(1) If there exists a function $\psi\colon \mathcal{W} \to \mathcal{V}$ such that $(\psi \circ \phi)(\mathbf{x}) = \mathbf{x}$ for all $\mathbf{x} \in \mathcal{V}$, then $(\phi \circ \psi)(\mathbf{y}) = \mathbf{y}$ for all $\mathbf{y} \in \mathcal{W}$.

(2) If there exists a function $\chi\colon \mathcal{W} \to \mathcal{V}$ such that $(\phi \circ \chi)(\mathbf{y}) = \mathbf{y}$ for all $\mathbf{y} \in \mathcal{W}$, then $(\chi \circ \phi)(\mathbf{x}) = \mathbf{x}$ for all $\mathbf{x} \in \mathcal{V}$.

PROOF. It remains to show that ϕ is invertible in both cases.
(1) If ψ exists, then ϕ must be one-to-one. Thus $\ker(\phi) = \{0\}$ by Lemma 6.11 and consequently ϕ is invertible by Theorem 6.13.
(2) We can use (1) to conclude that $\chi^{-1} = \phi$. Hence $\phi^{-1} = \chi$ and the statement follows. □

An immediate consequence of this Theorem is the existence of ψ or χ implies the existence of the other one. Consequently, this also implies that ϕ is invertible and $\phi^{-1} = \psi = \chi$.

6.2 Matrices and Linear Maps

In Section 5.3 we have seen that the $\mathbb{R}^n$ can be interpreted as *the* vector space of dimension n. Example 6.2 shows us that any $m \times n$ matrix $\mathbf{A}$ defines a linear map between $\mathbb{R}^n$ and $\mathbb{R}^m$. The following theorem tells us that there is also a one-to-one correspondence between matrices and linear maps. Thus matrices are *the* representations of linear maps.

Theorem 6.15 Let $\phi\colon \mathbb{R}^n \to \mathbb{R}^m$ be a linear map. Then there exists an $m \times n$ matrix $\mathbf{A}_\phi$ such that $\phi(\mathbf{x}) = \mathbf{A}_\phi \mathbf{x}$.

PROOF. Let $\mathbf{a}_i = \phi(\mathbf{e}_i)$ denote the images of the elements of the canonical basis $\{\mathbf{e}_1, \dots, \mathbf{e}_n\}$. Let $\mathbf{A} = (\mathbf{a}_1, \dots, \mathbf{a}_n)$ be the matrix with column vectors $\mathbf{a}_i$. Notice that $\mathbf{A}\mathbf{e}_i = \mathbf{a}_i$. Now we find for every $\mathbf{x} = (x_1, \dots, x_n)' \in \mathbb{R}^n$, $\mathbf{x} = \sum_{i=1}^n x_i \mathbf{e}_i$ and therefore

$$\phi(\mathbf{x}) = \phi\left(\sum_{i=1}^n x_i \mathbf{e}_i\right) = \sum_{i=1}^n x_i \phi(\mathbf{e}_i) = \sum_{i=1}^n x_i \mathbf{a}_i = \sum_{i=1}^n x_i \mathbf{A}\mathbf{e}_i = \mathbf{A}\sum_{i=1}^n x_i \mathbf{e}_i = \mathbf{A}\mathbf{x}$$

as claimed. □

Now assume that we have two linear maps $\phi\colon \mathbb{R}^n \to \mathbb{R}^m$ and $\psi\colon \mathbb{R}^m \to \mathbb{R}^k$ with corresponding matrices $\mathbf{A}$ and $\mathbf{B}$, resp. The map composition $\psi \circ \phi$ is then given by $(\psi \circ \phi)(\mathbf{x}) = \psi(\phi(\mathbf{x})) = \mathbf{B}(\mathbf{A}\mathbf{x}) = (\mathbf{B}\mathbf{A})\mathbf{x}$. Thus matrix multiplication corresponds to map composition.

If the linear map $\phi\colon \mathbb{R}^n \to \mathbb{R}^n, \mathbf{x} \mapsto \mathbf{A}\mathbf{x}$, is invertible, then matrix $\mathbf{A}$ is also invertible and $\mathbf{A}^{-1}$ describes the inverse map ϕ^{-1}.

By Theorem 6.15 a linear map ϕ and its corresponding matrix $\mathbf{A}$ are closely related. Thus all definitions and theorems about linear maps may be applied to matrices. For example, the kernel of matrix $\mathbf{A}$ is the set

$$\ker(\mathbf{A}) = \{\mathbf{x} \colon \mathbf{A}\mathbf{x} = 0\}\,.$$

The following result is an immediate consequence of our considerations and Theorem 6.14.

Theorem 6.16

Let $\mathbf{A}$ be some square matrix.

(a) If there exists a square matrix $\mathbf{B}$ such that $\mathbf{AB} = \mathbf{I}$, then $\mathbf{A}$ is invertible and $\mathbf{A}^{-1} = \mathbf{B}$.

(b) If there exists a square matrix $\mathbf{C}$ such that $\mathbf{CA} = \mathbf{I}$, then $\mathbf{A}$ is invertible and $\mathbf{A}^{-1} = \mathbf{C}$.

The following result is very convenient.

Corollary 6.17

Let $\mathbf{A}$ and $\mathbf{B}$ be two square matrices with $\mathbf{AB} = \mathbf{I}$. Then both $\mathbf{A}$ and $\mathbf{B}$ are invertible and $\mathbf{A}^{-1} = \mathbf{B}$ and $\mathbf{B}^{-1} = \mathbf{A}$.

6.3 Rank of a Matrix

Theorem 6.8 tells us that the columns of a matrix $\mathbf{A}$ span the image of a linear map ϕ induced by $\mathbf{A}$. Consequently, by Theorem 5.20 we get a basis of $\operatorname{Im}(\phi)$ by a maximal linearly independent subset of these column vectors. The dimension of the image is then the size of this subset. This motivates the following notion.

Definition 6.18

The **rank** of a matrix $\mathbf{A}$ is the maximal number of linearly independent columns of $\mathbf{A}$.

By the above considerations we immediately have the following lemmata.

Lemma 6.19

For any matrix $\mathbf{A}$,

$$\operatorname{rank}(\mathbf{A}) = \dim \operatorname{Im}(\mathbf{A})\,.$$

Lemma 6.20

Let $\mathbf{A}$ be an $m \times n$ matrix. If $\mathbf{T}$ is an invertible $m \times m$ matrix, then $\operatorname{rank}(\mathbf{TA}) = \operatorname{rank}(\mathbf{A})$.

PROOF. Let ϕ and ψ be the linear maps induced by $\mathbf{A}$ and $\mathbf{T}$, resp. Theorem 6.13 states that $\ker(\psi) = \{0\}$. Hence by Theorem 6.9, ψ is one-to-one. Hence $\operatorname{rank}(\mathbf{TA}) = \dim(\psi(\phi(\mathbb{R}^n))) = \dim(\phi(\mathbb{R}^n)) = \operatorname{rank}(\mathbf{A})$, as claimed. □

Sometimes we are interested in the dimension of the kernel of a matrix.

Definition 6.21

The **nullity** of matrix $\mathbf{A}$ is the dimension of the kernel (nullspace) of $\mathbf{A}$.

Theorem 6.22

Rank-nullity theorem. Let $\mathbf{A}$ be an $m \times n$ matrix. Then

$$\operatorname{rank}(\mathbf{A}) + \operatorname{nullity}(\mathbf{A}) = n.$$

PROOF. By Lemma 6.19 and Theorem 6.9 we find $\operatorname{rank}(\mathbf{A}) + \operatorname{nullity}(\mathbf{A}) = \dim \operatorname{Im}(\mathbf{A}) + \dim \ker(\mathbf{A}) = n$. □

Theorem 6.23

Let $\mathbf{A}$ be an $m \times n$ matrix and $\mathbf{B}$ be an $n \times k$ matrix. Then

$$\operatorname{rank}(\mathbf{AB}) \leq \min\{\operatorname{rank}(\mathbf{A}), \operatorname{rank}(\mathbf{B})\}.$$

PROOF. Let ϕ and ψ be the maps represented by $\mathbf{A}$ and $\mathbf{B}$, respectively. Recall that $\mathbf{AB}$ correspond to map composition $\phi \circ \psi$. Obviously, $\operatorname{Im}(\phi \circ \psi) \subseteq \operatorname{Im}\phi$. Hence $\operatorname{rank}(\mathbf{AB}) = \dim \operatorname{Im}(\phi \circ \psi) \leq \dim \operatorname{Im}(\phi) = \operatorname{rank}(\mathbf{A})$. Similarly, $\operatorname{Im}(\phi \circ \psi)$ is spanned by $\phi(S)$ where S is any linearly independent subset of $\operatorname{Im}(\psi)$. Hence $\operatorname{rank}(\mathbf{AB}) = \dim \operatorname{Im}(\phi \circ \psi) \leq \dim \operatorname{Im}(\psi) = \operatorname{rank}(\mathbf{B})$. Thus the result follows. □

Our notion of *rank* in Definition 6.18 is sometimes also referred to as *column rank* of matrix $\mathbf{A}$. One may also define the **row rank** of $\mathbf{A}$ as the maximal number of linearly independent rows of $\mathbf{A}$. However, column rank and row rank always coincide.

Theorem 6.24

For any matrix $\mathbf{A}$,

$$\operatorname{rank}(\mathbf{A}) = \operatorname{rank}(\mathbf{A}').$$

For the proof of this theorem we first show the following result.

Lemma 6.25

Let $\mathbf{A}$ be a $m \times n$ matrix. Then $\operatorname{rank}(\mathbf{A}'\mathbf{A}) = \operatorname{rank}(\mathbf{A})$.

PROOF. We show that $\ker(\mathbf{A}'\mathbf{A}) = \ker(\mathbf{A})$. Obviously, $\mathbf{x} \in \ker(\mathbf{A})$ implies $\mathbf{A}'\mathbf{A}\mathbf{x} = \mathbf{A}'0 = 0$ and thus $\ker(\mathbf{A}) \subseteq \ker(\mathbf{A}'\mathbf{A})$. Now assume that $\mathbf{x} \in \ker(\mathbf{A}'\mathbf{A})$. Then $\mathbf{A}'\mathbf{A}\mathbf{x} = 0$ and we find $0 = \mathbf{x}'\mathbf{A}'\mathbf{A}\mathbf{x} = (\mathbf{A}\mathbf{x})'(\mathbf{A}\mathbf{x})$ which implies that $\mathbf{A}\mathbf{x} = 0$ so that $\mathbf{x} \in \ker(\mathbf{A})$. Hence $\ker(\mathbf{A}'\mathbf{A}) \subseteq \ker(\mathbf{A})$ and, consequently, $\ker(\mathbf{A}'\mathbf{A}) = \ker(\mathbf{A})$. Now notice that $\mathbf{A}'\mathbf{A}$ is an $n \times n$ matrix. Theorem 6.22 then implies

$$\begin{aligned}\operatorname{rank}(\mathbf{A}'\mathbf{A}) - \operatorname{rank}(\mathbf{A}) &= \big(n - \operatorname{nullity}(\mathbf{A}'\mathbf{A})\big) - (n - \operatorname{nullity}(\mathbf{A})) \\ &= \operatorname{nullity}(\mathbf{A}) - \operatorname{nullity}(\mathbf{A}'\mathbf{A}) = 0\end{aligned}$$

and thus $\operatorname{rank}(\mathbf{A}'\mathbf{A}) = \operatorname{rank}(\mathbf{A})$, as claimed. □

PROOF OF THEOREM 6.24. By Theorem 6.23 and Lemma 6.25 we find

$$\operatorname{rank}(\mathbf{A}') \geq \operatorname{rank}(\mathbf{A}'\mathbf{A}) = \operatorname{rank}(\mathbf{A}).$$

As this statement remains true if we replace $\mathbf{A}$ by its transpose $\mathbf{A}'$ we have $\operatorname{rank}(\mathbf{A}) \geq \operatorname{rank}(\mathbf{A}')$ and thus the statement follows. □

Corollary 6.26

Let $\mathbf{A}$ be an $m \times n$ matrix. Then

$$\operatorname{rank}(\mathbf{A}) \leq \min\{m, n\}.$$

Finally, we give a necessary and sufficient condition for invertibility of a square matrix.

Definition 6.27

An $n \times n$ matrix $\mathbf{A}$ is called **regular** if it has **full rank**, i.e., if $\operatorname{rank}(\mathbf{A}) = n$.

Theorem 6.28

A square matrix $\mathbf{A}$ is invertible if and only if it is regular.

Proof. By Theorem 6.13 a matrix is invertible if and only if $\operatorname{nullity}(\mathbf{A}) = 0$ (i.e., $\ker(\mathbf{A}) = \{0\}$). Then $\operatorname{rank}(\mathbf{A}) = n - \operatorname{nullity}(\mathbf{A}) = n$ by Theorem 6.22. □

6.4 Similar Matrices

In Section 5.3 we have seen that every vector $\mathbf{x} \in \mathcal{V}$ in some vector space $\mathcal{V}$ of dimension $\dim \mathcal{V} = n$ can be uniformly represented by a coordinate vector $\mathbf{c}(\mathbf{x}) \in \mathbb{R}^n$. However, for this purpose we first have to choose an arbitrary but fixed basis for $\mathcal{V}$. In this sense every finitely generated vector space is "equivalent" (i.e., isomorphic) to the $\mathbb{R}^n$.

However, we also have seen that there is no such thing as *the* basis of a vector space and that coordinate vector $\mathbf{c}(\mathbf{x})$ changes when we change the underlying basis of $\mathcal{V}$. Of course vector $\mathbf{x}$ then remains the same.

In Section 6.2 above we have seen that matrices are the representations of linear maps between $\mathbb{R}^m$ and $\mathbb{R}^n$. Thus if $\phi\colon \mathcal{V} \to \mathcal{W}$ is a linear map, then there is a matrix $\mathbf{A}$ that represents the linear map between the coordinate vectors of vectors in $\mathcal{V}$ and those in $\mathcal{W}$. Obviously matrix $\mathbf{A}$ depends on the chosen bases for $\mathcal{V}$ and $\mathcal{W}$.

Suppose now that $\dim \mathcal{V} = \dim \mathcal{W} = \mathbb{R}^n$. Let $\mathbf{A}$ be an $n \times n$ square matrix that represents a linear map $\phi\colon \mathbb{R}^n \to \mathbb{R}^n$ with respect to some basis B_1. Let $\mathbf{x}$ be a coefficient vector corresponding to basis B_2 and let $\mathbf{U}$ denote the transformation matrix that transforms $\mathbf{x}$ into the coefficient vector corresponding to basis B_1. Then we find:

$$\begin{tikzcd}
\text{basis } B_1\colon & \mathbf{Ux} \arrow[r, "\mathbf{A}"] & \mathbf{AUx} \arrow[dd, "\mathbf{U}^{-1}"] & \\
& & & \text{hence} \quad \mathbf{C}\mathbf{x} = \mathbf{U}^{-1}\mathbf{AUx} \\
\text{basis } B_2\colon & \mathbf{x} \arrow[uu, "\mathbf{U}"] \arrow[r, "\mathbf{C}"] & \mathbf{U}^{-1}\mathbf{AUx} &
\end{tikzcd}$$

That is, if $\mathbf{A}$ represents a linear map corresponding to basis B_1, then $\mathbf{C} = \mathbf{U}^{-1}\mathbf{AU}$ represents the same linear map corresponding to basis B_2.

Definition 6.29

Two $n \times n$ matrices $\mathbf{A}$ and $\mathbf{C}$ are called **similar** if $\mathbf{C} = \mathbf{U}^{-1}\mathbf{AU}$ for some invertible $n \times n$ matrix $\mathbf{U}$.

— Summary

- A *Linear map* ϕ preserve the linear structure, i.e.,

$$\phi(\alpha\mathbf{x} + \beta\mathbf{y}) = \alpha\phi(\mathbf{x}) + \beta\phi(\mathbf{y}) .$$

- *Kernel* and *image* of a linear map $\phi\colon \mathcal{V} \to \mathcal{W}$ are subspaces of $\mathcal{V}$ and $\mathcal{W}$, resp.
- $\operatorname{Im}(\phi)$ is spanned by the images of a basis of $\mathcal{V}$.
- A linear map $\phi\colon \mathcal{V} \to \mathcal{W}$ is *invertible* if and only if $\dim \mathcal{V} = \dim \mathcal{W}$ and $\ker(\phi) = \{0\}$.
- Linear maps are represented by matrices. The corresponding matrix depends on the chosen bases of the vector spaces.
- Matrices are called *similar* if they describe the same linear map but w.r.t. different bases.
- The *rank* of a matrix is the dimension of the image of the corresponding linear map.
- Matrix multiplication corresponds to map composition. The inverse of a matrix corresponds to the corresponding inverse linear map.
- A matrix is invertible if and only if it is a square matrix and *regular*, i.e., has full rank.

— Exercises

6.1 Let $\mathscr{P}_2 = \{a_0 + a_1 x + a_2 x^2 \colon a_i \in \mathbb{R}\}$ be the vector space of all polynomials of order less than or equal to 2 equipped with point-wise addition and scalar multiplication. Then $B = \{1, x, x^2\}$ is a basis of $\mathscr{P}_2$ (see Example 5.29). Let $\phi = \frac{d}{dx} \colon \mathscr{P}_2 \to \mathscr{P}_2$ be the differential operator on $\mathscr{P}_2$ (see Example 6.3).

(a) What is the kernel of ϕ? Give a basis for $\ker(\phi)$.

(b) What is the image of ϕ? Give a basis for $\operatorname{Im}(\phi)$.

(c) For the given basis B represent map ϕ by a matrix $\mathbf{D}$.

(d) The first three so called Laguerre polynomials are $\ell_0(x) = 1$, $\ell_1(x) = 1 - x$, and $\ell_2(x) = \frac{1}{2}\left(x^2 - 4x + 2\right)$.
Then $B_\ell = \{\ell_0(x), \ell_1(x), \ell_2(x)\}$ also forms a basis of $\mathscr{P}_2$. What is the transformation matrix $\mathbf{U}_\ell$ that transforms the coefficient vector of a polynomial p with respect to basis B into its coefficient vector with respect to basis B_ℓ?

(e) For basis B_ℓ represent map ϕ by a matrix $\mathbf{D}_\ell$.

— Problems

6.2 Verify the claim in Example 6.2.

6.3 Show that the following statement is equivalent to Definition 6.1:

Let $\mathscr{V}$ and $\mathscr{W}$ be two vector spaces. A transformation $\phi \colon \mathscr{V} \to \mathscr{W}$ is called a *linear map* if for for all $\mathbf{x}, \mathbf{y} \in \mathscr{V}$ and $\alpha, \beta \in \mathbb{R}$ the following holds:

$$\phi(\alpha\mathbf{x} + \beta\mathbf{y}) = \alpha\phi(\mathbf{x}) + \beta\phi(\mathbf{y})$$

for all $\mathbf{x}, \mathbf{y} \in \mathscr{V}$ and all $\alpha, \beta \in \mathbb{R}$.

6.4 Let $\phi \colon \mathscr{V} \to \mathscr{W}$ be a linear map and let $B = \{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ be a basis of $\mathscr{V}$. Give a necessary and sufficient for $\operatorname{span}\left(\phi(\mathbf{v}_1), \ldots, \phi(\mathbf{v}_n)\right)$ being a basis of $\operatorname{Im}(\phi)$.

6.5 Prove Lemma 6.11.

HINT: We have to prove two statements:
(1) ϕ is one-to-one $\Rightarrow \ker(\phi) = \{0\}$.
(2) $\ker(\phi) = \{0\} \Rightarrow \phi$ is one-to-one.
For (2) use the fact that if $\phi(\mathbf{x}_1) = \phi(\mathbf{x}_2)$, then $\mathbf{x}_1 - \mathbf{x}_2$ must be an element of the kernel. (Why?)

6.6 Let $\phi \colon \mathbb{R}^n \to \mathbb{R}^m$, $\mathbf{x} \mapsto \mathbf{y} = \mathbf{A}\mathbf{x}$ be a linear map, where $\mathbf{A} = (\mathbf{a}_1, \ldots, \mathbf{a}_n)$. Show that the column vectors of matrix $\mathbf{A}$ span $\operatorname{Im}(\phi)$, i.e.,

$$\operatorname{span}(\mathbf{a}_1, \ldots, \mathbf{a}_n) = \operatorname{Im}(\phi)\,.$$

6.7 Let $\mathbf{A}$ be an $m \times n$ matrix. If $\mathbf{T}$ is an invertible $n \times n$ matrix, then $\operatorname{rank}(\mathbf{AT}) = \operatorname{rank}(\mathbf{A})$.

6.8 Prove Corollary 6.26.

6.9 Disprove the following statement:
For any $m \times n$ matrix $\mathbf{A}$ and any $n \times k$ matrix $\mathbf{B}$ it holds that $\operatorname{rank}(\mathbf{AB}) = \min\{\operatorname{rank}(\mathbf{A}), \operatorname{rank}(\mathbf{B})\}$.

6.10 Show that two similar matrices have the same rank.

6.11 Is the converse of the statement in Problem 6.10 true?
Prove or disprove:
If two $n \times n$ matrices have the same rank then they are similar.

HINT: Consider simple diagonal matrices.

6.12 Let $\mathbf{U}$ be the transformation matrix for a change of basis (see Definition 5.30). Argue why $\mathbf{U}$ is invertible.

HINT: Use the fact that $\mathbf{U}$ describes a linear map between the sets of coefficient vectors for two given bases. These have the same dimension.

7
Linear Equations

We want to compute *dimensions and bases of kernel and image.*

7.1 Linear Equations

Definition 7.1

A system of m linear equations in n unknowns is given by the following set of equations:

$$\begin{array}{ccccccccc} a_{11}x_1 & + & a_{12}x_2 & + & \cdots & + & a_{1n}x_n & = & b_1 \\ a_{21}x_1 & + & a_{22}x_2 & + & \cdots & + & a_{2n}x_n & = & b_2 \\ \vdots & & \vdots & & \ddots & & \vdots & & \vdots \\ a_{m1}x_1 & + & a_{m2}x_2 & + & \cdots & + & a_{mn}x_n & = & b_m \end{array}$$

By means of matrix algebra it can be written in much more compact form as (see Problem 7.2)

$$\mathbf{A}\mathbf{x} = \mathbf{b}\,.$$

The matrix

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{pmatrix}$$

is called the **coefficient matrix** and the vectors

$$\mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \qquad \text{and} \qquad \mathbf{b} = \begin{pmatrix} b_1 \\ \vdots \\ b_m \end{pmatrix}$$

contain the unknowns x_i and the constants b_j on the right hand side.

Definition 7.2

A linear equation $\mathbf{A}\mathbf{x} = 0$ is called **homogeneous**.
A linear equation $\mathbf{A}\mathbf{x} = \mathbf{b}$ with $\mathbf{b} \neq 0$ is called **inhomogeneous**.

Observe that the set of solutions of the homogeneous linear equation $\mathbf{A}\mathbf{x} = 0$ and is just the kernel of the coefficient matrix, $\ker(\mathbf{A})$, and thus forms a vector space. The set of solutions of an inhomogeneous linear equation $\mathbf{A}\mathbf{x} = \mathbf{b}$ can be derived from $\ker(\mathbf{A})$ as well.

Lemma 7.3

Let $\mathbf{x}_0$ and $\mathbf{y}_0$ be two solutions of the inhomogeneous equation $\mathbf{A}\mathbf{x} = \mathbf{b}$. Then $\mathbf{x}_0 - \mathbf{y}_0$ is a solution of the homogeneous equation $\mathbf{A}\mathbf{x} = 0$.

Theorem 7.4

Let $\mathbf{x}_0$ be a particular solution of the inhomogeneous equation $\mathbf{A}\mathbf{x} = \mathbf{b}$, then the set of all solutions of $\mathbf{A}\mathbf{x} = \mathbf{b}$ is given by

$$\mathscr{S} = \mathbf{x}_0 + \ker(\mathbf{A}) = \{\mathbf{x} = \mathbf{x}_0 + \mathbf{z} : \mathbf{z} \in \ker(\mathbf{A})\}\,.$$

PROOF. See Problem 7.3.

Set $\mathscr{S}$ is an example of an *affine subspace* of $\mathbb{R}^n$.

Definition 7.5

Let $\mathbf{x}_0 \in \mathbb{R}^n$ be a vector and $\mathscr{S} \subseteq \mathbb{R}^n$ be a subspace. Then the set $\mathbf{x}_0 + \mathscr{S} = \{\mathbf{x} = \mathbf{x}_0 + \mathbf{z} : \mathbf{z} \in \mathscr{S}\}$ is called an **affine subspace** of $\mathbb{R}^n$

7.2 Gauß Elimination

A linear equation $\mathbf{A}\mathbf{x} = \mathbf{b}$ can be solved by transforming it into a simpler form called *row echelon form*.

Definition 7.6

A matrix $\mathbf{A}$ is said to be in **row echelon form** if the following holds:

(i) All nonzero rows (rows with at least one nonzero element) are above any rows of all zeroes, and

(ii) The leading coefficient (i.e., the first nonzero number from the left, also called the **pivot**) of a nonzero row is always strictly to the right of the leading coefficient of the row above it.

It is sometimes convenient to work with an even simpler form.

Definition 7.7

A matrix $\mathbf{A}$ is said to be in **row reduced echelon form** if the following holds:

(i) All nonzero rows (rows with at least one nonzero element) are above any rows of all zeroes, and

(ii) The leading coefficient of a nonzero row is always strictly to the right of the leading coefficient of the row above it. It is 1 and is the only nonzero entry in its column. Such columns are then called **pivotal**.

Any coefficient matrix $\mathbf{A}$ can be transformed into a matrix $\mathbf{R}$ that is in row (reduced) echelon form by means of **elementary row operations** (see Problem 7.6):

(E1) Switch two rows.

(E2) Multiply some row with $\alpha \neq 0$.

(E3) Add some multiple of a row to another row.

These row operations can be performed by means of *elementary matrices*, i.e., matrices that differs from the identity matrix by one single elementary row operation. These matrices are always invertible, see Problem 7.4.

The procedure works due to the following lemma which tells use how we obtain equivalent linear equations that have the same solutions.

Lemma 7.8

Let $\mathbf{A}$ be an $m \times n$ coefficient matrix and $\mathbf{b}$ the vector of constants. If $\mathbf{T}$ is an invertible $m \times m$ matrix, then the linear equations

$$\mathbf{A}\mathbf{x} = \mathbf{b} \quad \text{and} \quad \mathbf{T}\mathbf{A}\mathbf{x} = \mathbf{T}\mathbf{b}$$

are equivalent. That is, they have the same solutions.

PROOF. See Problem 7.5.

Gauß elimination now iteratively applies elementary row operations until a row (reduced) echelon form is obtained. Mathematically spoken: In each step of the iteration we multiply a corresponding elementary matrix $\mathbf{T}_k$ from the left to the equation $\mathbf{T}_{k-1}\cdots\mathbf{T}_1\mathbf{A}\mathbf{x} = \mathbf{T}_{k-1}\cdots\mathbf{T}_1\mathbf{b}$. For practical reasons one usually uses the augmented coefficient matrix.

Theorem 7.9

For every matrix $\mathbf{A}$ there exists a sequence of elementary row operations $\mathbf{T}_1,\ldots\mathbf{T}_k$ such that $\mathbf{R} = \mathbf{T}_k\cdots\mathbf{T}_1\mathbf{A}$ is in row (reduced) echelon form.

PROOF. See Problem 7.6.

For practical reasons one augments the coefficient matrix $\mathbf{A}$ of a linear equation by the constant vector $\mathbf{b}$. Thus the row operations can be performed on $\mathbf{A}$ and $\mathbf{b}$ simultaneously.

Definition 7.10

Let $\mathbf{A}\mathbf{x} = \mathbf{b}$ be a linear equation with coefficient matrix $\mathbf{A} = (\mathbf{a}_1,\ldots\mathbf{a}_n)$. Then matrix $\mathbf{A}_b = (\mathbf{A},\mathbf{b}) = (\mathbf{a}_1,\ldots\mathbf{a}_n,\mathbf{b})$ is called the **augmented coefficient matrix** of the linear equation.

When the coefficient matrix is in row echelon form, then the solution $\mathbf{x}$ of the linear equation $\mathbf{A}\mathbf{x} = \mathbf{b}$ can be easily obtained by means of an iterative process called **back substitution**. When it is in row *reduced* echelon form it is even simpler: We get a particular solution $\mathbf{x}_0$ by setting all variables that belong to non-pivotal columns to 0. Then we solve the resulting linear equations for the variables that corresponds to the pivotal columns. This is easy as each row reduces to

$$\delta_i x_i = b_i \qquad \text{where } \delta_i \in \{0,1\}.$$

Obviously, these equations can be solved if and only if $\delta_i = 1$ or $b_i = 0$.

We then need a basis of $\ker(\mathbf{A})$ which we easily get from a row reduced echelon form of the homogeneous equation $\mathbf{Ax} = 0$. Notice, that $\ker(\mathbf{A}) = \{0\}$ if there are no non-pivotal columns.

7.3 Image, Kernel and Rank of a Matrix

Once the row reduced echelon form $\mathbf{R}$ is given for a matrix $\mathbf{A}$ we also can easily compute bases for its image and kernel.

Theorem 7.11

Let $\mathbf{R}$ be a row reduced echelon form of some matrix $\mathbf{A}$. Then $\operatorname{rank}(\mathbf{A})$ is equal to the number nonzero rows of $\mathbf{R}$.

PROOF. By Lemma 6.20 and Theorem 7.9, $\operatorname{rank}(\mathbf{R}) = \operatorname{rank}(\mathbf{A})$. It is easy to see that *non*-pivotal columns can be represented as linear combinations of pivotal columns. Hence the pivotal columns span $\operatorname{Im}(\mathbf{R})$. Moreover, the pivotal columns are linearly independent since no two of them have a common non-zero entry. The result then follows from the fact that the number of pivotal columns equal the number of nonzero elements. □

Theorem 7.12

Let $\mathbf{R}$ be a row reduced echelon form of some matrix $\mathbf{A}$. Then the columns of $\mathbf{A}$ that correspond to pivotal columns of $\mathbf{R}$ form a basis of $\operatorname{Im}(\mathbf{A})$.

PROOF. The columns of $\mathbf{A}$ span $\operatorname{Im}(\mathbf{A})$. Let $\mathbf{A}_p$ consists of all columns of $\mathbf{A}$ that correspond to pivotal columns of $\mathbf{R}$. If we apply the same elementary row operations on $\mathbf{A}_p$ as for $\mathbf{A}$ we obtain a row reduced echelon form $\mathbf{R}_p$ where all columns are pivotal. Hence the columns of $\mathbf{A}_p$ are linearly independent and $\operatorname{rank}(\mathbf{A}_p) = \operatorname{rank}(\mathbf{A})$. Thus the columns of $\mathbf{A}_p$ form a basis of $\operatorname{Im}(\mathbf{A})$, as claimed. □

At last we verify other observation about the existence of the solution of an inhomogeneous equation.

Theorem 7.13

Let $\mathbf{Ax} = \mathbf{b}$ be an inhomogeneous linear equation. Then there exists a solution $\mathbf{x}_0$ if and only if $\operatorname{rank}(\mathbf{A}) = \operatorname{rank}(\mathbf{A}_b)$.

PROOF. Recall that $\mathbf{A}_b$ denotes the augmented coefficient matrix. If there exists a solution $\mathbf{x}_0$, then $\mathbf{b} = \mathbf{Ax}_0 \in \operatorname{Im}(\mathbf{A})$ and thus

$$\operatorname{rank}(\mathbf{A}_b) = \dim\operatorname{span}(\mathbf{a}_1,\dots,\mathbf{a}_n,\mathbf{b}) = \dim\operatorname{span}(\mathbf{a}_1,\dots,\mathbf{a}_n) = \operatorname{rank}(\mathbf{A})\,.$$

On the other hand, if no such solution exists, then $\mathbf{b} \not\in \operatorname{span}(\mathbf{a}_1,\dots,\mathbf{a}_k)$ and thus $\operatorname{span}(\mathbf{a}_1,\dots,\mathbf{a}_n) \subset \operatorname{span}(\mathbf{a}_1,\dots,\mathbf{a}_n,\mathbf{b})$. Consequently,

$$\operatorname{rank}(\mathbf{A}_b) = \dim\operatorname{span}(\mathbf{a}_1,\dots,\mathbf{a}_n,\mathbf{b}) > \dim\operatorname{span}(\mathbf{a}_1,\dots,\mathbf{a}_n) = \operatorname{rank}(\mathbf{A})$$

and thus $\operatorname{rank}(\mathbf{A}_b) \neq \operatorname{rank}(\mathbf{A})$. □

— Summary

- A *linear equation* is one that can be written as $\mathbf{Ax} = \mathbf{b}$.
- The set of all solutions of a *homogeneous* linear equation forms a *vector space*.
  The set of all solutions of an *inhomogeneous* linear equation forms an *affine space*.
- Linear equations can be solved by transforming the *augmented coefficient matrix* into *row (reduced) echelon form*.
- This transformation is performed by (invertible) *elementary row operations*.
- Bases of image and kernel of a matrix $\mathbf{A}$ as well as its rank can be computed by transforming the matrix into row reduced echelon form.

— Exercises

7.1 Compute image, kernel and rank of

$$\mathbf{A} = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix}.$$

— Problems

7.2 Verify that a system of linear equations can indeed written in matrix form. Moreover show that each equation $\mathbf{A}\mathbf{x} = \mathbf{b}$ represents a system of linear equations.

7.3 Prove Lemma 7.3 and Theorem 7.4.

7.4 Let $\mathbf{A} = \begin{pmatrix} \mathbf{a}'_1 \\ \vdots \\ \mathbf{a}'_m \end{pmatrix}$ be an $m \times n$ matrix.

(1) Define matrix $\mathbf{T}_{i\leftrightarrow j}$ that switches rows $\mathbf{a}'_i$ and $\mathbf{a}'_j$.

(2) Define matrix $\mathbf{T}_i(\alpha)$ that multiplies row $\mathbf{a}'_i$ by α.

(3) Define matrix $\mathbf{T}_{i\leftarrow j}(\alpha)$ that adds row $\mathbf{a}'_j$ multiplied by α to row $\mathbf{a}'_i$.

For each of these matrices argue why these are invertible and state their respective inverse matrices.

HINT: Use the results from Exercise 4.14 to construct these matrices.

7.5 Prove Lemma 7.8.

7.6 Prove Theorem 7.9. Use a so called *constructive* proof. In this case this means to provide an algorithm that transforms every input matrix **A** into row reduce echelon form by means of elementary row operations. Describe such an algorithm (in words or pseudocode).

8

Euclidean Space

We need a ruler *and a* protractor.

8.1 Inner Product, Norm, and Metric

Inner product. Let $\mathbf{x},\mathbf{y} \in \mathbb{R}^n$. Then Definition 8.1

$$\mathbf{x}'\mathbf{y} = \sum_{i=1}^{n} x_i y_i$$

is called the **inner product** (**dot product**, **scalar product**) of $\mathbf{x}$ and $\mathbf{y}$.

Fundamental properties of inner products. Let $\mathbf{x},\mathbf{y},\mathbf{z} \in \mathbb{R}^n$ and $\alpha,\beta \in \mathbb{R}$. Then the following holds: Theorem 8.2

(1) $\mathbf{x}'\mathbf{y} = \mathbf{y}'\mathbf{x}$ (Symmetry)

(2) $\mathbf{x}'\mathbf{x} \geq 0$ where equality holds if and only if $\mathbf{x} = 0$ (Positive-definiteness)

(3) $(\alpha\mathbf{x} + \beta\mathbf{y})'\mathbf{z} = \alpha\mathbf{x}'\mathbf{z} + \beta\mathbf{y}'\mathbf{z}$ (Linearity)

PROOF. See Problem 8.1.

In our notation the inner product of two vectors $\mathbf{x}$ and $\mathbf{y}$ is just the usual matrix multiplication of the *row* vector $\mathbf{x}'$ with the *column* vector $\mathbf{y}$. However, the formal transposition of the first vector $\mathbf{x}$ is often omitted in the notation of the inner product. Thus one simply writes $\mathbf{x} \cdot \mathbf{y}$. Hence the name *dot* product. This is reflected in many computer algebra systems like *Maxima* where the symbol for matrix multiplication is used to multiply two (column) vectors.

Inner product space. The notion of an *inner product* can be generalized. Let $\mathcal{V}$ be some vector space. Then any function Definition 8.3

$$\langle\cdot,\cdot\rangle : \mathcal{V} \times \mathcal{V} \to \mathbb{R}$$

that satisfies the properties

(i) $\langle \mathbf{x}, \mathbf{y} \rangle = \langle \mathbf{y}, \mathbf{x} \rangle$,

(ii) $\langle \mathbf{x}, \mathbf{x} \rangle \geq 0$ where equality holds if and only if $\mathbf{x} = 0$,

(iii) $\langle \alpha\mathbf{x} + \beta\mathbf{y} \rangle \mathbf{z} = \alpha\langle \mathbf{x}, \mathbf{z} \rangle + \beta\langle \mathbf{y}, \mathbf{z} \rangle$,

is called an **inner product**. A vector space that is equipped with such an inner product is called an **inner product space**. In pure mathematics the symbol $\langle \mathbf{x}, \mathbf{y} \rangle$ is often used to denote the (abstract) inner product of two vectors $\mathbf{x}, \mathbf{y} \in \mathscr{V}$.

Example 8.4

Let $\mathscr{L}$ be the vector space of all random variables X on some given probability space with finite variance $\mathbb{V}(X)$. Then map

$$\langle \cdot, \cdot \rangle \colon \mathscr{L} \times \mathscr{L} \to \mathbb{R}, \; (X, Y) \mapsto \langle X, Y \rangle = \mathbb{E}(XY)$$

is an inner product in $\mathscr{L}$. ♢

Definition 8.5

Euclidean norm. Let $\mathbf{x} \in \mathbb{R}^n$. Then

$$\|\mathbf{x}\| = \sqrt{\mathbf{x}'\mathbf{x}} = \sqrt{\sum_{i=1}^{n} x_i^2}$$

is called the **Euclidean norm** (or *norm* for short) of $\mathbf{x}$.

Theorem 8.6

Cauchy-Schwarz inequality. Let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$. Then

$$|\mathbf{x}'\mathbf{y}| \leq \|\mathbf{x}\| \cdot \|\mathbf{y}\|$$

Equality holds if and only if $\mathbf{x}$ and $\mathbf{y}$ are linearly dependent.

PROOF. The inequality trivially holds if $\mathbf{x} = 0$ or $\mathbf{y} = 0$. Assume that $\mathbf{y} \neq 0$. Then we find for any $\lambda \in \mathbb{R}$,

$$0 \leq (\mathbf{x} - \lambda\mathbf{y})'(\mathbf{x} - \lambda\mathbf{y}) = \mathbf{x}'\mathbf{x} - \lambda\mathbf{x}'\mathbf{y} - \lambda\mathbf{y}'\mathbf{x} + \lambda^2\mathbf{y}'\mathbf{y} = \mathbf{x}'\mathbf{x} - 2\lambda\mathbf{x}'\mathbf{y} + \lambda^2\mathbf{y}'\mathbf{y}.$$

Using the special value $\lambda = \frac{\mathbf{x}'\mathbf{y}}{\mathbf{y}'\mathbf{y}}$ we obtain

$$0 \leq \mathbf{x}'\mathbf{x} - 2\frac{\mathbf{x}'\mathbf{y}}{\mathbf{y}'\mathbf{y}}\mathbf{x}'\mathbf{y} + \frac{(\mathbf{x}'\mathbf{y})^2}{(\mathbf{y}'\mathbf{y})^2}\mathbf{y}'\mathbf{y} = \mathbf{x}'\mathbf{x} - \frac{(\mathbf{x}'\mathbf{y})^2}{\mathbf{y}'\mathbf{y}}.$$

Hence

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

or equivalently

$$|\mathbf{x}'\mathbf{y}| \leq \|\mathbf{x}\| \, \|\mathbf{y}\|$$

as claimed. The proof for the last statement is left as an exercise, see Problem 8.2. □

Minkowski inequality. Let $\mathbf{x},\mathbf{y} \in \mathbb{R}^n$. Then Theorem 8.7

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

Proof. See Problem 8.3.

Fundamental properties of norms. Let $\mathbf{x},\mathbf{y} \in \mathbb{R}^n$ and $\alpha \in \mathbb{R}$. Then Theorem 8.8

(1) $\|\mathbf{x}\| \geq 0$ where equality holds if and only if $\mathbf{x} = 0$ (Positive-definiteness)

(2) $\|\alpha\mathbf{x}\| = |\alpha|\,\|\mathbf{x}\|$ (Positive scalability)

(3) $\|\mathbf{x}+\mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$ (Triangle inequality or subadditivity)

Proof. See Problem 8.5.

A vector $\mathbf{x} \in \mathbb{R}^n$ is called normalized if $\|\mathbf{x}\| = 1$. Definition 8.9

Normed vector space. The notion of a *norm* can be generalized. Let $\mathcal{V}$ be some vector space. Then any function Definition 8.10

$$\|\cdot\| : \mathcal{V} \to \mathbb{R}$$

that satisfies properties

(i) $\|\mathbf{x}\| \geq 0$ where equality holds if and only if $\mathbf{x} = 0$

(ii) $\|\alpha\mathbf{x}\| = |\alpha|\,\|\mathbf{x}\|$

(iii) $\|\mathbf{x}+\mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$

is called a **norm**. A vector space that is equipped with such a norm is called a **normed vector space**.

The *Euclidean* norm of a vector $\mathbf{x}$ in Definition 8.5 is often denoted by $\|\mathbf{x}\|_2$ and called the *2-norm* of $\mathbf{x}$ (because the coefficients of $\mathbf{x}$ are squared).

Other examples of norms of vectors $\mathbf{x} \in \mathbb{R}^n$ are the so called *1-norm* Example 8.11

$$\|\mathbf{x}\|_1 = \sum_{i=1}^{n} |x_i|$$

the *p-norm*

$$\|\mathbf{x}\|_p = \left(\sum_{i=1}^{n} |x_i|^p\right)^{\frac{1}{p}} \qquad \text{for } p \leq 1 < \infty,$$

and the *supremum norm*

$$\|\mathbf{x}\|_\infty = \max_{i=1,\ldots,n} |x_i| \,. \qquad \diamondsuit$$

Example 8.12

Let $\mathscr{L}$ be the vector space of all random variables X on some given probability space with finite variance $\mathbb{V}(X)$. Then map

$$\|\cdot\|_2 \colon \mathscr{L} \to [0,\infty),\ X \mapsto \|X\|_2 = \sqrt{\mathbb{E}(X^2)} = \sqrt{\langle X,X\rangle}$$

is a norm in $\mathscr{L}$. ♢

In Definition 8.5 we used the inner product (Definition 8.1) to define the Euclidean norm. In fact we only needed the properties of the inner product to derive the properties of the Euclidean norm in Theorem 8.8 and the Cauchy-Schwarz inequality (Theorem 8.6). That is, every inner product induces a norm. However, there are also other norms that are not induced by inner products, e.g., the p-norms $\|\mathbf{x}\|_p$ for $p \neq 2$.

Definition 8.13

Euclidean metric. Let $\mathbf{x},\mathbf{y} \in \mathbb{R}^n$, then $d_2(\mathbf{x},\mathbf{y}) = \|\mathbf{x}-\mathbf{y}\|_2$ defines the **Euclidean distance** between $\mathbf{x}$ and $\mathbf{y}$.

Theorem 8.14

Fundamental properties of metrics. Let $\mathbf{x},\mathbf{y},\mathbf{z} \in \mathbb{R}^n$. Then

(1) $d_2(\mathbf{x},\mathbf{y}) = d_2(\mathbf{y},\mathbf{x})$ (Symmetry)

(2) $d_2(\mathbf{x},\mathbf{y}) \geq 0$ where equality holds if and only if $\mathbf{x} = \mathbf{y}$ (Positive-definiteness)

(3) $d_2(\mathbf{x},\mathbf{z}) \leq d_2(\mathbf{x},\mathbf{y}) + d_2(\mathbf{y},\mathbf{z})$ (Triangle inequality)

PROOF. See Problem 8.8.

Definition 8.15

Metric space. The notion of a *metric* can be generalized. Let $\mathscr{V}$ be some vector space. Then any function

$$d(\cdot,\cdot) \colon \mathscr{V} \times \mathscr{V} \to \mathbb{R}$$

that satisfies properties

(i) $d(\mathbf{x},\mathbf{y}) = d(\mathbf{y},\mathbf{x})$

(ii) $d(\mathbf{x},\mathbf{y}) \geq 0$ where equality holds if and only if $\mathbf{x} = \mathbf{y}$

(iii) $d(\mathbf{x},\mathbf{z}) \leq d(\mathbf{x},\mathbf{y}) + d(\mathbf{y},\mathbf{z})$

is called a **metric**. A vector space that is equipped with a metric is called a **metric vector space**.

Definition 8.13 (and the proof of Theorem 8.14) shows us that any norm induces a metric. However, there also exist metrics that are not induced by some norm.

Example 8.16

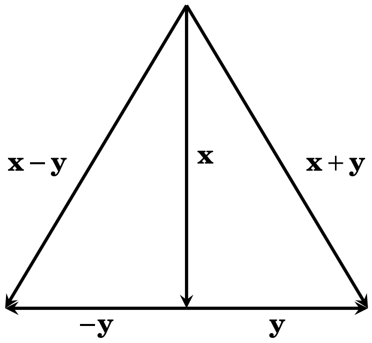
Let $\mathscr{L}$ be the vector space of all random variables X on some given probability space with finite variance $\mathbb{V}(X)$. Then the following maps are metrics in $\mathscr{L}$:

$$d_2 \colon \mathscr{L} \times \mathscr{L} \to [0,\infty),\ (X,Y) \mapsto \|X-Y\| = \sqrt{\mathbb{E}((X-Y)^2)}$$
$$d_E \colon \mathscr{L} \times \mathscr{L} \to [0,\infty),\ (X,Y) \mapsto d_E(X,Y) = \mathbb{E}(|X-Y|)$$
$$d_F \colon \mathscr{L} \times \mathscr{L} \to [0,\infty),\ (X,Y) \mapsto d_F(X,Y) = \max\left|F_X(z) - F_Y(z)\right|$$

where F_X denotes the cumulative distribution function of X. ♢

8.2 Orthogonality

Two vectors $\mathbf{x}$ and $\mathbf{y}$ are *perpendicular* if and only if the triangle shown below is isosceles, i.e., $\|\mathbf{x}+\mathbf{y}\| = \|\mathbf{x}-\mathbf{y}\|$.



The difference between the two sides of this triangle can be computed by means of an inner product (see Problem 8.9) as

$$\|\mathbf{x}+\mathbf{y}\|^2 - \|\mathbf{x}-\mathbf{y}\|^2 = 4\mathbf{x}'\mathbf{y} .$$

Two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ are called **orthogonal** to each other if $\mathbf{x}'\mathbf{y} = 0$. Definition 8.17

Pythagorean theorem. Let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ be two vectors that are orthogonal to each other. Then Theorem 8.18

$$\|\mathbf{x}+\mathbf{y}\|^2 = \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 .$$

PROOF. See Problem 8.10.

Let $\mathbf{v}_1, \dots, \mathbf{v}_k$ be non-zero vectors. If these vectors are pairwise orthogonal to each other, then they are linearly independent. Lemma 8.19

PROOF. Suppose $\mathbf{v}_1, \dots, \mathbf{v}_k$ are linearly dependent. Then w.l.o.g. there exist $\alpha_2, \dots, \alpha_k$ such that $\mathbf{v}_1 = \sum_{i=2}^k \alpha_i \mathbf{v}_i$. Then $\mathbf{v}_1'\mathbf{v}_1 = \mathbf{v}_1' \left(\sum_{i=2}^k \alpha_i \mathbf{v}_i\right) = \sum_{i=2}^k \alpha_i \mathbf{v}_1' \mathbf{v}_i = 0$, i.e., $\mathbf{v}_1 = 0$ by Theorem 8.2, a contradiction to our assumption that all vectors are non-zero. □

Orthonormal system. A set $\{\mathbf{v}_1, \dots, \mathbf{v}_n\} \subset \mathbb{R}^n$ is called an **orthonormal system** if the following holds: Definition 8.20

(i) the vectors are mutually orthogonal,

(ii) the vectors are normalized.

Orthonormal basis. A basis $\{\mathbf{v}_1, \dots, \mathbf{v}_n\} \subset \mathbb{R}^n$ is called an **orthonormal basis** if it forms an orthonormal system. Definition 8.21

Notice that we find for the elements of an orthonormal basis $B = \{\mathbf{v}_1,\dots,\mathbf{v}_n\}$:

$$\mathbf{v}'_i\mathbf{v}_j = \delta_{ij} = \begin{cases} 1, & \text{if } i = j, \\ 0, & \text{if } i \neq j. \end{cases}$$

Theorem 8.22

Let $B = \{\mathbf{v}_1,\dots,\mathbf{v}_n\}$ be an orthonormal basis of $\mathbb{R}^n$. Then the coefficient vector $\mathbf{c}(\mathbf{x})$ of some vector $\mathbf{x} \in \mathbb{R}^n$ with respect to B is given by

$$c_j(\mathbf{x}) = \mathbf{v}'_j\mathbf{x}.$$

PROOF. See Problem 8.11.

Definition 8.23

Orthogonal matrix. A square matrix $\mathbf{U}$ is called an **orthogonal matrix** if its columns form an orthonormal system.

Theorem 8.24

Let $\mathbf{U}$ be an $n \times n$ matrix. Then the following are equivalent:

(1) $\mathbf{U}$ is an orthogonal matrix.

(2) $\mathbf{U}'$ is an orthogonal matrix.

(3) $\mathbf{U}'\mathbf{U} = \mathbf{I}$, i.e., $\mathbf{U}^{-1} = \mathbf{U}'$.

(4) The linear map defined by $\mathbf{U}$ is an **isometry**, i.e., $\|\mathbf{U}\mathbf{x}\| = \|\mathbf{x}\|$ for all $\mathbf{x} \in \mathbb{R}^n$.

PROOF. Let $\mathbf{U} = (\mathbf{u}_1,\dots,\mathbf{u}_n)$.

(1)⇒(3) $[\mathbf{U}'\mathbf{U}]_{ij} = \mathbf{u}'_i\mathbf{u}_j = \delta_{ij} = [\mathbf{I}]_{ij}$, i.e., $\mathbf{U}'\mathbf{U} = \mathbf{I}$. By Theorem 6.16, $\mathbf{U}'\mathbf{U} = \mathbf{U}\mathbf{U}' = \mathbf{I}$ and thus $\mathbf{U}^{-1} = \mathbf{U}'$.

(3)⇒(4) $\|\mathbf{U}\mathbf{x}\|^2 = (\mathbf{U}\mathbf{x})'(\mathbf{U}\mathbf{x}) = \mathbf{x}'\mathbf{U}'\mathbf{U}\mathbf{x} = \mathbf{x}'\mathbf{x} = \|\mathbf{x}\|^2$.

(4)⇒(1) Let $\mathbf{x},\mathbf{y} \in \mathbb{R}^n$. Then by (4), $\|\mathbf{U}(\mathbf{x}-\mathbf{y})\| = \|\mathbf{x}-\mathbf{y}\|$, or equivalently

$$\mathbf{x}'\mathbf{U}'\mathbf{U}\mathbf{x} - \mathbf{x}'\mathbf{U}'\mathbf{U}\mathbf{y} - \mathbf{y}'\mathbf{U}'\mathbf{U}\mathbf{x} + \mathbf{y}'\mathbf{U}'\mathbf{U}\mathbf{y} = \mathbf{x}'\mathbf{x} - \mathbf{x}'\mathbf{y} - \mathbf{y}'\mathbf{x} + \mathbf{y}'\mathbf{y}.$$

If we again apply (4) we can cancel out some terms on both side of this equation and obtain

$$-\mathbf{x}'\mathbf{U}'\mathbf{U}\mathbf{y} - \mathbf{y}'\mathbf{U}'\mathbf{U}\mathbf{x} = -\mathbf{x}'\mathbf{y} - \mathbf{y}'\mathbf{x}.$$

Notice that $\mathbf{x}'\mathbf{y} = \mathbf{x}'\mathbf{y}$ by Theorem 8.2. Similarly, $\mathbf{x}'\mathbf{U}'\mathbf{U}\mathbf{y} = \left(\mathbf{x}'\mathbf{U}'\mathbf{U}\mathbf{y}\right)' = \mathbf{y}'\mathbf{U}'\mathbf{U}\mathbf{x}$, where the first equality holds as these are 1×1 matrices. The second equality follows from the properties of matrix multiplication (Theorem 4.15). Thus

$$\mathbf{x}'\mathbf{U}'\mathbf{U}\mathbf{y} = \mathbf{x}'\mathbf{y} = \mathbf{x}'\mathbf{I}\mathbf{y} \quad \text{for all } \mathbf{x},\mathbf{y} \in \mathbb{R}^n.$$

Recall that $\mathbf{e}'_i\mathbf{U}' = \mathbf{u}'_i$ and $\mathbf{U}\mathbf{e}_j = \mathbf{u}_j$. Thus if we set $\mathbf{x} = \mathbf{e}_i$ and $\mathbf{y} = \mathbf{e}_j$ we obtain

$$\mathbf{u}'_i\mathbf{u}_j = \mathbf{e}'_i\mathbf{U}'\mathbf{U}\mathbf{e}_j = \mathbf{e}'_i\mathbf{e}_j = \delta_{ij}$$

that is, the columns of **U** for an orthonormal system.

(2)⇒(3) Can be shown analogously to (1)⇒(3).

(3)⇒(2) Let $\mathbf{v}_1, \dots, \mathbf{v}_n$ denote the rows of **U**. Then

$$\mathbf{v}_i'\mathbf{v}_j = [\mathbf{U}\mathbf{U}']_{ij} = [\mathbf{I}]_{ij} = \delta_{ij}$$

i.e., the rows of **U** form an orthonormal system.

This completes the proof. □

— Summary

- An *inner product* is a bilinear symmetric positive definite function $\mathcal{V} \times \mathcal{V} \to \mathbb{R}$. It can be seen as a measure for the angle between two vectors.
- Two vectors **x** and **y** are *orthogonal* (perpendicular, normal) to each other, if their inner product is 0.
- A *norm* is a positive definite, positive scalable function $\mathcal{V} \to [0,\infty)$ that satisfies the triangle inequality $\|\mathbf{x}+\mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$. It can be seen as the length of a vector.
- Every inner product induces a norm: $\|x\| = \sqrt{\mathbf{x}'\mathbf{x}}$.

  Then the *Cauchy-Schwarz inequality* $|\mathbf{x}'\mathbf{y}| \leq \|\mathbf{x}\| \cdot \|\mathbf{y}\|$ holds for all $\mathbf{x}, \mathbf{y} \in \mathcal{V}$.

  If in addition $\mathbf{x}, \mathbf{y} \in \mathcal{V}$ are orthogonal, then the *Pythagorean theorem* $\|\mathbf{x}+\mathbf{y}\|^2 = \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2$ holds.
- A *metric* is a bilinear symmetric positive definite function $\mathcal{V} \times \mathcal{V} \to [0,\infty)$ that satisfies the triangle inequality. It measures the distance between two vectors.
- Every norm induces a metric.
- A metric that is induced by an inner product is called an *Euclidean metric*.
- Set of vectors that are mutually orthogonal and have norm 1 is called an *orthonormal system*.
- An *orthogonal matrix* is whose columns form an orthonormal system. Orthogonal maps preserve angles and norms.

— Problems

8.1 Prove Theorem 8.2.

8.2 Complete the proof of Theorem 8.6. That is, show that equality holds if and only if $\mathbf{x}$ and $\mathbf{y}$ are linearly dependent.

8.3 (a) The Minkowski inequality is also called **triangle inequality**. Draw a picture that illustrates this inequality.

(b) Prove Theorem 8.7.

(c) Give conditions where equality holds for the Minkowski inequality.

HINT: Compute $\|\mathbf{x}+\mathbf{y}\|^2$ and apply the Cauchy-Schwarz inequality.

8.4 Show that for any $\mathbf{x},\mathbf{y} \in \mathbb{R}^n$

$$\Big| \|\mathbf{x}\| - \|\mathbf{y}\| \Big| \le \|\mathbf{x}-\mathbf{y}\| \,.$$

HINT: Use the simple observation that $\mathbf{x} = (\mathbf{x}-\mathbf{y})+\mathbf{y}$ and $\mathbf{y} = (\mathbf{y}-\mathbf{x})+\mathbf{y}$ and apply the Minkowski inequality.

8.5 Prove Theorem 8.8. Draw a picture that illustrates property (iii).

HINT: Use Theorems 8.2 and 8.7.

8.6 Let $\mathbf{x} \in \mathbb{R}^n$ be a non-zero vector. Show that $\frac{\mathbf{x}}{\|\mathbf{x}\|}$ is a normalized vector. Is the condition $\mathbf{x} \neq 0$ necessary? Why? Why not?

8.7 (a) Show that $\|\mathbf{x}\|_1$ and $\|\mathbf{x}\|_\infty$ satisfy the properties of a norm.

(b) Draw the unit balls in $\mathbb{R}^2$, i.e., the sets $\{\mathbf{x} \in \mathbb{R}^2 \colon \|\mathbf{x}\| \le 1\}$ with respect to the norms $\|\mathbf{x}\|_1$, $\|\mathbf{x}\|_2$, and $\|\mathbf{x}\|_\infty$.

(c) Use a computer algebra system of your choice (e.g., *Maxima*) and draw unit balls with respect to the p-norm for various values of p. What do you observe?

8.8 Prove Theorem 8.14. Draw a picture that illustrates property (iii).

HINT: Use Theorem 8.8.

8.9 Show that $\|\mathbf{x}+\mathbf{y}\|^2 - \|\mathbf{x}-\mathbf{y}\|^2 = 4\mathbf{x}'\mathbf{y}$.

HINT: Use $\|\mathbf{x}\|^2 = \mathbf{x}'\mathbf{x}$.

8.10 Prove Theorem 8.18.

HINT: Use $\|\mathbf{x}\|^2 = \mathbf{x}'\mathbf{x}$.

8.11 Prove Theorem 8.22.

HINT: Represent $\mathbf{x}$ by means of $\mathbf{c}(\mathbf{x})$ and compute $\mathbf{x}'\mathbf{v}_j$.

8.12 Let $\mathbf{U} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$. Give conditions for the elements a,b,c,d that imply that $\mathbf{U}$ is an orthogonal matrix. Give an example for such an orthogonal matrix.

9

Projections

To them, I said, the truth would be literally nothing but the shadows *of the images.*

Suppose we are given a subspace $\mathcal{U} \subset \mathbb{R}^n$ and a vector $\mathbf{y} \in \mathbb{R}^n$. We want to find a vector $\mathbf{u} \in \mathcal{U}$ such that the "error" $\mathbf{r} = \mathbf{y} - \mathbf{u}$ is as small as possible. This procedure is of great importance when we want to reduce the number of dimensions in our model without loosing too much information.

9.1 Orthogonal Projection

We first look at the simplest case $\mathcal{U} = \operatorname{span}(\mathbf{x})$ where $\mathbf{x} \in \mathbb{R}^n$ is some fixed normalized vector, i.e., $\|\mathbf{x}\| = 1$. Then every $\mathbf{u} \in \mathcal{U}$ can be written as $\lambda\mathbf{x}$ for some $\lambda \in \mathbb{R}$.

Lemma 9.1

Let $\mathbf{y}, \mathbf{x} \in \mathbb{R}^n$ be fixed with $\|\mathbf{x}\| = 1$. Let $\mathbf{r} \in \mathbb{R}^n$ and $\lambda \in \mathbb{R}$ such that

$$\mathbf{y} = \lambda\mathbf{x} + \mathbf{r}\,.$$

Then for $\lambda = \lambda^*$ and $\mathbf{r} = \mathbf{r}^*$ the following statements are equivalent:

(1) $\|\mathbf{r}^*\|$ is minimal among all values for λ and $\mathbf{r}$.

(2) $\mathbf{x}'\mathbf{r}^* = 0$.

(3) $\lambda^* = \mathbf{x}'\mathbf{y}$.

PROOF. (2) ⇔ (3): Follows by a simple computation (see Problem 9.1). An immediate consequence is that there always exist $\mathbf{r}^*$ and λ^* such that $\mathbf{r}^* = \mathbf{y} - \lambda^*\mathbf{x}$ and $\mathbf{x}'\mathbf{r}^* = 0$.
(2) ⇒ (1): Assume that $\mathbf{x}'\mathbf{r}^* = 0$ and λ^* such that $\mathbf{r}^* = \mathbf{y} - \lambda^*\mathbf{x}$. Set $\mathbf{r}(\varepsilon) = \mathbf{y} - (\lambda^* + \varepsilon)\mathbf{x} = (\mathbf{y} - \lambda^*\mathbf{x}) - \varepsilon\mathbf{x} = \mathbf{r}^* - \varepsilon\mathbf{x}$ for $\varepsilon \in \mathbb{R}$. As $\mathbf{r}^*$ and $\mathbf{x}$ are orthogonal by our assumption the Pythagorean theorem implies $\|\mathbf{r}(\varepsilon)\|^2 = \|\mathbf{r}^*\|^2 + \varepsilon^2\|\mathbf{x}\|^2 = \|\mathbf{r}^*\|^2 + \varepsilon^2$. Thus $\|\mathbf{r}(\varepsilon)\| \geq \|\mathbf{r}^*\|$ where equality holds if and only if $\varepsilon = 0$. Thus $\mathbf{r}^*$ minimizes $\|\mathbf{r}\|$.

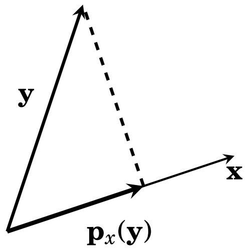
(1) ⇒ (2): Assume that $\mathbf{r}^*$ minimizes $\|\mathbf{r}\|$ and λ^* such that $\mathbf{r}^* = \mathbf{y} - \lambda^*\mathbf{x}$. Set $\mathbf{r}(\varepsilon) = \mathbf{y} - (\lambda^* + \varepsilon)\mathbf{x} = \mathbf{r}^* - \varepsilon\mathbf{x}$ for $\varepsilon \in \mathbb{R}$. Our assumption implies that $\|\mathbf{r}^*\|^2 \le \|\mathbf{r}^* - \varepsilon\mathbf{x}\|^2 = \|\mathbf{r}^*\|^2 - 2\varepsilon\mathbf{x}'\mathbf{r} + \varepsilon^2\|\mathbf{x}\|^2$ for all $\varepsilon \in \mathbb{R}$. Thus $2\varepsilon\mathbf{x}'\mathbf{r} \le \varepsilon^2\|\mathbf{x}\|^2 = \varepsilon^2$. Since ε may have positive and negative sign we find $-\frac{\varepsilon}{2} \le \mathbf{x}'\mathbf{r} \le \frac{\varepsilon}{2}$ for all $\varepsilon \ge 0$ and hence $\mathbf{x}'\mathbf{r} = 0$, as claimed. □

Definition 9.2

Orthogonal projection. Let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ be two vectors with $\|\mathbf{x}\| = 1$. Then

$$\mathbf{p}_x(\mathbf{y}) = (\mathbf{x}'\mathbf{y})\mathbf{x}$$

is called the **orthogonal projection** of $\mathbf{y}$ onto the linear span of $\mathbf{x}$.



Theorem 9.3

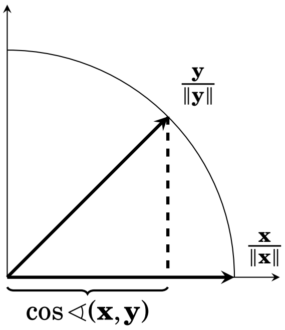
Orthogonal decomposition. Let $\mathbf{x} \in \mathbb{R}^n$ with $\|\mathbf{x}\| = 1$. Then every $\mathbf{y} \in \mathbb{R}^n$ can be uniquely decomposed as

$$\mathbf{y} = \mathbf{u} + \mathbf{v}$$

where $\mathbf{u} \in \operatorname{span}(\mathbf{x})$ and $\mathbf{v}$ is orthogonal to $\mathbf{u}$, that is $\mathbf{u}'\mathbf{v} = 0$. Such a representation is called an **orthogonal decomposition** of $\mathbf{y}$. Moreover, $\mathbf{u}$ is given by

$$\mathbf{u} = \mathbf{p}_x(\mathbf{y})\,.$$

PROOF. Let $\mathbf{u} = \lambda\mathbf{x} \in \operatorname{span}(\mathbf{x})$ with $\lambda = \mathbf{x}'\mathbf{y}$ and $\mathbf{v} = \mathbf{y} - \mathbf{u}$. Obviously, $\mathbf{u} + \mathbf{v} = \mathbf{y}$. By Lemma 9.1, $\mathbf{u}'\mathbf{v} = 0$ and $\mathbf{u} = \mathbf{p}_x(\mathbf{y})$. Moreover, no other value of λ has this property. □



$\frac{\mathbf{y}}{\|\mathbf{y}\|}$
$\frac{\mathbf{x}}{\|\mathbf{x}\|}$
$\cos\sphericalangle(\mathbf{x},\mathbf{y})$

Now let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ with $\|\mathbf{x}\| = \|\mathbf{y}\| = 1$ and let $\lambda = \mathbf{x}'\mathbf{y}$. Then $|\lambda| = \|\mathbf{p}_x(\mathbf{y})\|$ and λ is positive if $\mathbf{x}$ and $\mathbf{p}_x(\mathbf{y})$ have the same orientation and negative if $\mathbf{x}$ and $\mathbf{p}_x(\mathbf{y})$ have opposite orientation. Thus by a geometric argument, λ is just the cosine of the angle between these vectors, i.e., $\cos\sphericalangle(\mathbf{x},\mathbf{y}) = \mathbf{x}'\mathbf{y}$. If $\mathbf{x}$ and $\mathbf{y}$ are arbitrary non-zero vectors these have to be normalized. We then find

$$\cos\sphericalangle(\mathbf{x},\mathbf{y}) = \frac{\mathbf{x}'\mathbf{y}}{\|\mathbf{x}\|\,\|\mathbf{y}\|}\,.$$

Projection matrix. Let $\mathbf{x} \in \mathbb{R}^n$ be fixed with $\|\mathbf{x}\| = 1$. Then $\mathbf{y} \mapsto \mathbf{p}_x(\mathbf{y})$ is a linear map and $\mathbf{p}_x(\mathbf{y}) = \mathbf{P}_x\mathbf{y}$ where $\mathbf{P}_x = \mathbf{x}\mathbf{x}'$. Theorem 9.4

PROOF. Let $\mathbf{y}_1, \mathbf{y}_2 \in \mathbb{R}^n$ and $\alpha_1, \alpha_2 \in \mathbb{R}$. Then

$$\begin{aligned}\mathbf{p}_x(\alpha_1\mathbf{y}_1 + \alpha_2\mathbf{y}_2) &= \big(\mathbf{x}'(\alpha_1\mathbf{y}_1 + \alpha_2\mathbf{y}_2)\big)\mathbf{x} = \big(\alpha_1\mathbf{x}'\mathbf{y}_1 + \alpha_2\mathbf{x}'\mathbf{y}_2\big)\mathbf{x} \\ &= \alpha_1(\mathbf{x}'\mathbf{y}_1)\mathbf{x} + \alpha_2(\mathbf{x}'\mathbf{y}_2)\mathbf{x} = \alpha_1\mathbf{p}_x(\mathbf{y}_1) + \alpha_2\mathbf{p}_x(\mathbf{y}_2)\end{aligned}$$

and thus $\mathbf{p}_x$ is a linear map and there exists a matrix $\mathbf{P}_x$ such that $\mathbf{p}_x(\mathbf{y}) = \mathbf{P}_x\mathbf{y}$ by Theorem 6.15.
Notice that $\alpha\mathbf{x} = \mathbf{x}\alpha$ for $\alpha \in \mathbb{R} = \mathbb{R}^1$. Thus $\mathbf{P}_x\mathbf{y} = (\mathbf{x}'\mathbf{y})\mathbf{x} = \mathbf{x}(\mathbf{x}'\mathbf{y}) = (\mathbf{x}\mathbf{x}')\mathbf{y}$ for all $\mathbf{y} \in \mathbb{R}^n$ and the result follows. □

If we project some vector $\mathbf{y} \in \operatorname{span}(\mathbf{x})$ onto $\operatorname{span}(\mathbf{x})$ then $\mathbf{y}$ remains unchanged, i.e., $\mathbf{P}_x(\mathbf{y}) = \mathbf{y}$. Thus the projection matrix $\mathbf{P}_x$ has the property that $\mathbf{P}_x^2\mathbf{z} = \mathbf{P}_x(\mathbf{P}_x\mathbf{z}) = \mathbf{P}_x\mathbf{z}$ for every $\mathbf{z} \in \mathbb{R}^n$ (see Problem 9.3).

A square matrix $\mathbf{A}$ is called **idempotent** if $\mathbf{A}^2 = \mathbf{A}$. Definition 9.5

9.2 Gram-Schmidt Orthonormalization

Theorem 8.22 shows that we can easily compute the coefficient vector $\mathbf{c}(\mathbf{x})$ of a vector $\mathbf{x}$ by means of projections when the given basis $\{\mathbf{v}_1, \dots, \mathbf{v}_n\}$ forms an orthonormal system:

$$\mathbf{x} = \sum_{i=1}^{n} c_i(\mathbf{x})\mathbf{v}_i = \sum_{i=1}^{n}(\mathbf{v}_i'\mathbf{x})\mathbf{v}_i = \sum_{i=1}^{n}\mathbf{p}_{v_i}(\mathbf{x})\,.$$

Hence orthonormal bases are quite convenient. Theorem 9.3 allows us to transform any two linearly independent vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ into two orthogonal vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$ which then can be normalized. This idea can be generalized to any number of linear independent vectors by means of a recursion, called **Gram-Schmidt process**.

Gram-Schmidt orthonormalization. Let $\{\mathbf{u}_1, \dots, \mathbf{u}_n\}$ be a basis of some subspace $\mathscr{U}$. Define $\mathbf{v}_k$ recursively for $k = 1, \dots, n$ by Theorem 9.6

$$\begin{aligned}
\mathbf{w}_1 &= \mathbf{u}_1, & \mathbf{v}_1 &= \frac{\mathbf{w}_1}{\|\mathbf{w}_1\|} \\
\mathbf{w}_2 &= \mathbf{u}_2 - \mathbf{p}_{v_1}(\mathbf{u}_2), & \mathbf{v}_2 &= \frac{\mathbf{w}_2}{\|\mathbf{w}_2\|} \\
\mathbf{w}_3 &= \mathbf{u}_3 - \mathbf{p}_{v_1}(\mathbf{u}_3) - \mathbf{p}_{v_2}(\mathbf{u}_3), & \mathbf{v}_3 &= \frac{\mathbf{w}_3}{\|\mathbf{w}_3\|} \\
&\vdots & &\vdots \\
\mathbf{w}_n &= \mathbf{u}_n - \sum_{j=1}^{n-1}\mathbf{p}_{v_j}(\mathbf{u}_n), & \mathbf{v}_n &= \frac{\mathbf{w}_n}{\|\mathbf{w}_n\|}
\end{aligned}$$

where $\mathbf{p}_{v_j}$ is the orthogonal projection from Definition 9.2, that is, $\mathbf{p}_{v_j}(\mathbf{u_k}) = (\mathbf{v}_j'\mathbf{u}_k)\mathbf{v}_j$. Then set $\{\mathbf{v}_1, \dots, \mathbf{v}_n\}$ forms an orthonormal basis for $\mathscr{U}$.

PROOF. We proceed by induction on k and show that $\{\mathbf{v}_1,\dots,\mathbf{v}_k\}$ form an orthonormal basis for $\operatorname{span}(\mathbf{u}_1,\dots,\mathbf{u}_k)$ for all $k=1,\dots,n$.
For $k=1$ the statement is obvious as $\operatorname{span}(\mathbf{v}_1)=\operatorname{span}(\mathbf{u}_1)$ and $\|\mathbf{v}_1\|=1$. Now suppose the result holds for $k\geq 1$. By the induction hypothesis, $\{\mathbf{v}_1,\dots,\mathbf{v}_k\}$ forms an orthonormal basis for $\operatorname{span}(\mathbf{u}_1,\dots,\mathbf{u}_k)$. In particular we have $\mathbf{v}_j'\mathbf{v}_i=\delta_{ji}$. Let

$$\mathbf{w}_{k+1}=\mathbf{u}_{k+1}-\sum_{j=1}^{k}\mathbf{p}_{v_j}(\mathbf{u}_{k+1})=\mathbf{u}_{k+1}-\sum_{j=1}^{k}(\mathbf{v}_j'\mathbf{u}_{k+1})\mathbf{v}_j\,.$$

First we show that $\mathbf{w}_{k+1}$ and $\mathbf{v}_i$ are orthogonal for all $i=1,\dots,k$. By construction we have

$$\begin{aligned}
\mathbf{w}_{k+1}'\mathbf{v}_i &= \left(\mathbf{u}_{k+1}-\sum_{j=1}^{k}(\mathbf{v}_j'\mathbf{u}_{k+1})\mathbf{v}_j\right)'\mathbf{v}_i\\
&= \mathbf{u}_{k+1}'\mathbf{v}_i-\sum_{j=1}^{k}(\mathbf{v}_j'\mathbf{u}_{k+1})\mathbf{v}_j'\mathbf{v}_i\\
&= \mathbf{u}_{k+1}'\mathbf{v}_i-\sum_{j=1}^{k}(\mathbf{v}_j'\mathbf{u}_{k+1})\delta_{ji}\\
&= \mathbf{u}_{k+1}'\mathbf{v}_i-\mathbf{v}_i'\mathbf{u}_{k+1}\\
&= 0\,.
\end{aligned}$$

Now $\mathbf{w}_{k+1}$ cannot be 0 since otherwise $\mathbf{u}_{k+1}-\sum_{j=1}^{k}\mathbf{p}_{v_j}(\mathbf{u}_{k+1})=0$ and consequently $\mathbf{u}_{k+1}\in\operatorname{span}(\mathbf{v}_1,\dots,\mathbf{v}_k)=\operatorname{span}(\mathbf{u}_1,\dots,\mathbf{u}_k)$, a contradiction to our assumption that $\{\mathbf{u}_1,\dots,\mathbf{u}_k,\mathbf{u}_{k+1}\}$ is a subset of a basis of $\mathcal{U}$. Thus we may take $\mathbf{v}_{k+1}=\frac{\mathbf{w}_{k+1}}{\|\mathbf{w}_{k+1}\|}$. Then by Lemma 8.19 the vectors $\{\mathbf{v}_1,\dots,\mathbf{v}_{k+1}\}$ are linearly independent and consequently form a basis for $\operatorname{span}(\mathbf{u}_1,\dots,\mathbf{u}_{k+1})$ by Theorem 5.21. Thus the result holds for $k+1$, and by the principle of induction, for all $k=1,\dots,n$ and in particular for $k=n$. □

9.3 Orthogonal Complement

We want to generalize Theorem 9.3 and Lemma 9.1. Thus we need the concepts of the *direct sum* of two vector spaces and of the *orthogonal complement*.

Definition 9.7

Direct sum. Let $\mathcal{U},\mathcal{V}\subseteq\mathbb{R}^n$ be two subspaces with $\mathcal{U}\cap\mathcal{V}=\{0\}$. Then

$$\mathcal{U}\oplus\mathcal{V}=\{\mathbf{u}+\mathbf{v}:\mathbf{u}\in\mathcal{U},\mathbf{v}\in\mathcal{V}\}$$

is called the **direct sum** of $\mathcal{U}$ and $\mathcal{V}$.

Lemma 9.8

Let $\mathcal{U},\mathcal{V}\subseteq\mathbb{R}^n$ be two subspaces with $\mathcal{U}\cap\mathcal{V}=\{0\}$ and $\dim(\mathcal{U})=k\geq 1$ and $\dim(\mathcal{V})=l\geq 1$. Let $\{\mathbf{u}_1,\dots,\mathbf{u}_k\}$ and $\{\mathbf{v}_1,\dots,\mathbf{v}_l\}$ be bases of $\mathcal{U}$ and $\mathcal{V}$, respectively. Then $\{\mathbf{u}_1,\dots,\mathbf{u}_k\}\cup\{\mathbf{v}_1,\dots,\mathbf{v}_l\}$ is a basis of $\mathcal{U}\oplus\mathcal{V}$.

Proof. Obviously $\{\mathbf{u}_1,\dots,\mathbf{u}_k\} \cup \{\mathbf{v}_1,\dots,\mathbf{v}_l\}$ is a generating set of $\mathcal{U} \oplus \mathcal{V}$. We have to show that this set is linearly independent. Suppose it is linearly dependent. Then we find $\alpha_1,\dots,\alpha_k \in \mathbb{R}$ not all zero and $\beta_1,\dots,\beta_l \in \mathbb{R}$ not all zero such that $\sum_{i=1}^k \alpha_i \mathbf{u}_i + \sum_{i=1}^l \beta_i \mathbf{v}_i = 0$. Then $\mathbf{u} = \sum_{i=1}^k \alpha_i \mathbf{u}_i \neq 0$ and $\mathbf{v} = -\sum_{i=1}^l \beta_i \mathbf{v}_i \neq 0$ where $\mathbf{u} \in \mathcal{U}$ and $\mathbf{v} \in \mathcal{V}$ and $\mathbf{u} - \mathbf{v} = 0$. But then $\mathbf{u} = \mathbf{v}$, a contradiction to the assumption that $\mathcal{U} \cap \mathcal{V} = \{0\}$. □

Decomposition of a vector. Let $\mathcal{U}, \mathcal{V} \subseteq \mathbb{R}^n$ be two subspaces with $\mathcal{U} \cap \mathcal{V} = \{0\}$ and $\mathcal{U} \oplus \mathcal{V} = \mathbb{R}^n$. Then every $\mathbf{x} \in \mathbb{R}^n$ can be uniquely decomposed into Lemma 9.9

$$\mathbf{x} = \mathbf{u} + \mathbf{v}$$

where $\mathbf{u} \in \mathcal{U}$ and $\mathbf{v} \in \mathcal{V}$.

Proof. See Problem 9.6.

Orthogonal complement. Let $\mathcal{U}$ be a subspace of $\mathbb{R}^n$. Then the **orthogonal complement** of $\mathcal{U}$ in $\mathbb{R}^n$ is the set of vectors $\mathbf{v}$ that are orthogonal to all vectors in $\mathbb{R}^n$, that is, Definition 9.10

$$\mathcal{U}^\perp = \{\mathbf{v} \in \mathbb{R}^n : \mathbf{u}'\mathbf{v} = 0 \text{ for all } \mathbf{u} \in \mathcal{U}\}.$$

Let $\mathcal{U}$ be a subspace of $\mathbb{R}^n$. Then the orthogonal complement $\mathcal{U}^\perp$ is also a subspace of $\mathbb{R}^n$. Furthermore, $\mathcal{U} \cap \mathcal{U}^\perp = \{0\}$. Lemma 9.11

Proof. See Problem 9.7.

Let $\mathcal{U}$ be a subspace of $\mathbb{R}^n$. Then Lemma 9.12

$$\mathbb{R}^n = \mathcal{U} \oplus \mathcal{U}^\perp .$$

Proof. See Problem 9.8.

Orthogonal decomposition. Let $\mathcal{U}$ be a subspace of $\mathbb{R}^n$. Then every $\mathbf{y} \in \mathbb{R}^n$ can be uniquely decomposed into Theorem 9.13

$$\mathbf{y} = \mathbf{u} + \mathbf{u}^\perp$$

where $\mathbf{u} \in \mathcal{U}$ and $\mathbf{u}^\perp \in \mathcal{U}^\perp$. $\mathbf{u}$ is called the **orthogonal projection** of $\mathbf{y}$ into $\mathcal{U}$. We denote this projection by $\mathbf{p}_U(\mathbf{y})$.

Proof. See Problem 9.9.

It remains derive a formula for computing this orthogonal projection. Thus we derive a generalization of and Lemma 9.1.

Theorem 9.14

Projection into subspace. Let $\mathcal{U}$ be a subspace of $\mathbb{R}^n$ with generating set $\{\mathbf{u}_1,\ldots,\mathbf{u}_k\}$ and $\mathbf{U}=(\mathbf{u}_1,\ldots,\mathbf{u}_k)$. For a fixed vector $\mathbf{y}\in\mathbb{R}^n$ let $\mathbf{r}\in\mathbb{R}^n$ and $\boldsymbol{\lambda}\in\mathbb{R}^k$ such that

$$\mathbf{y}=\mathbf{U}\boldsymbol{\lambda}+\mathbf{r}\,.$$

Then for $\boldsymbol{\lambda}=\boldsymbol{\lambda}^*$ and $\mathbf{r}=\mathbf{r}^*$ the following statements are equivalent:

(1) $\|\mathbf{r}^*\|$ is minimal among all possible values for $\boldsymbol{\lambda}$ and $\mathbf{r}$.

(2) $\mathbf{U}'\mathbf{r}^*=0$, that is, $\mathbf{r}^*\in\mathcal{U}^\perp$.

(3) $\mathbf{U}'\mathbf{U}\boldsymbol{\lambda}^*=\mathbf{U}'\mathbf{y}$.

Notice that $\mathbf{u}^*=\mathbf{U}\boldsymbol{\lambda}^*\in\mathcal{U}$.

PROOF. Equivalence of (2) and (3) follows by a straightforward computation (see Problem 9.10).
Now by Theorem 9.13, $\mathbf{y}=\mathbf{u}^*+\mathbf{r}^*$ where $\mathbf{u}^*=\mathbf{U}\boldsymbol{\lambda}^*\in\mathcal{U}$ and $\mathbf{r}^*\in\mathcal{U}^\perp$. Then for every $\boldsymbol{\varepsilon}\in\mathbb{R}^k$ define $\mathbf{r}(\boldsymbol{\varepsilon})=\mathbf{y}-\mathbf{U}(\boldsymbol{\lambda}^*+\boldsymbol{\varepsilon})=\mathbf{r}^*-\mathbf{U}\boldsymbol{\varepsilon}$. As $\mathbf{U}\boldsymbol{\varepsilon}\in\mathcal{U}$, $\mathbf{r}(\boldsymbol{\varepsilon})\in\mathcal{U}^\perp$ if and only if $\mathbf{U}\boldsymbol{\varepsilon}=0$, i.e., $\boldsymbol{\varepsilon}\in\ker(\mathbf{U})$. Now the Pythagorean Theorem implies $\|\mathbf{r}(\boldsymbol{\varepsilon})\|^2=\|\mathbf{r}^*\|^2+\|\mathbf{U}\boldsymbol{\varepsilon}\|^2\geq\|\mathbf{r}^*\|^2$ where equality holds if and only if $\boldsymbol{\varepsilon}\in\ker(\mathbf{U})$. Thus equivalence of (1) and (2) follows. □

Equation (3) in Theorem 9.14 can be transformed simplified when $\{\mathbf{u}_1,\ldots,\mathbf{u}_k\}$ is linearly independent, i.e., when it forms a basis for $\mathcal{U}$. Then the $n\times k$ matrix $\mathbf{U}=(\mathbf{u}_1,\ldots,\mathbf{u}_k)$ has rank k. Then by Lemma 6.25 the $k\times k$ matrix $\mathbf{U}'\mathbf{U}$ also has rank k is thus invertible.

Theorem 9.15

Let $\mathcal{U}$ be a subspace of $\mathbb{R}^n$ with basis $\{\mathbf{u}_1,\ldots,\mathbf{u}_k\}$ and $\mathbf{U}=(\mathbf{u}_1,\ldots,\mathbf{u}_k)$. Then the orthogonal projection $\mathbf{y}\in\mathbb{R}^n$ onto $\mathcal{U}$ is given by

$$\mathbf{p}_U(\mathbf{y})=\mathbf{U}(\mathbf{U}'\mathbf{U})^{-1}\mathbf{U}'\mathbf{y}\,.$$

If in addition $\{\mathbf{u}_1,\ldots,\mathbf{u}_k\}$ forms an orthonormal system we find

$$\mathbf{p}_U(\mathbf{y})=\mathbf{U}\mathbf{U}'\mathbf{y}\,.$$

PROOF. See Problem 9.12.

9.4 Approximate Solutions of Linear Equations

Let $\mathbf{A}$ be an $n\times k$ matrix and $\mathbf{b}\in\mathbb{R}^n$. Suppose there is no $\mathbf{x}\in\mathbb{R}^k$ such that

$$\mathbf{A}\mathbf{x}=\mathbf{b}$$

that is, the linear equation $\mathbf{A}\mathbf{x}=\mathbf{b}$ does not have a solution. Nevertheless, we may want to find an *approximate* solution $\mathbf{x}_0\in\mathbb{R}^k$ that minimizes the error $\mathbf{r}=\mathbf{b}-\mathbf{A}\mathbf{x}$ among all $\mathbf{x}\in\mathbb{R}^k$.

By Theorem 9.14 this task can be solved by means of orthogonal projections $\mathbf{p}_A(\mathbf{b})$ onto the linear span $\mathscr{A}$ of the column vectors of $\mathbf{A}$, i.e., we have to find $\mathbf{x}_0$ such that

$$\mathbf{A}'\mathbf{A}\mathbf{x}_0 = \mathbf{A}'\mathbf{b}\,. \tag{9.1}$$

Notice that by Theorem 9.13 there always exists an $\mathbf{r}$ such that $\mathbf{b} = \mathbf{p}_A(\mathbf{b}) + \mathbf{r}$ with $\mathbf{r} \in \mathscr{A}^\perp$ and hence an $\mathbf{x}_0$ exists such that $\mathbf{p}_A(\mathbf{b}) = \mathbf{A}\mathbf{x}_0$. Thus Equation (9.1) always has a solution by Theorem 9.14.

9.5 Applications in Statistics

Let $\mathbf{x} = (x_1, \dots, x_n)'$ be a given set of data and let $\mathbf{j} = (1, \dots, 1)'$ denote a vector of length n of ones. Notice that $\|\mathbf{j}\|^2 = n$. Then we can express the arithmetic mean $\bar{x}$ of the x_i as

$$\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i = \frac{1}{n}\mathbf{j}'\mathbf{x}$$

and we find

$$\mathbf{p}_j(\mathbf{x}) = \left(\frac{1}{\sqrt{n}}\mathbf{j}'\mathbf{x}\right)\left(\frac{1}{\sqrt{n}}\mathbf{j}\right) = \left(\frac{1}{n}\mathbf{j}'\mathbf{x}\right)\mathbf{j} = \bar{x}\mathbf{j}\,.$$

That is, the arithmetic mean $\bar{x}$ is $\frac{1}{\sqrt{n}}$ times the length of the orthogonal projection of $\mathbf{x}$ onto the constant vector. For the length of its orthogonal complement $\mathbf{p}_j(\mathbf{x})^\perp$ we then obtain

$$\|\mathbf{x} - \bar{x}\mathbf{j}\|^2 = (\mathbf{x} - \bar{x}\mathbf{j})'(\mathbf{x} - \bar{x}\mathbf{j}) = \|\mathbf{x}\|^2 - \bar{x}\mathbf{j}'\mathbf{x} - \bar{x}\mathbf{x}'\mathbf{j} + \bar{x}^2\mathbf{j}'\mathbf{j} = \|\mathbf{x}\|^2 - n\bar{x}^2$$

where the last equality follows from the fact that $\mathbf{j}'\mathbf{x} = \mathbf{x}'\mathbf{j} = \bar{x}n$ and $\mathbf{j}'\mathbf{j} = n$. On the other hand recall that $\|\mathbf{x} - \bar{x}\mathbf{j}\|^2 = \sum_{i=1}^n (x_i - \bar{x})^2 = n\sigma_x^2$ where σ_x^2 denotes the variance of data x_i. Consequently the standard deviation of the data is $\frac{1}{\sqrt{n}}$ times the length of the orthogonal complement of $\mathbf{x}$ with respect to the constant vector.

Now assume that we are also given data $\mathbf{y} = (y_1, \dots, y_n)'$. Again $\mathbf{y} - \bar{y}\mathbf{j}$ is the complement of the orthogonal projection of $\mathbf{y}$ onto the constant vector. Then the inner product of these two orthogonal complements is

$$(\mathbf{x} - \bar{x}\mathbf{j})'(\mathbf{y} - \bar{y}\mathbf{j}) = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = n\sigma_{xy}$$

where σ_{xy} denotes the covariance between $\mathbf{x}$ and $\mathbf{y}$.

Now suppose that we are given a set of data $(y_i, x_{i1}, \dots, x_{ik})$, $i = 1, \dots, n$. We assume a linear regression model, i.e.,

$$y_i = \beta_0 + \sum_{s=1}^k \beta_s x_{is} + \epsilon_i\,.$$

These n equations can be stacked together using matrix notation as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where

$$\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_k \end{pmatrix}, \quad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}.$$

$\mathbf{X}$ is then called the **design matrix** of the linear regression model, $\boldsymbol{\beta}$ are the **model parameters** and $\boldsymbol{\epsilon}$ are random errors (“noise”) called **residuals**.

The parameters $\boldsymbol{\beta}$ can be estimated by means of the **least square principle** where the sum of squared errors,

$$\sum_{i=1}^{n} \epsilon_i^2 = \|\boldsymbol{\epsilon}\|^2 = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2$$

is minimized. Therefore by Theorem 9.14 the estimated parameter $\hat{\boldsymbol{\beta}}$ satisfies the normal equation

$$\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y} \tag{9.2}$$

and hence

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.$$

— Summary

- For every subspace $\mathcal{U} \subset \mathbb{R}^n$ we find $\mathbb{R}^n = \mathcal{U} \oplus \mathcal{U}^\perp$, where $\mathcal{U}^\perp$ denotes the orthogonal complement of $\mathcal{U}$.

- Every $\mathbf{y} \in \mathbb{R}^n$ can be decomposed as $\mathbf{y} = \mathbf{u} + \mathbf{u}^\perp$ where $\mathbf{u} \in \mathcal{U}$ and $\mathbf{u}^\perp \in \mathcal{U}^\perp$. $\mathbf{u}$ is called the *orthogonal projection* of $\mathbf{y}$ into $\mathcal{U}$.

- If $\{\mathbf{u}_1, \dots, \mathbf{u}_k\}$ is a basis of $\mathcal{U}$ and $\mathbf{U} = (\mathbf{u}_1, \dots, \mathbf{u}_k)$, then $\mathbf{U}\mathbf{u}^\perp = 0$ and $\mathbf{u} = \mathbf{U}\boldsymbol{\lambda}$ where $\boldsymbol{\lambda} \in \mathbb{R}^k$ satisfies $\mathbf{U}'\mathbf{U}\boldsymbol{\lambda} = \mathbf{U}'\mathbf{y}$.

- If $\mathbf{y} = \mathbf{u} + \mathbf{v}$ with $\mathbf{u} \in \mathcal{U}$ then $\mathbf{v}$ has minimal length for fixed $\mathbf{y}$ if and only if $\mathbf{v} \in \mathcal{U}^\perp$.

- If the linear equation $\mathbf{A}\mathbf{x} = \mathbf{b}$ does not have a solution, then the solution of $\mathbf{A}'\mathbf{A}\mathbf{x} = \mathbf{A}'\mathbf{b}$ minimizes the error $\|\mathbf{b} - \mathbf{A}\mathbf{x}\|$.

— Problems

9.1 Assume that $\mathbf{y} = \lambda\mathbf{x} + \mathbf{r}$ where $\|\mathbf{x}\| = 1$.
Show: $\mathbf{x}'\mathbf{r} = 0$ if and only if $\lambda = \mathbf{x}'\mathbf{y}$.

HINT: Use $\mathbf{x}'(\mathbf{y} - \lambda\mathbf{x}) = 0$.

9.2 Let $\mathbf{r} = \mathbf{y} - \lambda\mathbf{x}$ where $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ and $\mathbf{x} \neq 0$. Which values of $\lambda \in \mathbb{R}$ minimize $\|\mathbf{r}\|$?

HINT: Use the normalized vector $\mathbf{x}_0 = \mathbf{x}/\|\mathbf{x}\|$ and apply Lemma 9.1.

9.3 Let $\mathbf{P}_x = \mathbf{x}\mathbf{x}'$ for some $\mathbf{x} \in \mathbb{R}^n$ with $\|\mathbf{x}\| = 1$.

(a) What is the dimension of $\mathbf{P}_x$?
(b) What is the rank of $\mathbf{P}_x$?
(c) Show that $\mathbf{P}_x$ is symmetric.
(d) Show that $\mathbf{P}_x$ is idempotent.
(e) Describe the rows and columns of $\mathbf{P}_x$.

9.4 Show that the direct sum $\mathscr{U} \oplus \mathscr{V}$ of two subspaces $\mathscr{U}, \mathscr{V} \subseteq \mathbb{R}^n$ is a subspace of $\mathbb{R}^n$.

9.5 Prove or disprove: Let $\mathscr{U}, \mathscr{V} \subseteq \mathbb{R}^n$ be two subspaces with $\mathscr{U} \cap \mathscr{V} = \{0\}$ and $\mathscr{U} \oplus \mathscr{V} = \mathbb{R}^n$. Let $\mathbf{u} \in \mathscr{U}$ and $\mathbf{v} \in \mathscr{V}$. Then $\mathbf{u}'\mathbf{v} = 0$.

9.6 Prove Lemma 9.9.

HINT: Use Lemma 9.8.

9.7 Prove Lemma 9.11.

9.8 Prove Lemma 9.12.

HINT: The union of respective orthonormal bases of $\mathscr{U}$ and $\mathscr{U}^\perp$ is an orthonormal basis for $\mathscr{U} \oplus \mathscr{U}^\perp$. (Why?) Now suppose that $\mathscr{U} \oplus \mathscr{U}^\perp$ is a proper subset of $\mathbb{R}^n$ and derive a contradiction by means of Theorem 5.18 and Theorem 9.6.

9.9 Prove Theorem 9.13.

9.10 Assume that $\mathbf{y} = \mathbf{U}\boldsymbol{\lambda} + \mathbf{r}$.
Show: $\mathbf{U}'\mathbf{r} = 0$ if and only if $\mathbf{U}'\mathbf{U}\boldsymbol{\lambda} = \mathbf{U}'\mathbf{y}$.

9.11 Let $\mathscr{U}$ be a subspace of $\mathbb{R}^n$ with generating set $\{\mathbf{u}_1, \ldots, \mathbf{u}_k\}$ and $\mathbf{U} = (\mathbf{u}_1, \ldots, \mathbf{u}_k)$. Show:

(a) $\mathbf{u} \in \mathscr{U}$ if and only if there exists an $\boldsymbol{\lambda} \in \mathbb{R}^k$ such that $\mathbf{u} = \mathbf{U}\boldsymbol{\lambda}$.
(b) $\mathbf{v} \in \mathscr{U}^\perp$ if and only if $\mathbf{U}'\mathbf{v} = 0$.
(c) The projection $\mathbf{y} \mapsto \mathbf{p}_U(\mathbf{y})$ is a linear map onto $\mathscr{U}$.
(d) If $\operatorname{rank}(\mathbf{U}) = k$, then the Projection matrix is given by $\mathbf{P}_U = \mathbf{U}(\mathbf{U}'\mathbf{U})^{-1}\mathbf{U}'$.

In addition:

(e) Could we simplify $\mathbf{P}_U$ in the following way?
$\mathbf{P}_U = \mathbf{U}(\mathbf{U}'\mathbf{U})^{-1}\mathbf{U}' = \mathbf{U}\mathbf{U}^{-1} \cdot \mathbf{U}'^{-1}\mathbf{U}' = \mathbf{I} \cdot \mathbf{I} = \mathbf{I}$.

(f) Let $\mathbf{P}_U$ be the matrix for projection $\mathbf{y} \mapsto \mathbf{p}_U(\mathbf{y})$. Compute the projection matrix $\mathbf{P}_{U^\perp}$ for the projection onto $\mathscr{U}^\perp$.

9.12 Prove Theorem 9.15.

9.13 Let $\mathbf{p}$ be a projection into some subspace $\mathscr{U} \subseteq \mathbb{R}^n$. Let $\mathbf{x}_1, \dots, \mathbf{x}_k \in \mathbb{R}^n$.
Show: If $\mathbf{p}(\mathbf{x}_1), \dots, \mathbf{p}(\mathbf{x}_k)$ are linearly independent, then the vectors $\mathbf{x}_1, \dots, \mathbf{x}_k$ are linearly independent.
Show that the converse is false.

9.14 (a) Give necessary and sufficient conditions such that the "normal equation" (9.2) has a uniquely determined solution.

(b) What happens when this condition is violated? (There is no solution at all? The solution exists but is not uniquely determined? How can we find solutions in the latter case? What is the statistical interpretation in all these cases?) Demonstrate your considerations by (simple) examples.

(c) Show that for each solution of Equation (9.2) the arithmetic mean of the error is zero, that is, $\bar{\boldsymbol{\varepsilon}} = 0$. Give a statistical interpretation of this result.

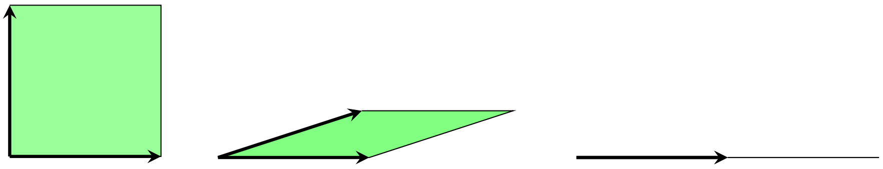
(d) Let $\mathbf{x}_i = (x_{i1}, \dots, x_{in})'$ be the i-th column of $\mathbf{X}$. Show that for each solution of Equation (9.2) $\mathbf{x}_i'\boldsymbol{\varepsilon} = 0$. Give a statistical interpretation of this result.

10 Determinant

What is the volume *of a skewed box?*

10.1 Linear Independence and Volume

We want to "measure" whether two vectors in $\mathbb{R}^2$ are linearly independent or not. Thus we may look at the parallelogram that is created by these two vectors. We may find the following cases:

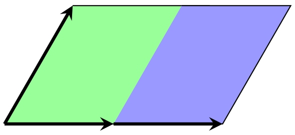


The two vectors are linearly dependent if and only if the corresponding parallelogram has area 0. The same holds for three vectors in $\mathbb{R}^3$ which form a parallelepiped and generally for n vectors in $\mathbb{R}^n$.

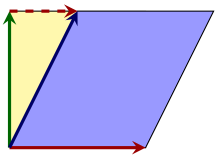
Idea: Use the n-dimensional volume to check whether n vectors in $\mathbb{R}^n$ are linearly independent.

Thus we need to compute this volume. Therefore we first look at the properties of the area of a parallelogram and the volume of a parallelepiped, respectively, and use these properties to define a "volume function".

(1) If we multiply one of the vectors by a number $\alpha \in \mathbb{R}$, then we obtain a parallelepiped (parallelogram) with the α-fold volume.



(2) If we add a multiple of one vector to one of the other vectors, then the volume remains unchanged.



(3) If two vectors are equal, then the volume is 0.

(4) The volume of the unit-cube has volume 1.

10.2 Determinant

Motivated by the above considerations we define the determinant as a *normed alternating multilinear form*.

Definition 10.1

Determinant. The **determinant** is a function det: $\mathbb{R}^{n\times n} \to \mathbb{R}$ that assigns a real number to an $n \times n$ matrix $\mathbf{A} = (\mathbf{a}_1,\dots,\mathbf{a}_n)$ with following properties:

(D1) The determinant is *multilinear*, i.e., it is linear in each column:

$$\begin{aligned}\det(&\mathbf{a}_1,\dots,\mathbf{a}_{i-1},\alpha\mathbf{a}_i+\beta\mathbf{b}_i,\mathbf{a}_{i+1},\dots,\mathbf{a}_n)\\ &= \alpha\det(\mathbf{a}_1,\dots,\mathbf{a}_{i-1},\mathbf{a}_i,\mathbf{a}_{i+1},\dots,\mathbf{a}_n)\\ &\quad+\beta\det(\mathbf{a}_1,\dots,\mathbf{a}_{i-1},\mathbf{b}_i,\mathbf{a}_{i+1},\dots,\mathbf{a}_n)\,.\end{aligned}$$

(D2) The determinant is *alternating*, i.e.,

$$\begin{aligned}&\det(\mathbf{a}_1,\dots,\mathbf{a}_{i-1},\mathbf{a}_i,\mathbf{a}_{i+1},\dots,\mathbf{a}_{k-1},\mathbf{b}_k,\mathbf{a}_{k+1},\dots,\mathbf{a}_n)\\ &=-\det(\mathbf{a}_1,\dots,\mathbf{a}_{i-1},\mathbf{b}_k,\mathbf{a}_{i+1},\dots,\mathbf{a}_{k-1},\mathbf{a}_i,\mathbf{a}_{k+1},\dots,\mathbf{a}_n)\,.\end{aligned}$$

(D3) The determinant is *normalized*, i.e.,

$$\det(\mathbf{I}) = 1\,.$$

Do not mix up with the absolute value of a number.

We denote the determinant of $\mathbf{A}$ by $\det(\mathbf{A})$ or $|\mathbf{A}|$.

This definition sounds like a “wish list”. We define the function by its properties. However, such an approach is quite common in mathematics. But of course we have to answer the following questions:

- Does such a function exist?
- Is this function uniquely defined?
- How can we evaluate the determinant of a particular matrix $\mathbf{A}$?

We proceed by deriving an explicit formula for the determinant that answers these questions. We begin with a few more properties of the determinant (provided that such a function exists). Their proofs are straightforward and left as an exercise (see Problems 10.10, 10.11, and 10.12).

Lemma 10.2

The determinant is zero if two columns are equal, i.e.,

$$\det(\dots,\mathbf{a},\dots,\mathbf{a},\dots) = 0\,.$$

Lemma 10.3

The determinant is zero, $\det(\mathbf{A}) = 0$, if the columns of $\mathbf{A}$ are linearly dependent.

Lemma 10.4

The determinant remains unchanged if we add a multiple of one column the one of the other columns:

$$\det(\dots,\mathbf{a}_i+\alpha\,\mathbf{a}_k,\dots,\mathbf{a}_k,\dots) = \det(\dots,\mathbf{a}_i,\dots,\mathbf{a}_k,\dots)\,.$$

Now let $\{\mathbf{v}_1,\dots,\mathbf{v}_n\}$ be a basis of $\mathbb{R}^n$. Then we can represent each column of $n\times n$ matrix $\mathbf{A}=(\mathbf{a}_1,\dots,\mathbf{a}_n)$ as

$$\mathbf{a}_j=\sum_{i=1}^{n} c_{ij}\mathbf{v}_i\,, \qquad \text{for } j=1,\dots,n,$$

where $c_{ij}\in\mathbb{R}$. We then find

$$\begin{aligned}
\det(\mathbf{a}_1,\mathbf{a}_2,\dots,\mathbf{a}_n) &= \det\left(\sum_{i_1=1}^{n} c_{i_1 1}\mathbf{v}_{i_1},\sum_{i_2=1}^{n} c_{i_2 2}\mathbf{v}_{i_2},\dots,\sum_{i_n=1}^{n} c_{i_n n}\mathbf{v}_{i_n}\right)\\
&= \sum_{i_1=1}^{n} c_{i_1 1}\det\left(\mathbf{v}_{i_1},\sum_{i_2=1}^{n} c_{i_2 2}\mathbf{v}_{i_2},\dots,\sum_{i_n=1}^{n} c_{i_n n}\mathbf{v}_{i_n}\right)\\
&= \sum_{i_1=1}^{n}\sum_{i_2=1}^{n} c_{i_1 1}c_{i_2 2}\det\left(\mathbf{v}_{i_1},\mathbf{v}_{i_2},\dots,\sum_{i_n=1}^{n} c_{i_n n}\mathbf{v}_{i_n}\right)\\
&\;\;\vdots\\
&= \sum_{i_1=1}^{n}\sum_{i_2=1}^{n}\cdots\sum_{i_n=1}^{n} c_{i_1 1}c_{i_2 2}\dots c_{i_n n}\det\left(\mathbf{v}_{i_1},\dots,\mathbf{v}_{i_n}\right)
\end{aligned}$$

There are n^n terms in this sum. However, Lemma 10.2 implies that $\det\left(\mathbf{v}_{i_1},\mathbf{v}_{i_2},\dots,\mathbf{v}_{i_n}\right)=0$ when at least two columns coincide. Thus only those determinants remain which contain all basis vectors $\{\mathbf{v}_1,\dots,\mathbf{v}_n\}$ (in different orders), i.e., those tuples $(i_i,i_2,\dots,i_n)$ which are a *permutation* of the numbers $(1,2,\dots,n)$.

We can define a **permutation** σ as a bijection from the set $\{1,2,\dots,n\}$ onto itself. We denote the set of these permutations by $\mathfrak{S}_n$. It has the following properties which we state without a formal proof.

- The *compound* of two permutations $\sigma,\tau\in\mathfrak{S}_n$ is again a permutation, $\sigma\tau\in\mathfrak{S}_n$.
- There is a *neutral* (or *identity*) permutation that does not change the ordering of $(1,2,\dots,n)$.
- Each permutation $\sigma\in\mathfrak{S}_n$ has a unique *inverse* permutation $\sigma^{-1}\in\mathfrak{S}_n$.

We then say that $\mathfrak{S}_n$ forms a **group**.

Using this concept we can remove the vanishing terms from the above expression for the determinant. As only determinants remain where the columns are permutations of the columns of $\mathbf{A}$ we can write

$$\det(\mathbf{a}_1,\dots,\mathbf{a}_n)=\sum_{\sigma\in\mathfrak{S}_n}\det\left(\mathbf{v}_{\sigma(1)},\dots,\mathbf{v}_{\sigma(n)}\right)\prod_{i=1}^{n} c_{\sigma(i),i}\,.$$

The simplest permutation is a **transposition** that just flips two elements.

- Every permutation can be composed of a sequence of transpositions, i.e., for every $\sigma\in\mathfrak{S}$ there exist $\tau_1,\dots,\tau_k\in\mathfrak{S}$ such that $\sigma=\tau_k\cdots\tau_1$.

Notice that a transposition of the columns of a determinant changes its sign by property (D2). An immediate consequence is that the determinants $\det\left(\mathbf{v}_{\sigma(1)},\mathbf{v}_{\sigma(2)},\ldots,\mathbf{v}_{\sigma(n)}\right)$ only differ in their signs. Moreover, the sign is given by the number of transitions into which a permutation σ is decomposed. So we have

$$\det\left(\mathbf{v}_{\sigma(1)},\ldots,\mathbf{v}_{\sigma(n)}\right) = \operatorname{sgn}(\sigma)\det(\mathbf{v}_1,\ldots,\mathbf{v}_n)$$

where $\operatorname{sgn}(\sigma) = +1$ if the number of transpositions into which σ can be decomposed is even, and where $\operatorname{sgn}(\sigma) = -1$ if the number of transpositions is odd. We remark (without proof) that $\operatorname{sgn}(\sigma)$ is well-defined although this sequence of transpositions is not unique.

We summarize our considerations in the following proposition.

Lemma 10.5

Let $\{\mathbf{v}_1,\ldots,\mathbf{v}_n\}$ be a basis of $\mathbb{R}^n$ and $\mathbf{A} = (\mathbf{a}_1,\ldots,\mathbf{a}_n)$ an $n \times n$ matrix. Let $c_{ij} \in \mathbb{R}$ such that $\mathbf{a}_j = \sum_{i=1}^n c_{ij}\mathbf{v}_i$ for $j = 1,\ldots,n$. Then

$$\det(\mathbf{a}_1,\ldots,\mathbf{a}_n) = \det(\mathbf{v}_1,\ldots,\mathbf{v}_n)\sum_{\sigma\in\mathfrak{S}_n}\operatorname{sgn}(\sigma)\prod_{i=1}^n c_{\sigma(i),i}\,.$$

This lemma allows us that we can compute $\det(\mathbf{A})$ provided that the determinant of a regular matrix is known. This equation in particular holds if we use the canonical basis $\{\mathbf{e}_1,\ldots,\mathbf{e}_n\}$. We then have $c_{ij} = a_{ij}$ and

$$\det(\mathbf{v}_1,\ldots,\mathbf{v}_n) = \det(\mathbf{e}_1,\ldots,\mathbf{e}_n) = \det(\mathbf{I}) = 1$$

where the last equality is just property (D3).

Theorem 10.6

Leibniz formula for determinant. The determinant of a $n \times n$ matrix $\mathbf{A}$ is given by

$$\det(\mathbf{A}) = \det(\mathbf{a}_1,\ldots,\mathbf{a}_n) = \sum_{\sigma\in\mathfrak{S}_n}\operatorname{sgn}(\sigma)\prod_{i=1}^n a_{\sigma(i),i}\,. \tag{10.1}$$

Corollary 10.7

Existence and uniqueness. The determinant as given in Definition 10.1 exists and is uniquely defined.

Leibniz formula (10.1) is often used as definition of the determinant. Of course we then have to derive properties (D1)–(D3) from (10.1), see Problem 10.13.

10.3 Properties of the Determinant

Theorem 10.8

Transpose. The determinant remains unchanged if a matrix is transposed, i.e.,

$$\det(\mathbf{A}') = \det(\mathbf{A})\,.$$

PROOF. Recall that $[\mathbf{A}']_{ij} = [\mathbf{A}]_{ji}$ and that each $\sigma \in \mathfrak{S}_n$ has a unique inverse permutation $\sigma^{-1} \in \mathfrak{S}_n$. Moreover, $\text{sgn}(\sigma^{-1}) = \text{sgn}(\sigma)$. Then by Theorem 10.6,

$$\begin{aligned}\det(\mathbf{A}') &= \sum_{\sigma\in\mathfrak{S}_n} \text{sgn}(\sigma) \prod_{i=1}^{n} a_{i,\sigma(i)} = \sum_{\sigma\in\mathfrak{S}_n} \text{sgn}(\sigma) \prod_{i=1}^{n} a_{\sigma^{-1}(i),i} \\ &= \sum_{\sigma\in\mathfrak{S}_n} \text{sgn}(\sigma^{-1}) \prod_{i=1}^{n} a_{\sigma^{-1}(i),i} = \sum_{\sigma\in\mathfrak{S}_n} \text{sgn}(\sigma) \prod_{i=1}^{n} a_{\sigma(i),i} = \det(\mathbf{A})\end{aligned}$$

where the forth equality holds as $\{\sigma^{-1} : \sigma \in \mathfrak{S}_n\} = \mathfrak{S}_n$. □

Product. The determinant of the product of two matrices equals the product of their determinants, i.e., Theorem 10.9

$$\det(\mathbf{A}\cdot\mathbf{B}) = \det(\mathbf{A})\cdot\det(\mathbf{B})\,.$$

PROOF. Let **A** and **B** be two $n \times n$ matrices. If **A** does not have full rank, then $\text{rank}(\mathbf{A}) < n$ and Lemma 10.3 implies $\det(\mathbf{A}) = 0$ and thus $\det(\mathbf{A})\cdot\det(\mathbf{B}) = 0$. On the other hand by Theorem 6.23 $\text{rank}(\mathbf{AB}) \le \text{rank}(\mathbf{A}) < n$ and hence $\det(\mathbf{AB}) = 0$.

If **A** has full rank, then the columns of **A** form a basis of $\mathbb{R}^n$ and we find for the columns of **AB**, $[\mathbf{AB}]_j = \sum_{i=1}^{n} b_{ij}\mathbf{a}_i$. Consequently, Lemma 10.5 and Theorem 10.6 immediately imply

$$\det(\mathbf{AB}) = \det(\mathbf{a}_1,\dots,\mathbf{a}_n) \sum_{\sigma\in\mathfrak{S}_n} \text{sgn}(\sigma) \prod_{i=1}^{n} b_{\sigma(i),i} = \det(\mathbf{A})\cdot\det(\mathbf{B})$$

as claimed. □

Singular matrix. Let **A** be an $n \times n$ matrix. Then the following are equivalent: Theorem 10.10

(1) $\det(\mathbf{A}) = 0$.

(2) The columns of **A** are linearly dependent.

(3) **A** does not have full rank.

(4) **A** is singular.

PROOF. The equivalence of (2), (3) and (4) has already been shown in Section 6.3. Implication (2) ⇒ (1) is stated in Lemma 10.3. For implication (1) ⇒ (4) see Problem 10.14. This finishes the proof. □

An $n \times n$ matrix **A** is invertible if and only if $\det(\mathbf{A}) \neq 0$. Corollary 10.11

We can use the determinant to estimate the rank of a matrix.

Theorem 10.12

Rank of a matrix. The rank of an $m \times n$ matrix $\mathbf{A}$ is r if and only if there is an $r \times r$ subdeterminant

$$\begin{vmatrix} a_{i_1 j_1} & \dots & a_{i_1 j_r} \\ \vdots & \ddots & \vdots \\ a_{i_r j_1} & \dots & a_{i_r j_r} \end{vmatrix} \neq 0$$

but all $(r+1)\times(r+1)$ subdeterminants vanish.

PROOF. By Gauß elimination we can find an invertible $r \times r$ submatrix but not an invertible $(r+1)\times(r+1)$ submatrix. □

Theorem 10.13

Inverse matrix. The determinant of the inverse of a regular matrix is the reciprocal value of the determinant of the matrix, i.e.,

$$\det(\mathbf{A}^{-1}) = \frac{1}{\det(\mathbf{A})} .$$

PROOF. See Problem 10.15.

Finally we return to the volume of a parallelepiped which we used as motivation for the definition of the determinant. Since we have no formal definition of the *volume* yet, we state the last theorem without proof.

Theorem 10.14

Volume. Let $\mathbf{a}_1, \dots, \mathbf{a}_n \in \mathbb{R}^n$. Then the volume of the n-dimensional parallelepiped created by these vectors is given by the absolute value of the determinant,

$$\operatorname{Vol}(\mathbf{a}_1, \dots, \mathbf{a}_n) = \left|\det(\mathbf{a}_1, \dots, \mathbf{a}_n)\right| .$$

10.4 Evaluation of the Determinant

Leibniz formula (10.1) provides an explicit expression for evaluating the determinant of a matrix. For small matrices one may expand sum and products and finds an easy to use scheme, known as **Sarrus' rule** (see Problems 10.17 and 10.18):

$$\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = a_{11}a_{22} - a_{21}a_{12} .$$

$$\begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} = \begin{array}{c} a_{11}a_{22}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32} \\ -a_{31}a_{22}a_{13} - a_{32}a_{23}a_{11} - a_{33}a_{21}a_{12} . \end{array} \tag{10.2}$$

For larger matrices Leibniz formula (10.1) expands to much longer expressions. For an $n \times n$ matrix we find a sum of $n!$ products of n factors. However, for triangular matrices this formula reduces to the product of the diagonal entries, see Problem 10.19.

Triangular matrix. Let $\mathbf{A}$ be an $n \times n$ (upper or lower) triangular matrix. Then Theorem 10.15

$$\det(\mathbf{A}) = \prod_{i=1}^{n} a_{ii} \, .$$

In Section 7.2 we have seen that we can transform a matrix $\mathbf{A}$ into a row echelon form $\mathbf{R}$ by a series of elementary row operations (Theorem 7.9), $\mathbf{R} = \mathbf{T}_k \cdots \mathbf{T}_1 \mathbf{A}$. Notice that for a square matrix we then obtain an upper triangular matrix. By Theorems 10.9 and 10.15 we find

$$\det(\mathbf{A}) = \big(\det(\mathbf{T}_k) \cdots \det(\mathbf{T}_1)\big)^{-1} \prod_{i=1}^{n} r_{ii} \, .$$

As $\det(\mathbf{T}_i)$ is easy to evaluate we obtain a fast algorithm for computing $\det(\mathbf{A})$, see Problems 10.20 and 10.21.

Another approach is to replace (10.1) by a recursion formula, known as **Laplace expansion**.

Minor. Let $\mathbf{A}$ be an $n \times n$ matrix. Let $\mathbf{M}_{ij}$ denote the $(n-1) \times (n-1)$ matrix that we obtain by deleting the i-th row and the j-th column from $\mathbf{A}$. Then $M_{ij} = \det(\mathbf{M}_{ij})$ is called the (i,j) **minor** of $\mathbf{A}$. Definition 10.16

Laplace expansion. Let $\mathbf{A}$ be an $n \times n$ matrix and M_{ik} its (i,k) minor. Then Theorem 10.17

$$\det(\mathbf{A}) = \sum_{i=1}^{n} a_{ik} \cdot (-1)^{i+k} M_{ik} = \sum_{k=1}^{n} a_{ik} \cdot (-1)^{i+k} M_{ik} \, .$$

The first expression is expansion along the k-th column. The second expression is expansion along the i-th row.

Cofactor. The term $C_{ik} = (-1)^{i+k} M_{ik}$ is called the **cofactor** of a_{ik}. Definition 10.18

With this notation Laplace expansion can also be written as

$$\det(\mathbf{A}) = \sum_{i=1}^{n} a_{ik} C_{ik} = \sum_{k=1}^{n} a_{ik} C_{ik} \, .$$

PROOF. As $\det(\mathbf{A}') = \det(\mathbf{A})$ we only need to prove first statement. Notice that $\mathbf{a}_k = \sum_{i=1}^{n} a_{ik} \mathbf{e}_i$. Therefore,

$$\begin{aligned} \det(\mathbf{A}) &= \det(\mathbf{a}_1, \dots, \mathbf{a}_k, \dots, \mathbf{a}_n) \\ &= \det\left(\mathbf{a}_1, \dots, \sum_{i=1}^{n} a_{ik} \mathbf{e}_i, \dots, \mathbf{a}_n\right) = \sum_{i=1}^{n} a_{ik} \det(\mathbf{a}_1, \dots, \mathbf{e}_i, \dots, \mathbf{a}_n) \, . \end{aligned}$$

It remains to show that $\det(\mathbf{a}_1, \dots, \mathbf{e}_i, \dots, \mathbf{a}_n) = C_{ik}$. Observe that we can transform matrix $(\mathbf{a}_1, \dots, \mathbf{e}_i, \dots, \mathbf{a}_n)$ into $\mathbf{B} = \begin{pmatrix} 1 & * \\ 0 & \mathbf{M}_{ik} \end{pmatrix}$ by a series of $j-1$

transpositions of rows and $k-1$ transpositions of columns and thus we find by property (D2), Theorem 10.8 and Leibniz formula

$$\begin{aligned}\det(\mathbf{a}_1,\dots,\mathbf{e}_i,\dots,\mathbf{a}_n) &= (-1)^{j+k-2}\begin{vmatrix}1 & * \\ 0 & \mathbf{M}_{ik}\end{vmatrix} = (-1)^{j+k-2}\left|\mathbf{B}\right| \\ &= (-1)^{j+k}\sum_{\sigma\in\mathfrak{S}_n}\operatorname{sgn}(\sigma)\prod_{i=1}^{n} b_{\sigma(i),i}\end{aligned}$$

Observe that $b_{11}=1$ and $b_{\sigma(1),i}=0$ for all permutations where $\sigma(1)=1$ and $i\neq 0$. Hence

$$\begin{aligned}(-1)^{j+k}\sum_{\sigma\in\mathfrak{S}_n}\operatorname{sgn}(\sigma)\prod_{i=1}^{n} b_{\sigma(i),i} &= (-1)^{j+k} b_{11}\sum_{\sigma\in\mathfrak{S}_{n-1}}\operatorname{sgn}(\sigma)\prod_{i=1}^{n-1} b_{\sigma(i)+1,i+1} \\ &= (-1)^{j+k}\left|\mathbf{M}_{ik}\right| = C_{ik}\end{aligned}$$

This finishes the proof. □

10.5 Cramer's Rule

Definition 10.19

Adjugate matrix. The matrix of cofactors for an $n\times n$ matrix $\mathbf{A}$ is the matrix $\mathbf{C}$ whose entry in the i-th row and k-th column is the cofactor C_{ik}. The **adjugate matrix** of $\mathbf{A}$ is the transpose of the matrix of cofactors of $\mathbf{A}$,

$$\operatorname{adj}(\mathbf{A}) = \mathbf{C}'\,.$$

Theorem 10.20

Let $\mathbf{A}$ be an $n\times n$ matrix. Then

$$\operatorname{adj}(\mathbf{A})\cdot\mathbf{A} = \det(\mathbf{A})\,\mathbf{I}\,.$$

PROOF. A straightforward computation and Laplace expansion (Theorem 10.17) yields

$$\begin{aligned}[\operatorname{adj}(\mathbf{A})\cdot\mathbf{A}]_{ij} &= \sum_{k=1}^{n} C'_{ik}\cdot a_{kj} = \sum_{k=1}^{n} a_{kj}\cdot C_{ki} \\ &= \det(\mathbf{a}_1,\dots,\mathbf{a}_{i-1},\mathbf{a}_j,\mathbf{a}_{i+1},\dots,\mathbf{a}_n) \\ &= \begin{cases}\det(\mathbf{A}), & \text{if } j=i, \\ 0, & \text{if } j\neq i,\end{cases}\end{aligned}$$

as claimed. □

Corollary 10.21

Inverse matrix. Let $\mathbf{A}$ be a regular $n\times n$ matrix. Then

$$\mathbf{A}^{-1} = \frac{1}{\det(\mathbf{A})}\operatorname{adj}(\mathbf{A})\,.$$

This formula is quite convenient as it provides an explicit expression for the inverse of a matrix. However, for numerical computations it is too expensive. Gauss-Jordan procedure, for example, is much faster. Nevertheless, it provides a nice rule for very small matrices.

The inverse of a regular 2 × 2 matrix **A** is given by Corollary 10.22

$$\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}^{-1} = \frac{1}{|\mathbf{A}|} \cdot \begin{pmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{pmatrix} .$$

We can use Corollary 10.21 to solve the linear equation

$$\mathbf{A} \cdot \mathbf{x} = \mathbf{b}$$

when **A** is an invertible matrix. We then find

$$\mathbf{x} = \mathbf{A}^{-1} \cdot \mathbf{b} = \frac{1}{|\mathbf{A}|} \operatorname{adj}(\mathbf{A}) \cdot \mathbf{b} .$$

Therefore we find for the i-th component of the solution **x**,

$$\begin{aligned} x_i &= \frac{1}{|\mathbf{A}|} \sum_{k=1}^{n} C'_{ik} \cdot b_k = \frac{1}{|\mathbf{A}|} \sum_{k=1}^{n} b_k \cdot C_{ki} \\ &= \frac{1}{|\mathbf{A}|} \det(\mathbf{a}_1, \ldots, \mathbf{a}_{i-1}, \mathbf{b}, \mathbf{a}_{i+1}, \ldots, \mathbf{a}_n) . \end{aligned}$$

So we get the following explicit expression for the solution of a linear equation.

Cramer's rule. Let **A** be an invertible matrix and **x** a solution of the linear equation $\mathbf{A} \cdot \mathbf{x} = \mathbf{b}$. Let $\mathbf{A}_i$ denote the matrix where the i-th column of **A** is replaced by **b**. Then Theorem 10.23

$$x_i = \frac{\det(\mathbf{A}_i)}{\det(\mathbf{A})} .$$

— Summary

- The determinant is a *normed alternating multilinear form*.
- The determinant is 0 if and only if it is singular.
- The determinant of the product of two matrices is the product of the determinants of the matrices.
- The Leibniz formula gives an explicit expression for the determinant.
- The Laplace expansion is a recursive formula for evaluating the determinant.
- The determinant can efficiently computed by a method similar to Gauß elimination.
- Cramer's rule allows to compute the inverse of matrices and the solutions of special linear equations.

— Exercises

10.1 Compute the following determinants by means of Sarrus" rule or by transforming into an upper triangular matrix:

(a) $\begin{pmatrix}1 & 2\\2 & 1\end{pmatrix}$ (b) $\begin{pmatrix}-2 & 3\\1 & 3\end{pmatrix}$ (c) $\begin{pmatrix}4 & -3\\0 & 2\end{pmatrix}$

(d) $\begin{pmatrix}3 & 1 & 1\\0 & 1 & 0\\3 & 2 & 1\end{pmatrix}$ (e) $\begin{pmatrix}2 & 1 & -4\\2 & 1 & 4\\3 & 4 & -4\end{pmatrix}$ (f) $\begin{pmatrix}0 & -2 & 1\\2 & 2 & 1\\4 & -3 & 3\end{pmatrix}$

(g) $\begin{pmatrix}1 & 2 & 3 & -2\\0 & 4 & 5 & 0\\0 & 0 & 6 & 3\\0 & 0 & 0 & 2\end{pmatrix}$ (h) $\begin{pmatrix}2 & 0 & 0 & 1\\0 & 1 & 0 & 2\\0 & 0 & 7 & 0\\1 & 2 & 0 & 1\end{pmatrix}$ (i) $\begin{pmatrix}1 & 1 & 1 & 1\\1 & 1 & 1 & 1\\1 & 1 & 1 & 1\\1 & 1 & 1 & 1\end{pmatrix}$

10.2 Compute the determinants from Exercise 10.1 by means of Laplace expansion.

10.3

(a) Estimate the ranks of the matrices from Exercise 10.1.

(b) Which of these matrices are regular?

(c) Which of these matrices are invertible?

(d) Are the column vectors of these matrices linear indpendent?

10.4 Let

$$\mathbf{A}=\begin{pmatrix}3 & 1 & 0\\0 & 1 & 0\\1 & 0 & 1\end{pmatrix},\quad \mathbf{B}=\begin{pmatrix}3 & 2\times 1 & 0\\0 & 2\times 1 & 0\\1 & 2\times 0 & 1\end{pmatrix}\quad\text{and}\quad \mathbf{C}=\begin{pmatrix}3 & 5\times 3+1 & 0\\0 & 5\times 0+1 & 0\\1 & 5\times 1+0 & 1\end{pmatrix}$$

Compute by means of the properties of determinants:

(a) $\det(\mathbf{A})$ (b) $\det(5\mathbf{A})$ (c) $\det(\mathbf{B})$ (d) $\det(\mathbf{A}')$

(e) $\det(\mathbf{C})$ (f) $\det(\mathbf{A}^{-1})$ (g) $\det(\mathbf{A}\cdot\mathbf{C})$ (h) $\det(\mathbf{I})$

10.5 Let $\mathbf{A}$ be a 3×4 matrix. Estimate $|\mathbf{A}'\cdot\mathbf{A}|$ and $|\mathbf{A}\cdot\mathbf{A}'|$.

10.6 Compute area of the parallelogram and volume of the parelelepiped, respectively, which are created by the following vectors:

(a) $\begin{pmatrix}-2\\3\end{pmatrix},\begin{pmatrix}1\\3\end{pmatrix}$ (b) $\begin{pmatrix}-2\\1\end{pmatrix},\begin{pmatrix}3\\3\end{pmatrix}$

(c) $\begin{pmatrix}2\\1\\-4\end{pmatrix},\begin{pmatrix}2\\1\\4\end{pmatrix},\begin{pmatrix}3\\4\\-4\end{pmatrix}$ (d) $\begin{pmatrix}2\\2\\3\end{pmatrix},\begin{pmatrix}1\\1\\4\end{pmatrix},\begin{pmatrix}-4\\4\\-4\end{pmatrix}$

10.7 Compute the matrix of cofactors, the adjugate matrix and the inverse of the following matrices:

(a) $\begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}$ (b) $\begin{pmatrix} -2 & 3 \\ 1 & 3 \end{pmatrix}$ (c) $\begin{pmatrix} 4 & -3 \\ 0 & 2 \end{pmatrix}$

(d) $\begin{pmatrix} 3 & 1 & 1 \\ 0 & 1 & 0 \\ 0 & 2 & 1 \end{pmatrix}$ (e) $\begin{pmatrix} 2 & 1 & -4 \\ 2 & 1 & 4 \\ 3 & 4 & -4 \end{pmatrix}$ (f) $\begin{pmatrix} 0 & -2 & 1 \\ 2 & 2 & 1 \\ 4 & -3 & 3 \end{pmatrix}$

10.8 Compute the inverse of the following matrices:

(a) $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ (b) $\begin{pmatrix} x_1 & y_1 \\ x_2 & y_2 \end{pmatrix}$ (c) $\begin{pmatrix} \alpha & \beta \\ \alpha^2 & \beta^2 \end{pmatrix}$

10.9 Solve the linear equation

$$\mathbf{A} \cdot \mathbf{x} = \mathbf{b}$$

by means of Cramer's rule for $\mathbf{b} = (1,2)'$ and $\mathbf{b} = (1,2,3)$, respetively, and the following matrices:

(a) $\begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}$ (b) $\begin{pmatrix} -2 & 3 \\ 1 & 3 \end{pmatrix}$ (c) $\begin{pmatrix} 4 & -3 \\ 0 & 2 \end{pmatrix}$

(d) $\begin{pmatrix} 3 & 1 & 1 \\ 0 & 1 & 0 \\ 0 & 2 & 1 \end{pmatrix}$ (e) $\begin{pmatrix} 2 & 1 & -4 \\ 2 & 1 & 4 \\ 3 & 4 & -4 \end{pmatrix}$ (f) $\begin{pmatrix} 0 & -2 & 1 \\ 2 & 2 & 1 \\ 4 & -3 & 3 \end{pmatrix}$

— Problems

10.10 Proof Lemma 10.2 using properties (D1)–(D3).

10.11 Proof Lemma 10.3 using properties (D1)–(D3).

10.12 Proof Lemma 10.4 using properties (D1)–(D3).

10.13 Derive properties (D1) and (D3) from Expression (10.1) in Theorem 10.6.

10.14 Show that an $n \times n$ matrix $\mathbf{A}$ is singular if $\det(\mathbf{A}) = 0$.
Does Lemma 10.3 already imply this result?

HINT: Try an indirect proof and use equation $\mathbf{I} = \det(\mathbf{AA}^{-1})$.

10.15 Prove Theorem 10.13.

10.16 Show that the determinants of similar square matrices are equal.

10.17 Derive formula

$$\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = a_{11}a_{22} - a_{21}a_{12}$$

directly from properties (D1)–(D3) and Lemma 10.4.

HINT: Use a method similar to Gauß elimination.

10.18 Derive Sarrus' rule (10.2) from Leibniz formula (10.1).

10.19 Let $\mathbf{A}$ be an $n \times n$ upper triangular matrix. Show that

$$\det(\mathbf{A}) = \prod_{i=1}^{n} a_{ii}\,.$$

HINT: Use Leibniz formula (10.1) and show that there is only one permutation σ with $\sigma(i) \leq i$ for all i.

10.20 Compute the determinants of the elementary row operations from Problem 7.4.

10.21 Modify the algorithm from Problem 7.6 such that it computes the determinant of a square matrix.

11

Eigenspace

We want to estimate the sign of a matrix and compute its square root.

11.1 Eigenvalues and Eigenvectors

Eigenvalues and eigenvectors. Let $\mathbf{A}$ be an $n \times n$ matrix. Then a non-zero vector $\mathbf{x}$ is called an **eigenvector** corresponding to **eigenvalue** λ if

Definition 11.1

$$\mathbf{A}\mathbf{x} = \lambda\mathbf{x}\,, \quad \mathbf{x} \neq 0\,. \tag{11.1}$$

Observe that a scalar λ is an eigenvalue if and only if $(\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = 0$ has a non-trivial solution, i.e., if $(\mathbf{A} - \lambda\mathbf{I})$ is not invertible or, equivalently, if and only if

$$\det(\mathbf{A} - \lambda\mathbf{I}) = 0\,.$$

The Leibniz formula for determinants (or, equivalently, Laplace expansion) implies that this determinant is a polynomial of degree n in λ.

Characteristic polynomial. The polynomial

Definition 11.2

$$p_{\mathbf{A}}(t) = \det(\mathbf{A} - t\mathbf{I})$$

is called the **characteristic polynomial** of $\mathbf{A}$. For this reason the eigenvalues of $\mathbf{A}$ are also called its **characteristic roots** and the corresponding eigenvectors the **characteristic vectors** of $\mathbf{A}$.

Notice that by the Fundamental Theorem of Algebra a polynomial of degree n has exactly n roots (in the sense we can factorize the polynomial into a product of n linear terms), i.e., we can write

$$p_{\mathbf{A}}(t) = (-1)^n(t-\lambda_1)\cdots(t-\lambda_n) = (-1)^n \prod_{i=1}^{n}(t-\lambda_i)\,.$$

However, some of these roots λ_i may be complex numbers.

If an eigenvalue λ_i appears m times ($m \geq 2$) as a linear factor, i.e., if it is a multiple root of the characteristic polynomial $p_{\mathbf{A}}(t)$, then we say that λ_i has **algebraic multiplicity** m.

Definition 11.3

Spectrum. The list of all eigenvalues of a square matrix $\mathbf{A}$ is called the **spectrum** of $\mathbf{A}$. It is denoted by $\sigma(\mathbf{A})$.

Obviously, the eigenvectors corresponding to eigenvalue λ are the solutions of the homogeneous linear equation $(\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = 0$. Therefore, the set of all eigenvectors with the same eigenvalue λ together with the zero vector is the subspace $\ker(\mathbf{A} - \lambda\mathbf{I})$.

Definition 11.4

Eigenspace. Let λ be an eigenvalue of the $n \times n$ matrix $\mathbf{A}$. The subspace

$$\mathcal{E}_\lambda = \ker(\mathbf{A} - \lambda\mathbf{I})$$

is called the **eigenspace** of $\mathbf{A}$ corresponding to eigenvalue λ.

Computer programs for computing eigenvectors thus always compute bases of the corresponding eigenspaces. Since bases of a subspace are not unique, see Section 5.2, their results may differ.

Example 11.5

Diagonal matrix. For every $n \times n$ diagonal matrix $\mathbf{D}$ and every $i = 1, \ldots, n$ we find

$$\mathbf{D}\mathbf{e}_i = d_{ii}\mathbf{e}_i\,.$$

That is, each of its diagonal entries d_{ii} is an eigenvalue affording eigenvectors $\mathbf{e}_i$. Its spectrum is just the set of its diagonal entries. ♢

11.2 Properties of Eigenvalues

Theorem 11.6

Transpose. $\mathbf{A}$ and $\mathbf{A}'$ have the same spectrum.

PROOF. See Problem 11.14.

Theorem 11.7

Matrix power. If $\mathbf{x}$ is an eigenvector of $\mathbf{A}$ corresponding to eigenvalue λ, then $\mathbf{x}$ is also an eigenvector of $\mathbf{A}^k$ corresponding to eigenvalue λ^k for every $k \in \mathbb{N}$.

PROOF. See Problem 11.15.

Theorem 11.8

Inverse matrix. If $\mathbf{x}$ is an eigenvector of the regular matrix $\mathbf{A}$ corresponding to eigenvalue λ, then $\mathbf{x}$ is also an eigenvector of $\mathbf{A}^{-1}$ corresponding to eigenvalue λ^{-1}.

PROOF. See Problem 11.16.

Eigenvalues and determinant. Let $\mathbf{A}$ be an $n \times n$ matrix with eigenvalues $\lambda_1, \dots, \lambda_n$ (counting multiplicity). Then Theorem 11.9

$$\det(\mathbf{A}) = \prod_{i=1}^{n} \lambda_i \,.$$

PROOF. A straightforward computation shows that $\prod_{i=1}^{n} \lambda_i$ is the constant term of the characteristic polynomial $p_{\mathbf{A}}(t) = (-1)^n \prod_{i=1}^{n}(t - \lambda_i)$. On the other hand, we show that the constant term of $p_{\mathbf{A}}(t) = \det(\mathbf{A} - t\mathbf{I})$ equals $\det(\mathbf{A})$. Observe that by multilinearity of the determinant we have

$$\det(\dots, \mathbf{a}_i - t\mathbf{e}_i, \dots) = \det(\dots, \mathbf{a}_i, \dots) - t \det(\dots, \mathbf{e}_i, \dots)\,.$$

As this holds for every columns we find

$$\det(\mathbf{A} - t\mathbf{I}) = \sum_{(\delta_1, \dots, \delta_n) \in \{0,1\}^n} (-t)^{\sum_{i=1}^{n} \delta_i} \det\big((1-\delta_1)\mathbf{a}_1 + \delta_1 \mathbf{e}_1, \dots, \\ \dots, (1-\delta_n)\mathbf{a}_n + \delta_n \mathbf{e}_n\big)\,.$$

Obviously, the only term that does not depend on t is where $\delta_1 = \dots = \delta_n = 0$, i.e., $\det(\mathbf{A})$. This completes the proof. □

There is also a similar remarkable result on the sum of the eigenvalues.

The **trace** of an $n \times n$ matrix $\mathbf{A}$ is the sum of its diagonal elements, i.e., Definition 11.10

$$\operatorname{tr}(\mathbf{A}) = \sum_{i=1}^{n} a_{ii}\,.$$

Eigenvalues and trace. Let $\mathbf{A}$ be an $n \times n$ matrix with eigenvalues $\lambda_1, \dots, \lambda_n$ (counting multiplicity). Then Theorem 11.11

$$\operatorname{tr}(\mathbf{A}) = \sum_{i=1}^{n} \lambda_i\,.$$

PROOF. See Problem 11.17.

11.3 Diagonalization and Spectral Theorem

In Section 6.4 we have called two matrices $\mathbf{A}$ and $\mathbf{B}$ similar if there exists a transformation matrix $\mathbf{U}$ such that $\mathbf{B} = \mathbf{U}^{-1}\mathbf{A}\mathbf{U}$.

Similar matrices. The spectra of two similar matrices $\mathbf{A}$ and $\mathbf{B}$ coincide. Theorem 11.12

PROOF. See Problem 11.18.

Now one may ask whether we can find a basis such that the corresponding matrix is as simple as possible. Motivated by Example 11.5 we even may try to find a basis such that $\mathbf{A}$ becomes a diagonal matrix. We find that this is indeed the case for symmetric matrices.

Theorem 11.13

Spectral theorem for symmetric matrices. Let $\mathbf{A}$ be a symmetric $n \times n$ matrix. Then all eigenvalues are real and there exists an orthonormal basis $\{\mathbf{u}_1, \dots, \mathbf{u}_n\}$ of $\mathbb{R}^n$ consisting of eigenvectors of $\mathbf{A}$.

Furthermore, let $\mathbf{D}$ be the $n \times n$ diagonal matrix with the eigenvalues of $\mathbf{A}$ as its entries and let $\mathbf{U} = (\mathbf{u}_1, \dots, \mathbf{u}_n)$ be the orthogonal matrix of eigenvectors. Then matrices $\mathbf{A}$ and $\mathbf{D}$ are similar with transformation matrix $\mathbf{U}$, i.e.,

$$\mathbf{U}'\mathbf{A}\mathbf{U} = \mathbf{D} = \begin{pmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_n \end{pmatrix} . \qquad (11.2)$$

We call this process the **diagonalization** of $\mathbf{A}$.

A proof of the first part of Theorem 11.13 is out of the scope of this manuscript. Thus we only show the following partial result (Lemma 11.14). For the second part recall that for an orthogonal matrix $\mathbf{U}$ we have $\mathbf{U}^{-1} = \mathbf{U}'$ by Theorem 8.24. Moreover, observe that

$$\mathbf{U}'\mathbf{A}\mathbf{U}\,\mathbf{e}_i = \mathbf{U}'\mathbf{A}\,\mathbf{u}_i = \mathbf{U}'\,\lambda_i \mathbf{u}_i = \lambda_i \mathbf{U}'\,\mathbf{u}_i = \lambda_i \mathbf{e}_i = \mathbf{D}\,\mathbf{e}_i$$

for all $i = 1, \dots, n$.

Lemma 11.14

Let $\mathbf{A}$ be a symmetric $n \times n$ matrix. If $\mathbf{u}_i$ and $\mathbf{u}_j$ are eigenvectors to distinct eigenvalues λ_i and λ_j, respectively, then $\mathbf{u}_i$ and $\mathbf{u}_j$ are orthogonal, i.e., $\mathbf{u}_i'\mathbf{u}_j = 0$.

PROOF. By the symmetry of $\mathbf{A}$ and eigenvalue equation (11.1) we find

$$\lambda_i \mathbf{u}_i'\mathbf{u}_j = (\mathbf{A}\mathbf{u}_i)'\mathbf{u}_j) = (\mathbf{u}_i'\mathbf{A}')\mathbf{u}_j) = \mathbf{u}_i'(\mathbf{A}\mathbf{u}_j) = \mathbf{u}_i'(\lambda_j \mathbf{u}_j) = \lambda_j \mathbf{u}_i'\mathbf{u}_j .$$

Consequently, if $\lambda_i \neq \lambda_j$ then $\mathbf{u}_i'\mathbf{u}_j = 0$, as claimed □

Theorem 11.13 immediately implies Theorem 11.9 for the special case where $\mathbf{A}$ is symmetric, see Problem 11.19.

11.4 Quadratic Forms

Up to this section we only have dealt with linear functions. Now we want to look to more advanced functions, in particular at quadratic functions.

Quadratic form. Let $\mathbf{A}$ be a symmetric $n \times n$ matrix. Then the function Definition 11.15

$$q_{\mathbf{A}} \colon \mathbb{R}^n \to \mathbb{R}, \mathbf{x} \mapsto q_{\mathbf{A}}(\mathbf{x}) = \mathbf{x}'\mathbf{A}\mathbf{x}$$

is called a **quadratic form**.

Observe that we have

$$q_{\mathbf{A}}(\mathbf{x}) = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij} x_i x_j \,.$$

In the second part of this course we need to characterize stationary points of arbitrary differentiable multivariate functions. We then will see that the sign of such quadratic forms will play a prominent rôle in our investigations. Hence we introduce the concept of the *definiteness* of a quadratic form.

Definiteness. A quadratic form $q_{\mathbf{A}}$ is called Definition 11.16

- **positive definite**, if $q_{\mathbf{A}}(\mathbf{x}) > 0$ for all $\mathbf{x} \neq 0$;
- **positive semidefinite**, if $q_{\mathbf{A}}(\mathbf{x}) \geq 0$ for all $\mathbf{x}$;
- **negative definite**, if $q_{\mathbf{A}}(\mathbf{x}) < 0$ for all $\mathbf{x} \neq 0$;
- **negative semidefinite**, if $q_{\mathbf{A}}(\mathbf{x}) \leq 0$ for all $\mathbf{x}$;
- **indefinite** in all other cases.

In abuse of language we call $\mathbf{A}$ *positive* (*negative*) (*semi*) *definite* if the corresponding quadratic form has this property.

Notice that we can reduce the definition of *negative definite* to that of *positive definite*, see Problem 11.21. Thus the treatment of the negative definite case could be omitted at all.

The quadratic form $q_{\mathbf{A}}$ is negative definite if and only if $q_{-\mathbf{A}}$ is positive definite. Lemma 11.17

By Theorem 11.13 a symmetric matrix $\mathbf{A}$ is similar to a diagonal matrix $\mathbf{D}$ and we find $\mathbf{U}'\mathbf{A}\mathbf{U} = \mathbf{D}$. Thus if $\mathbf{c}$ is the coefficient vector of a vector $\mathbf{x}$ with respect to the orthonormal basis of eigenvectors of $\mathbf{A}$, then we find

$$\mathbf{x} = \sum_{i=1}^{n} c_i \mathbf{u}_i = \mathbf{U}\mathbf{c}$$

and thus

$$q_{\mathbf{A}}(\mathbf{x}) = \mathbf{x}'\mathbf{A}\mathbf{x} = (\mathbf{U}\mathbf{c})'\mathbf{A}(\mathbf{U}\mathbf{c}) = \mathbf{c}'\mathbf{U}'\mathbf{A}\mathbf{U}\mathbf{c} = \mathbf{c}'\mathbf{D}\mathbf{c}$$

that is,

$$q_{\mathbf{A}}(\mathbf{x}) = \sum_{i=1}^{n} \lambda_i c_i^2 \,.$$

Obviously, the definiteness of $q_{\mathbf{A}}$ solely depends on the signs of the eigenvalues of $\mathbf{A}$.

Theorem 11.18

Definiteness and eigenvalues. Let $\mathbf{A}$ be symmetric matrix with eigenvalues $\lambda_1,\dots,\lambda_n$. Then the quadratic form $q_{\mathbf{A}}$ is

- *positive definite* if and only if all $\lambda_i > 0$;
- *positive semidefinite* if and only if all $\lambda_i \geq 0$;
- *negative definite* if and only if all $\lambda_i < 0$;
- *negative semidefinite* if and only if all $\lambda_i \leq 0$;
- *indefinite* if and only if there are positive and negative eigenvalues.

Computing eigenvalues requires to find all roots of a polynomial. While this is quite simple for a quadratic term, it becomes cumbersome for cubic and quartic equations and there is no explicit solution for polynomials of degree 5 or higher. Then only numeric methods are available. Fortunately, there exists an alternative method for determine the definiteness of a matrix, called *Sylvester's criterion*, that requires the computation of so called minors.

Definition 11.19

Leading principle minor. Let $\mathbf{A}$ be an $n \times n$ matrix. For $k = 1,\dots,n$, the k-th *leading principle submatrix* is the $k \times k$ submatrix formed from the first k rows and first k columns of $\mathbf{A}$. The k-th **leading principle minor** is the determinant of this submatrix, i.e.,

$$H_k = \begin{vmatrix} a_{1,1} & a_{1,2} & \dots & a_{1,k} \\ a_{2,1} & a_{2,2} & \dots & a_{2,k} \\ \vdots & \vdots & \ddots & \vdots \\ a_{k,1} & a_{k,2} & \dots & a_{k,k} \end{vmatrix}$$

Theorem 11.20

Sylvester's criterion. A symmetric $n \times n$ matrix $\mathbf{A}$ is positive definite if and only if all its leading principle minors are positive.

It is easy to prove that positive leading principle minors are a necessary condition for the positive definiteness of $\mathbf{A}$, see Problem 11.22. For the sufficiency of this condition we first show an auxiliary result[1].

Lemma 11.21

Let $\mathbf{A}$ be a symmetric $n \times n$ matrix. If $\mathbf{x}'\mathbf{A}\mathbf{x} > 0$ for all nonzero vectors $\mathbf{x}$ in a k-dimensional subspace $\mathcal{V}$ of $\mathbb{R}^n$, then $\mathbf{A}$ has at least k positive eigenvalues (counting multiplicity).

PROOF. Suppose that $m < k$ eigenvalues are positive but the rest are not. Let $\mathbf{u}_{m+1},\dots,\mathbf{u}_n$ be the eigenvectors corresponding to the non-positive eigenvalues $\lambda_{m+1},\dots,\lambda_n \leq 0$ and let Let $\mathcal{U} = \operatorname{span}(\mathbf{u}_{m+1},\dots,\mathbf{u}_n)$. Since $\mathcal{V} + \mathcal{U} \subseteq \mathbb{R}^n$ the formula from Problem 5.14 implies that

$$\begin{aligned} \dim(\mathcal{V} \cap \mathcal{U}) &= \dim(\mathcal{V}) + \dim(\mathcal{U}) - \dim(\mathcal{V} + \mathcal{U}) \\ &\geq k + (n-m) - n = k - m > 0\,. \end{aligned}$$

[1] We essentially follow a proof by G. T. Gilbert (1991), *Positive definite matrices and Sylvester's criterion*, The American Mathematical Monthly 98(1): 44–46, DOI: 10.2307/2324036.

Hence $\mathcal{V}$ and $\mathcal{U}$ have non-trivial intersection and there exists a non-zero vector $\mathbf{v} \in \mathcal{V}$ that can be written as

$$\mathbf{v} = \sum_{i=m+1}^{n} c_i \mathbf{u}_i$$

and we have

$$\mathbf{v}'\mathbf{A}\mathbf{v} = \sum_{i=m+1}^{n} \lambda_i c_i^2 \le 0$$

a contradiction. Thus $m \ge k$, as desired. □

PROOF OF THEOREM 11.20. We complete the proof of sufficiency by induction. For $n = 1$, the result is trivial. Assume the sufficiency of positive leading principle minors of $(n-1)\times(n-1)$ matrices. So if $\mathbf{A}$ is a symmetric $n \times n$ matrix, its $(n-1)$st leading principle submatrix is positive definite. Then for any non-zero vector $\mathbf{v}$ with $v_n = 0$ we find $\mathbf{v}'\mathbf{A}\mathbf{v} > 0$. As the subspace of all such vectors has dimension $n-1$ Lemma 11.21 implies that $\mathbf{A}$ has at least $n-1$ positive eigenvalues (counting multiplicities). Since $\det(\mathbf{A}) > 0$ we conclude by Theorem 11.9 that all n eigenvalues of $\mathbf{A}$ are positive and hence $\mathbf{A}$ is positive definite by Theorem 11.18. This completes the proof. □

By means of Sylvester's criterion we immediately get the following characterizations, see Problem 11.23.

Definiteness and leading principle minors. A symmetric $n \times n$ matrix $\mathbf{A}$ is Theorem 11.22

- *positive definite* if and only if all $H_k > 0$ for $1 \le k \le n$;
- *negative definite* if and only if all $(-1)^k H_k > 0$ for $1 \le k \le n$; and
- *indefinite* if $\det(\mathbf{A}) \ne 0$ but $\mathbf{A}$ is neither positive nor negative definite.

Unfortunately, for a characterization of positive and negative semidefinite matrices the sign of leading principle minors is not sufficient, see Problem 11.25. We then have to look at the sign of a lot more determinants.

Principle minor. Let $\mathbf{A}$ be an $n \times n$ matrix. For $k = 1, \dots, n$, a k-th **principle minor** is the determinant of the $k \times k$ submatrix formed from the same set of rows and columns of $\mathbf{A}$, i.e., for $1 \le i_1 < i_2 < \cdots < i_k \le n$ we obtain the minor Definition 11.23

$$M_{i_1,\dots,i_k} = \begin{vmatrix} a_{i_1,i_1} & a_{i_1,i_2} & \dots & a_{i_1,i_k} \\ a_{i_2,i_1} & a_{i_2,i_2} & \dots & a_{i_2,i_k} \\ \vdots & \vdots & \ddots & \vdots \\ a_{i_k,i_1} & a_{i_k,i_2} & \dots & a_{i_k,i_k} \end{vmatrix}$$

Notice that there are $\binom{n}{k}$ many k-th principle minors which gives a total of $2^n - 1$. The following criterion we state without a formal proof.

Theorem 11.24

Semifiniteness and principle minors. A symmetric $n \times n$ matrix $\mathbf{A}$ is

- *positive semidefinite* if and only if all principle minors are non-negative, i.e., $M_{i_1,\dots,i_k} \geq 0$ for all $1 \leq k \leq n$ and all $1 \leq i_1 < i_2 < \cdots < i_k \leq n$.
- *positive semidefinite* if and only if $(-1)^k M_{i_1,\dots,i_k} \geq 0$ for all $1 \leq k \leq n$ and all $1 \leq i_1 < i_2 < \cdots < i_k \leq n$.
- *indefinite* in all other cases.

11.5 Spectral Decomposition and Functions of Matrices

We may state the Spectral Theorem 11.13 in a different way. Observe that Equation (11.2) implies

$$\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{U}' .$$

Observe that $[\mathbf{U}'\mathbf{x}]_i = \mathbf{u}_i'\mathbf{x}$ and thus $\mathbf{U}'\mathbf{x} = \sum_{i=1}^n (\mathbf{u}_i'\mathbf{x})\,\mathbf{e}_i$. Then a straightforward computation yields

$$\begin{aligned}\mathbf{A}\mathbf{x} = \mathbf{U}\mathbf{D}\mathbf{U}'\mathbf{x} &= \mathbf{U}\mathbf{D}\sum_{i=1}^n (\mathbf{u}_i'\mathbf{x})\,\mathbf{e}_i = \sum_{i=1}^n (\mathbf{u}_i'\mathbf{x})\mathbf{U}\mathbf{D}\,\mathbf{e}_i = \sum_{i=1}^n (\mathbf{u}_i'\mathbf{x})\mathbf{U}\lambda_i\mathbf{e}_i \\ &= \sum_{i=1}^n \lambda_i(\mathbf{u}_i'\mathbf{x})\mathbf{U}\mathbf{e}_i = \sum_{i=1}^n \lambda_i(\mathbf{u}_i'\mathbf{x})\mathbf{u}_i = \sum_{i=1}^n \lambda_i\mathbf{p}_i(\mathbf{x})\end{aligned}$$

where $\mathbf{p}_i$ is just the orthogonal projection onto $\text{span}(\mathbf{u}_i)$, see Definition 9.2. By Theorem 9.4 there exists a projection matrix $\mathbf{P}_i = \mathbf{u}_i\mathbf{u}_i'$, such that $\mathbf{p}_i(\mathbf{x}) = \mathbf{P}_i\mathbf{x}$. Therefore we arrive at the following **spectral decomposition**,

$$\mathbf{A} = \sum_{i=1}^n \lambda_i\mathbf{P}_i \,. \tag{11.3}$$

A simple computation gives that $\mathbf{A}^k = \mathbf{U}\mathbf{D}^k\mathbf{U}'$, see Problem 11.20, or using Equation (11.3)

$$\mathbf{A}^k = \sum_{i=1}^n \lambda_i^k\mathbf{P}_i \,.$$

Thus by means of the spectral decomposition we can compute integer powers of a matrix. Similarly, we find

$$\mathbf{A}^{-1} = \mathbf{U}\mathbf{D}^{-1}\mathbf{U}' = \sum_{i=1}^n \lambda_i^{-1}\mathbf{P}_i \,.$$

Can we compute other functions of a symmetric matrix as well as, e.g., its square root?

Square root. A matrix **B** is called the **square root** of a symmetric matrix **A** if $\mathbf{B}^2 = \mathbf{A}$.

Definition 11.25

Let $\mathbf{B} = \sum_{i=1}^{n} \sqrt{\lambda_i}\mathbf{P}_i$ then $\mathbf{B}^2 = \sum_{i=1}^{n} \left(\sqrt{\lambda_i}\right)^2 \mathbf{P}_i = \sum_{i=1}^{n} \lambda_i \mathbf{P}_i = \mathbf{A}$, provided that all eigenvalues of **A** are positive.

This motivates to define any function of a matrix in the following way: Let $f : \mathbb{R} \to \mathbb{R}$ some function. Then

$$f(\mathbf{A}) = \sum_{i=1}^{n} f(\lambda_i)\mathbf{P}_i = \mathbf{U} \begin{pmatrix} f(\lambda_1) & 0 & \dots & 0 \\ 0 & f(\lambda_2) & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & f(\lambda_n) \end{pmatrix} \mathbf{U}' .$$

— Summary

- An *eigenvalue* and its corresponding *eigenvector* of an $n \times n$ matrix **A** satisfy the equation $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$.
- The polynomial $\det(\mathbf{A} - \lambda\mathbf{I}) = 0$ is called the *characteristic polynomial* of **A** and has degree n.
- The set of all eigenvectors corresponding to an eigenvalue λ forms a subspace and is called *eigenspace*.
- The product and sum of all eigenvalue equals the *determinant* and *trace*, resp., of the matrix.
- *Similar* matrices have the same spectrum.
- Every *symmetric* matrix is similar to diagonal matrix with its eigenvalues as entries. The transformation matrix is an orthogonal matrix that contains the corresponding eigenvectors.
- The definiteness of a *quadratic form* can be determined by means of the eigenvalues of the underlying symmetric matrix.
- Alternatively, it can be computed by means of principle minors.
- Spectral decompositions allows to compute functions of symmetric matrices.

— Exercises

11.1 Compute eigenvalues and eigenvectors of the following matrices:

(a) $\mathbf{A}=\begin{pmatrix}3&2\\2&6\end{pmatrix}$ (b) $\mathbf{B}=\begin{pmatrix}2&3\\4&13\end{pmatrix}$ (c) $\mathbf{C}=\begin{pmatrix}-1&5\\5&-1\end{pmatrix}$

11.2 Compute eigenvalues and eigenvectors of the following matrices:

(a) $\mathbf{A}=\begin{pmatrix}1&-1&0\\-1&1&0\\0&0&2\end{pmatrix}$ (b) $\mathbf{B}=\begin{pmatrix}4&0&1\\-2&1&0\\-2&0&1\end{pmatrix}$

(c) $\mathbf{C}=\begin{pmatrix}1&2&2\\1&2&-1\\-1&1&4\end{pmatrix}$ (d) $\mathbf{D}=\begin{pmatrix}-3&0&0\\0&-5&0\\0&0&-9\end{pmatrix}$

(e) $\mathbf{E}=\begin{pmatrix}3&1&1\\0&1&0\\3&2&1\end{pmatrix}$ (f) $\mathbf{F}=\begin{pmatrix}11&4&14\\4&-1&10\\14&10&8\end{pmatrix}$

11.3 Compute eigenvalues and eigenvectors of the following matrices:

(a) $\mathbf{A}=\begin{pmatrix}1&0&0\\0&1&0\\0&0&1\end{pmatrix}$ (b) $\mathbf{B}=\begin{pmatrix}1&1&1\\0&1&1\\0&0&1\end{pmatrix}$

11.4 Estimate the definiteness of the matrices from Exercises 11.1a, 11.1c, 11.2a, 11.2d, 11.2f and 11.3a.

What can you say about the definiteness of the other matrices from Exercises 11.1, 11.2 and 11.3?

11.5 Let $\mathbf{A}=\begin{pmatrix}3&2&1\\2&-2&0\\1&0&-1\end{pmatrix}$. Give the quadratic form that is generated by $\mathbf{A}$.

11.6 Let $q(\mathbf{x})=5x_1^2+6x_1x_2-2x_1x_3+x_2^2-4x_2x_3+x_3^2$ be a quadratic form. Give its corresponding matrix $\mathbf{A}$.

11.7 Compute the eigenspace of matrix

$$\mathbf{A}=\begin{pmatrix}1&-1&0\\-1&1&0\\0&0&2\end{pmatrix}.$$

11.8 Demonstrate the following properties of eigenvalues:

(1) Quadratic matrices $\mathbf{A}$ and $\mathbf{A}'$ have the same spectrum.
(Do they have the same eigenvectors as well?)

(2) Let $\mathbf{A}$ and $\mathbf{B}$ be two $n \times n$ matrices. Then $\mathbf{A} \cdot \mathbf{B}$ and $\mathbf{B} \cdot \mathbf{A}$ have the same eigenvalues.
(Do they have the same eigenvectors as well?)

(3) If $\mathbf{x}$ is an eigenvector of $\mathbf{A}$ corresponding to eigenvalue λ, then $\mathbf{x}$ is also an eigenvector of $\mathbf{A}^k$ corresponding to eigenvalue λ^k.

(4) If $\mathbf{x}$ is an eigenvector of regular $\mathbf{A}$ corresponding to eigenvalue λ, then $\mathbf{x}$ is also an eigenvector of $\mathbf{A}^{-1}$ corresponding to eigenvalue λ^{-1}.

(5) The determinant of an $n \times n$ matrix $\mathbf{A}$ is equal to the product of all its eigenvalues: $\det(\mathbf{A}) = \prod_{i=1}^{n} \lambda_i$.

(6) The trace of an $n \times n$ matrix $\mathbf{A}$ (i.e., the sum of its diagonal entries) is equal to the sum of all its eigenvalues: $\det(\mathbf{A}) = \sum_{i=1}^{n} \lambda_i$.

11.9 Compute all leading principle minors of the symmetric matrices from Exercises 11.1, 11.2 and 11.3 and determine their definiteness.

11.10 Compute all principle minors of the symmetric matrices from Exercises 11.1, 11.2 and 11.3 and determine their definiteness.

11.11 Compute a symmetric 2×2 matrix $\mathbf{A}$ with eigenvalues $\lambda_1 = 1$ and $\lambda_2 = 3$ and corresponding eigenvectors $\mathbf{v}_1 = (1,1)'$ and $\mathbf{v}_2 = (-1,1)'$.

HINT: Use the Spectral Theorem. Recall that one needs an orthonormal basis.

11.12 Let $\mathbf{A}$ be the matrix in Problem 11.11. Compute $\sqrt{\mathbf{A}}$.

11.13 Let $\mathbf{A} = \begin{pmatrix} -1 & 3 \\ 3 & -1 \end{pmatrix}$. Compute $e^{\mathbf{A}}$.

— Problems

11.14 Prove Theorem 11.6.

HINT: Compare the characteristic polynomials of $\mathbf{A}$ and $\mathbf{A}'$.

11.15 Prove Theorem 11.7 by induction on power k.

HINT: Use Definition 11.1.

11.16 Prove Theorem 11.8.

HINT: Use Definition 11.1.

11.17 Prove Theorem 11.11.

HINT: Use a direct computation similar to the proof of Theorem 11.9 on p. 95.

11.18 Prove Theorem 11.12.

Show that the converse is false, i.e., if two matrices have the same spectrum then they need not be similar.

HINT: Compare the characteristic polynomials of $\mathbf{A}$ and $\mathbf{B}$. See Problem 6.11 for the converse statement.

11.19 Derive Theorem 11.9 immediately from Theorem 11.13.

11.20 Let $\mathbf{A}$ and $\mathbf{B}$ be similar $n \times n$ matrices with transformation matrix $\mathbf{U}$ such that $\mathbf{A} = \mathbf{U}^{-1}\mathbf{B}\mathbf{U}$. Show that $\mathbf{A}^k = \mathbf{U}^{-1}\mathbf{B}^k\mathbf{U}$ for every $k \in \mathbb{N}$.

HINT: Use induction.

11.21 Show that $q_{\mathbf{A}}$ is negative definite if and only if $q_{-\mathbf{A}}$ is positive definite.

11.22 Let $\mathbf{A}$ be a symmetric $n \times n$ matrix. Show that the positivity of all leading principle minors is a necessary condition for the positive definiteness of $\mathbf{A}$.

HINT: Compute $\mathbf{y}'\mathbf{A}_k\mathbf{y}$ where $\mathbf{A}_k$ be the k-th leading principle submatrix of $\mathbf{A}$ and $\mathbf{y} \in \mathbb{R}^k$. Notice that $\mathbf{y}$ can be extended to a vector $\mathbf{z} \in \mathbb{R}^n$ where $z_i = y_i$ if $1 \le i \le k$ and $z_i = 0$ for $k+1 \le i \le n$.

11.23 Prove Theorem 11.22.

HINT: Use Sylvester's criterion and Lemmata 11.17 and 11.21.

11.24 Derive a criterion for the positive or negative (semi) definiteness of a symmetric 2×2 matrix in terms of its determinant and trace.

11.25 Suppose that all leading principle minors of some matrix $\mathbf{A}$ are non-negative. Show that $\mathbf{A}$ need not be positive semidefinite.

HINT: Construct a 2×2 matrix where all leading principle minors are 0 and where the two eigenvalues are 0 and -1, respectively.

11.26 Let $\mathbf{v}_1, \ldots, \mathbf{v}_k \in \mathbb{R}^n$ and $\mathbf{V} = (\mathbf{v}_1, \ldots, \mathbf{v}_k)$. Then the **Gram matrix** of these vectors is defined as

$$\mathbf{G} = \mathbf{V}'\mathbf{V}\,.$$

Prove the following statements:

(a) $[G]_{ij} = \mathbf{v}_i'\mathbf{v}_j$.

(b) $\mathbf{G}$ is symmetric.

(c) $\mathbf{G}$ is positive semidefinite for all $\mathbf{X}$.

(d) $\mathbf{G}$ is regular if and only if the vectors $\mathbf{x}_1, \ldots, \mathbf{x}_k$ are linearly independent.

HINT: Use Definition 11.16 for statement (c). Use Lemma 6.25 for statement (d).

11.27 Let $\mathbf{v}_1, \ldots, \mathbf{v}_k \in \mathbb{R}^n$ be linearly independent vectors. Let $\mathbf{P}$ be the projection matrix for an orthogonal projection onto $\operatorname{span}(\mathbf{v}_1, \ldots, \mathbf{v}_k)$.

(a) Compute all eigenvalues of $\mathbf{P}$.

(b) Give bases for each of the eigenspace corresponding to non-zero eigenvalues.

HINT: Recall that $\mathbf{P}$ is idempotent, i.e., $\mathbf{P}^2 = \mathbf{P}$.

11.28 Let $\mathbf{U}$ be an orthogonal matrix. Show that all eigenvalues λ of $\mathbf{U}$ have absolute value 1, i.e., $|\lambda| = 1$.

HINT: Use Theorem 8.24.

11.29 Let $\mathbf{U}$ be an orthogonal 3×3 matrix. Show that there exists a vector $\mathbf{x}$ such that either $\mathbf{U}\mathbf{x} = \mathbf{x}$ or $\mathbf{U}\mathbf{x} = -\mathbf{x}$.

HINT: Use the result from Problem 11.28.

Part III

Analysis

12

Sequences and Series

What happens when we proceed ad infinitum*?*

12.1 Limits of Sequences

Definition 12.1

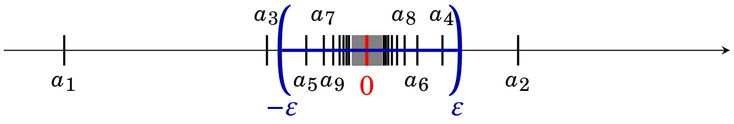
Sequence. A **sequence** $(x_n)_{n=1}^{\infty}$ of real numbers is an ordered list of real numbers. Formally it can be defined as a *function* that maps the natural numbers into $\mathbb{R}$. Number x_n is called the nth term of the sequence. We write (x_n) for short to denote a squence if there is no risk if confusion. Sequences can also be seen as vectors of infinite length.

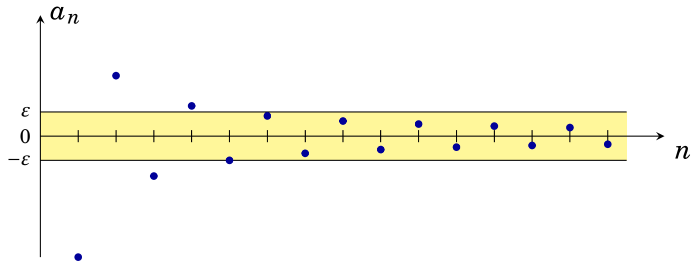
$x\colon \mathbb{N} \to \mathbb{R}, n \mapsto x_n$

Definition 12.2

Convergence and divergence. A sequence $(x_n)_{n=1}^{\infty}$ in $\mathbb{R}$ **converges** to a number x if for every $\varepsilon > 0$ there exists an index $N = N(\varepsilon)$ such that $|x_n - x| < \varepsilon$ for all $n \geq N$, or equivalently $x_n \in (x-\varepsilon, x+\varepsilon)$. The number x is then called the **limit** of the sequence. We write

$$x_n \to x \quad \text{as} \quad n \to \infty, \quad \text{or} \quad \lim_{n\to\infty} x_n = x\,.$$





A sequence that has a limit is called **convergent**. Otherwise it is called **divergent**.

Notice that the limit of a convergent sequence is uniquely determined, see Problem 12.5.

$$
\begin{aligned}
&\lim_{n\to\infty} c = c \text{ for all } c \in \mathbb{R}\\
&\lim_{n\to\infty} n^\alpha = \begin{cases} \infty, & \text{for } \alpha > 0,\\ 1, & \text{for } \alpha = 0,\\ 0, & \text{for } \alpha < 0. \end{cases}\\
&\lim_{n\to\infty} q^n = \begin{cases} \infty, & \text{for } q > 1,\\ 1, & \text{for } q = 1,\\ 0, & \text{for } -1 < q < 1,\\ \nexists, & \text{for } q \le -1. \end{cases}\\
&\lim_{n\to\infty} \tfrac{n^a}{q^n} = \begin{cases} 0, & \text{for } |q| > 1,\\ \infty, & \text{for } 0 < q < 1,\\ \nexists, & \text{for } -1 < q < 0, \end{cases} \qquad \text{for } |q| \notin \{0,1\}.\\
&\lim_{n\to\infty} \left(1+\tfrac{1}{n}\right)^n = e = 2.7182818\ldots
\end{aligned}
$$

Table 12.5

Limits of impo
sequences

Example 12.3

The sequences

$$(a_n)_{n=1}^\infty = \left(\frac{1}{2^n}\right)_{n=1}^\infty = \left(\frac{1}{2},\frac{1}{4},\frac{1}{8},\frac{1}{16},\ldots\right) \to 0$$

$$(b_n)_{n=1}^\infty = \left(\frac{n-1}{n+1}\right)_{n=1}^\infty = \left(0,\frac{1}{3},\frac{2}{4},\frac{3}{5},\frac{4}{6},\frac{5}{7},\ldots\right) \to 1$$

converge as $n \to \infty$, i.e.,

$$\lim_{n\to\infty}\left(\frac{1}{2^n}\right) = 0 \quad \text{and} \quad \lim_{n\to\infty}\left(\frac{n-1}{n+1}\right) = 1\,. \qquad \diamondsuit$$

Example 12.4

The sequence

$$(c_n)_{n=1}^\infty = \left((-1)^n\right)_{n=1}^\infty = (-1,1,-1,1,-1,1,\ldots)$$

$$(d_n)_{n=1}^\infty = \left(2^n\right)_{n=1}^\infty = (2,4,8,16,32,\ldots)$$

diverge. However, in the last example the sequence is increasing and not bounded from above. Thus we may write in *abuse of language*

$$\lim_{n\to\infty} 2^n = \infty\,. \qquad \diamondsuit$$

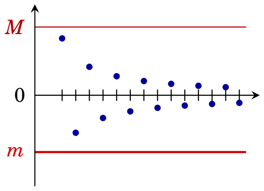
Computing limits can be a very challenging task. Thus we only look at a few examples. Table 12.5 lists limits of some important sequences. Notice that the limit of $\lim_{n\to\infty} \frac{n^a}{q^n}$ just says that in a product of a power sequence with an exponential sequence the latter dominates the limits.

We prove one of these limits in Lemma 12.12 below. For this purpose we need a few more notions.

Definition 12.6

Bounded sequence. A sequence $(x_n)_{n=1}^{\infty}$ of real numbers is called **bounded** if there exists an M such that

$$|x_n| \leq M \quad \text{for all } n \in \mathbb{N}.$$

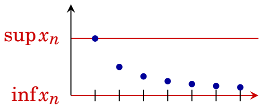


Two numbers m and M are called **lower** and **upper bound**, respectively, if

$$m \leq x_n \leq M, \quad \text{for all } n \in \mathbb{N}.$$

The greatest lower bound and the smallest upper bound are called **infimum** and **supremum** of the sequence, respectively, denoted by

$$\inf_{n\in\mathbb{N}} x_n \quad \text{and} \quad \sup_{n\in\mathbb{N}} x_n, \quad \text{respectively.}$$



Lemma 12.7

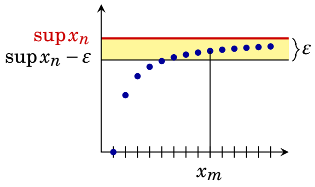
Notice that for a bounded sequence (x_n),

$$x_n \leq \sup_{k\in\mathbb{N}} x_k \quad \text{for all } n \in \mathbb{N}$$

and for all $\varepsilon > 0$, there exists an $m \in \mathbb{N}$ such that

$$x_m > \left(\sup_{k\in\mathbb{N}} x_k\right) - \varepsilon$$

since otherwise $\left(\sup_{k\in\mathbb{N}} x_k\right) - \varepsilon$ were a smaller upper bound, a contradiction to the definition of the supremum.

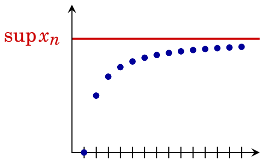


Example 12.8

Do not mix up supremum (or infimum) with the maximal (and minimal) value of a sequence. If a sequence (x_n) has a maximal value, then obviously $\max_{n\in\mathbb{N}} x_n = \sup_{n\in\mathbb{N}} x_n$. However, a maximal value need not exist. The sequence $\left(1 - \frac{1}{n}\right)_{n=1}^{\infty}$ is bounded and we have

$$\sup_{n\in\mathbb{N}} \left(1 - \frac{1}{n}\right) = 1\,.$$

However, 1 is never attained by this sequence and thus it does not have a maximum. ♢



Definition 12.9

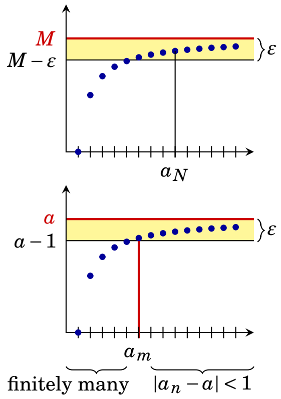
Monotone sequence. A sequence $(a_n)_{n=1}^{\infty}$ is called **monotone** if either $a_{n+1} \geq a_n$ (*increasing*) or $a_{n+1} \leq a_n$ (*decreasing*) for all $n \in \mathbb{N}$.

Lemma 12.10

Convergence of a monotone sequence. A monotone sequence $(a_n)_{n=1}^{\infty}$ is convergent if and only if it is bounded. We then find $\lim_{n\to\infty} a_n = \sup_{n\in\mathbb{N}} a_n$ if (a_n) is increasing, and $\lim_{n\to\infty} a_n = \inf_{n\in\mathbb{N}} a_n$ if (a_n) is decreasing.

PROOF IDEA. If (a_n) is increasing and bounded, then there is only a finite number of elements that are less than $\sup_{n\in\mathbb{N}} a_n - \varepsilon$.

If (a_n) is increasing and convergent, then there is only a finite number of elements greater than $\lim_{n\to\infty} a_n + \varepsilon$ or less than $\lim_{n\to\infty} a_n - \varepsilon$. These have a maximum and minimum value, respectively.



PROOF. We consider the case where (a_n) is increasing. Assume that (a_n) is bounded and $M = \sup_{n\in\mathbb{N}} a_n$. Then for every $\varepsilon > 0$, there exists an N such that $a_N > M - \varepsilon$ (Lemma 12.7). Since (a_n) is increasing we find $M \geq a_n > M - \varepsilon$ and thus $|a_n - M| < \varepsilon$ for all $n \geq N$. Consequently, $(a_n) \to M$ as $n \to \infty$.

Conversely, if (a_n) converges to a, then there is only a finite number of elements $a_1, \dots, a_m$ which do not satisfy $|a_n - a| < 1$. Thus $a_n < M = \max\{a+1, a_1, \dots, a_m\} < \infty$ for all $n \in \mathbb{N}$. Moreover, since (a_n) is increasing we also find $a_n \geq a_1$. Thus the sequence is bounded. The case where the sequence is decreasing follows completely analogously. □

Definition 12.11

For any $q \in \mathbb{R}$, the sequence $(q^n)_{n=0}^{\infty}$ is called a **geometric sequence**.

Lemma 12.12

Convergence of geometric sequence. $\lim\limits_{n\to\infty} q^n = 0$ for all $q \in (-1, 1)$.

PROOF. Observe that for $0 \leq q < 1$ we find $0 \leq q^n = q \cdot q^{n-1} \leq q^{n-1}$ for all $n \geq 2$ and hence q^n is decreasing and bounded from below. Hence it converges by Lemma 12.10 and $\lim\limits_{n\to\infty} q^n = \inf\limits_{n\geq 1} q^n$.

Now suppose that $m = \inf\limits_{n\geq 1} q^n > 0$ for some $0 < q < 1$ and let $\varepsilon = m(1/q - 1) > 0$. By Lemma 12.7 there exists a k such that $q^k < m + \varepsilon$. Then $q^{k+1} = q \cdot q^k < q(m + m(1/q - 1)) = m$, a contradiction. Hence $\lim\limits_{n\to\infty} q^n = 0$.

If $-1 < q < 0$, then $\lim\limits_{n\to\infty} |q^n| = 0$ and hence $\lim\limits_{n\to\infty} q^n = 0$ (Problem 12.6). □

Lemma 12.13

Divergence of geometric sequence. For $|q| > 1$ the geometric sequence diverges. Moreover, for $q > 1$ we find $\lim\limits_{n\to\infty} q^n = \infty$.

PROOF. Suppose $M = \sup\limits_{n\in\mathbb{N}} |q^n| < \infty$. Then $|(1/q)^n| \geq 1/M > 0$ for all $n \in \mathbb{N}$ and $M = \inf\limits_{n\in\mathbb{N}} |(1/q)^n| > 0$, a contradiction to Lemma 12.12, as $|1/q| < 1$. □

Limits of sequences with more complex terms can be reduced to the limits listed in Table 12.5 by means of the rules listed in Theorem 12.14 below. Notice that Rule (1) implies that taking the limit of a sequence is a linear operator on the set of all convergent sequences.

Theorem 12.14

Rules for limits. Let $(a_n)_{n=1}^{\infty}$ and $(b_n)_{n=1}^{\infty}$ be convergent sequences in $\mathbb{R}$ and $(c_n)_{n=1}^{\infty}$ be a bounded sequence in $\mathbb{R}$. Then

(1) $\lim\limits_{n\to\infty} (\alpha a_n + \beta b_n) = \alpha \lim\limits_{n\to\infty} a_n + \beta \lim\limits_{n\to\infty} b_n \qquad$ for all $\alpha, \beta \in \mathbb{R}$

(2) $\lim\limits_{n\to\infty} (a_n \cdot b_n) = \lim\limits_{n\to\infty} a_n \cdot \lim\limits_{n\to\infty} b_n$

(3) $\lim\limits_{n\to\infty} \dfrac{a_n}{b_n} = \dfrac{\lim_{n\to\infty} a_n}{\lim_{n\to\infty} b_n} \qquad$ (if $\lim\limits_{n\to\infty} b_n \neq 0$)

(4) $\lim\limits_{n\to\infty} a_n^r = \left(\lim\limits_{n\to\infty} a_n\right)^r$

(5) $\lim\limits_{n\to\infty} (a_n \cdot c_n) = 0 \qquad$ (if $\lim\limits_{n\to\infty} a_n = 0$)

For the proof of these (and other) properties of sequences the triangle inequality plays a prominent rôle.

Triangle inequality. For two real numbers a and b we find Lemma 12.15

$$|a+b| \le |a| + |b| .$$

PROOF. See Problem 12.4. □

Here we just prove Rule (1) from Theorem 12.14 (see also Problem 12.7). The other rules remain stated without proof.

Sum of covergent sequences. Let (a_n) and (b_n) be two sequences in $\mathbb{R}$ that converge to a and b, resp. Then Lemma 12.16

$$\lim_{n\to\infty} (a_n + b_n) = a + b .$$

PROOF IDEA. Use the triangle inequality for each term $(a_n+b_n)-(a+b)$.

PROOF. Let $\varepsilon > 0$ be arbitrary. Since both $(a_n) \to a$ and $(b_n) \to b$ there exists an $N = N(\varepsilon)$ such that $|a_n - a| < \varepsilon/2$ and $|b_n - b| < \varepsilon/2$ for all $n > N$. Then we find

$$\begin{aligned} |(a_n + b_n) - (a+b)| &= |(a_n - a) - (b_n - b)| \\ &\le |a_n - a| + |b_n - b| < \tfrac{\varepsilon}{2} + \tfrac{\varepsilon}{2} = \varepsilon \end{aligned}$$

for all $n > N$. But this means that $(a_n + b_n) \to (a+b)$, as claimed. □

The rules from Theorem 12.14 allow to reduce limits of composite terms to the limits listed in Table 12.5. Example 12.17

$$\lim_{n\to\infty}\left(2+\frac{3}{n^2}\right) = 2 + 3 \underbrace{\lim_{n\to\infty} n^{-2}}_{=0} = 2 + 3\cdot 0 = 2$$

$$\lim_{n\to\infty} (2^{-n}\cdot n^{-1}) = \lim_{n\to\infty} \frac{n^{-1}}{2^n} = 0$$

$$\lim_{n\to\infty} \frac{1+\frac{1}{n}}{2-\frac{3}{n^2}} = \frac{\lim_{n\to\infty}\left(1+\frac{1}{n}\right)}{\lim_{n\to\infty}\left(2-\frac{3}{n^2}\right)} = \frac{1}{2}$$

$$\lim_{n\to\infty} \underbrace{\sin(n)}_{\text{bounded}} \cdot \underbrace{\frac{1}{n^2}}_{\to 0} = 0$$ ♢

Exponential function. Theorem 12.14 allows to compute e^x as the limit of a sequence: Example 12.18

$$\begin{aligned} e^x &= \left(\lim_{m\to\infty}\left(1+\frac{1}{m}\right)^m\right)^x = \lim_{m\to\infty}\left(1+\frac{1}{m}\right)^{mx} = \lim_{n\to\infty}\left(1+\frac{1}{n/x}\right)^n \\ &= \lim_{n\to\infty}\left(1+\frac{x}{n}\right)^n \end{aligned}$$

where we have set $n = mx$. ♢

12.2 Series

Definition 12.19

Series. Let $(x_n)_{n=1}^{\infty}$ be a sequence of real numbers. Then the associated **series** is defined as the ordered formal sum

$$\sum_{n=1}^{\infty} x_n = x_1 + x_2 + x_3 + \dots .$$

The sequence of **partial sums** associated to series $\sum_{n=1}^{\infty} x_n$ is defined as

$$S_n = \sum_{i=1}^{n} x_i \qquad \text{for } n \in \mathbb{N}.$$

The series **converges** to a limit S if sequence $(S_n)_{n=1}^{\infty}$ converges to S, i.e.,

$$S = \sum_{i=1}^{\infty} x_i \quad \text{if and only if} \quad S = \lim_{n\to\infty} S_n = \lim_{n\to\infty} \sum_{i=1}^{n} x_i .$$

Otherwise, the series is called **divergent**.

We have already seen that a geometric sequence converges if $|q| < 1$, see Lemma 12.12. The same holds for the associated geometric series.

Lemma 12.20

Geometric series. The *geometric series* converges if and only if $|q| < 1$ and we find

$$\sum_{n=0}^{\infty} q^n = 1 + q + q^2 + q^3 + \cdots = \frac{1}{1-q} .$$

PROOF IDEA. We first find a closed form for the terms of the geometric series and then compute the limit.

PROOF. We first show that for any $n \geq 0$,

$$S_n = \sum_{k=0}^{n} q^k = \frac{1-q^{n+1}}{1-q} .$$

In fact,

$$\begin{aligned} S_n(1-q) = S_n - qS_n &= \sum_{k=0}^{n} q^k - q\sum_{k=0}^{n} q^k = \sum_{k=0}^{n} q^k - \sum_{k=0}^{n} q^{k+1} \\ &= \sum_{k=0}^{n} q^k - \sum_{k=1}^{n+1} q^k = q^0 - q^{n+1} = 1 - q^{n+1} \end{aligned}$$

and thus the result follows. Now by the rules for limits of sequences we find by Lemma 12.12 $\lim\limits_{n\to\infty} \sum_{k=0}^{n} q^n = \lim\limits_{n\to\infty} \frac{1-q^{n+1}}{1-q} = \frac{1}{1-q}$ if $|q| < 1$. Conversely, if $|q| > 1$, the sequence diverges by Lemma 12.13. If $q = 1$, the we trivially have $\sum_{n=0}^{\infty} 1 = \infty$. For $q = -1$ the sequence of partial sums is given by $S_n = \sum_{k=0}^{n} (-1)^k = 1 + (-1)^n$ which obviously does not converge. This completes the proof. □

Harmonic series. The so called *harmonic series* diverges, Lemma 12.21

$$\sum_{n=1}^{\infty} \frac{1}{n} = 1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \cdots = \infty\,.$$

PROOF IDEA. We construct a new series which is component-wise smaller than or equal to the harmonic series. This series is then transformed by adding some its terms into a series with constant terms which is obviously divergent.

PROOF. We find

$$\begin{aligned}\sum_{n=1}^{\infty} \frac{1}{n} &= 1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \frac{1}{5} + \frac{1}{6} + \frac{1}{7} + \frac{1}{8} + \frac{1}{9} + \cdots + \frac{1}{16} + \frac{1}{17} + \ldots \\ &> 1 + \frac{1}{2} + \frac{1}{4} + \frac{1}{4} + \frac{1}{8} + \frac{1}{8} + \frac{1}{8} + \frac{1}{8} + \frac{1}{16} + \cdots + \frac{1}{16} + \frac{1}{32} + \ldots \\ &= 1 + \frac{1}{2} + \left(\frac{1}{4} + \frac{1}{4}\right) + \left(\frac{1}{8} + \frac{1}{8} + \frac{1}{8} + \frac{1}{8}\right) + \left(\frac{1}{16} + \cdots + \frac{1}{16}\right) + \left(\frac{1}{32} + \ldots \right.\\ &= 1 + \frac{1}{2} + \left(\frac{1}{2}\right) + \left(\frac{1}{2}\right) + \left(\frac{1}{2}\right) + \left(\frac{1}{2}\right) + \ldots \\ &= \infty\,.\end{aligned}$$

More precisely, we have $\sum_{n=1}^{2^k} \frac{1}{n} > 1 + \frac{k}{2} \to \infty$ as $k \to \infty$. □

The trick from the above proof is called the **comparison test** as we compare our series with a divergent series. Analogously one also may compare the sequence with a convergent one.

Comparison test. Let $\sum_{n=1}^{\infty} a_n$ and $\sum_{n=1}^{\infty} b_n$ be two series with $0 \le a_n \le b_n$ for all $n \in \mathbb{N}$. Lemma 12.22

(a) If $\sum_{n=1}^{\infty} b_n$ converges, then $\sum_{n=1}^{\infty} a_n$ converges.

(b) If $\sum_{n=1}^{\infty} a_n$ diverges, then $\sum_{n=1}^{\infty} b_n$ diverges.

PROOF. (a) Suppose that $B = \sum_{k=1}^{\infty} b_k < \infty$ exists. Then by our assumptions $0 \le \sum_{k=1}^{n} a_k \le \sum_{k=1}^{n} b_k \le B$ for all $n \in \mathbb{N}$. Hence $\sum_{k=1}^{n} a_k$ is increasing and bounded and thus the series converges by Lemma 12.10.

(b) On the other hand, if $\sum_{k=1}^{\infty} a_k$ diverges, then for every M there exists an N such that $M \le \sum_{k=1}^{n} a_k \le \sum_{k=1}^{n} b_k$ for all $n \ge N$. Hence $\sum_{k=1}^{\infty} b_k$ diverges, too. □

Such tests are very important as it allows to verify whether a series converges or diverges by comparing it to a series where the answer is much simpler. However, it does not provide a limit when (b_n) converges (albeit it provides an upper bound for the limit). Nevertheless, the proof of existence is also of great importance. The following example demonstrates that using expressions in a naïve way without checking their existence may result in contradictions.

Example 12.23

Grandi's series. Consider the following series that has been extensively discussed during the 18th century. What is the value of

$$S = \sum_{n=0}^{\infty} (-1)^n = 1 - 1 + 1 - 1 + 1 - 1 + \ldots ?$$

One might argue in the following way:

$$\begin{aligned} 1 - S &= 1 - (1 - 1 + 1 - 1 + 1 - 1 + \ldots) = 1 - 1 + 1 - 1 + 1 - 1 + 1 - \ldots \\ &= 1 - 1 + 1 - 1 + 1 - \ldots = S \end{aligned}$$

and hence $2S = 1$ and $S = \frac{1}{2}$. Notice that this series is just a special case of the geometric series with $q = -1$. Thus we get the same result if we misleadingly use the formula from Lemma 12.20.

However, we also may proceed in a different way. By putting parentheses we obtain

$$\begin{aligned} &S = (1-1) + (1-1) + (1-1) + \ldots = 0 + 0 + 0 + \ldots = 0\,, \quad \text{and} \\ &S = 1 + (-1+1) + (-1+1) + (-1+1) + \ldots = 1 + 0 + 0 + 0 + \ldots = 1\,. \end{aligned}$$

Combining these three computations gives

$$S = \tfrac{1}{2} = 0 = 1$$

which obviously is not what we expect from real numbers. The error in all these computation is that the expression S cannot be treated like a number since the series diverges. ♢

If we are given a convergent sequence $(a_n)_{n=1}^{\infty}$ then the sequence of its absolute values also converges (Problem 12.6). The converse, however, may not hold. For the associated series we have an opposite result.

Lemma 12.24

Let $\sum_{n=1}^{\infty} a_n$ be some series. If $\sum_{n=1}^{\infty} |a_n|$ converges, then $\sum_{n=1}^{\infty} a_n$ also converges.

PROOF IDEA. We split the series into a positive and a negative part.

PROOF. Let $\mathscr{P} = \{n \in \mathbb{N} \colon a_n \geq 0\}$ and $\mathscr{N} = \{n \in \mathbb{N} \colon a_n < 0\}$. Then

$$m_+ = \sum_{n \in \mathscr{P}} |a_n| \leq \sum_{n=1}^{\infty} |a_n| < \infty \quad \text{and} \quad m_- = \sum_{n \in \mathscr{N}} |a_n| \leq \sum_{n=1}^{\infty} |a_n| < \infty$$

and therefore

$$\sum_{n=1}^{\infty} a_n = \sum_{n \in \mathscr{P}} |a_n| - \sum_{n \in \mathscr{N}} |a_n| = m_+ - m_-$$

exists. □

Notice that the converse does not hold. If series $\sum_{n=1}^{\infty} a_n$ converges then $\sum_{n=1}^{\infty} |a_n|$ may diverge.

Example 12.25

It can be shown that the **alternating harmonic series**

$$\sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n} = 1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \frac{1}{5} - \ldots = \ln 2$$

converges, whereas we already have seen in Lemma 12.21 that the harmonic series $\sum_{n=1}^{\infty} \left|\frac{(-1)^{n+1}}{n}\right| = \sum_{n=1}^{\infty} \frac{1}{n}$ does not not. ♢

Definition 12.26

A series $\sum_{n=1}^{\infty} a_n$ is called **absolutely convergent** if $\sum_{n=1}^{\infty} |a_n|$ converges.

Lemma 12.27

Ratio test. A series $\sum_{n=1}^{\infty} a_n$ converges if there exists a $q < 1$ and an $N < \infty$ such that

$$\left|\frac{a_{n+1}}{a_n}\right| \leq q < 1 \quad \text{for all } n \geq N.$$

Similarly, if there exists an $r > 1$ and an $N < \infty$ such that

$$\left|\frac{a_{n+1}}{a_n}\right| \geq r > 1 \quad \text{for all } n \geq N$$

then the series diverges.

PROOF IDEA. We compare the series with a geometric series and apply the comparison test.

PROOF. For the first statement observe that $|a_{n+1}| < |a_n|\,q$ implies $|a_{N+k}| < |a_N|\,q^k$. Hence

$$\sum_{n=1}^{\infty} |a_n| = \sum_{n=1}^{N} |a_n| + \sum_{k=1}^{\infty} |a_{N+k}| < \sum_{n=1}^{N} |a_n| + |a_N| \sum_{k=1}^{\infty} q^k < \infty$$

where the two inequalities follows by Lemmata 12.22 and 12.20. Thus $\sum_{n=1}^{\infty} a_n$ converges by Lemma 12.24. The second statement follows similarly but requires more technical details and is thus omitted. □

There exist different variants of this test. We give a convenient version for a special case.

Lemma 12.28

Ratio test. Let $\sum_{n=1}^{\infty} a_n$ be a series where $\lim_{n\to\infty} \left|\frac{a_{n+1}}{a_n}\right|$ exists. Then $\sum_{n=1}^{\infty} a_n$ converges if

$$\lim_{n\to\infty} \left|\frac{a_{n+1}}{a_n}\right| < 1 .$$

It diverges if

$$\lim_{n\to\infty} \left|\frac{a_{n+1}}{a_n}\right| > 1 .$$

PROOF. Assume that $L = \lim_{n\to\infty} \left|\frac{a_{n+1}}{a_n}\right|$ exists and $L < 1$. Then there exists an N such that $\left|\frac{a_{n+1}}{a_n}\right| < q = 1 - \frac{1}{2}(1-L) < 1$ for all $n \geq N$. Thus the series converges by Lemma 12.27. The proof for the second statement is completely analogous. □

— Exercises

12.1 Compute the following limits:

(a) $\lim_{n\to\infty}\left(7+\left(\frac{1}{2}\right)^n\right)$ (b) $\lim_{n\to\infty}\frac{2n^3-6n^2+3n-1}{7n^3-16}$

(c) $\lim_{n\to\infty}\frac{n \bmod 10}{(-2)^n}$ (d) $\lim_{n\to\infty}\frac{n^2+1}{n+1}$

(e) $\lim_{n\to\infty}\left(n^2-(-1)^n n^3\right)$ (f) $\lim_{n\to\infty}\left(\frac{7n}{2n-1}-\frac{4n^2-1}{5-3n^2}\right)$

12.2 Compute the limits of sequence $(a_n)_{n=1}^{\infty}$ with the following terms:

(a) $a_n=(-1)^n\left(1+\frac{1}{n}\right)$ (b) $a_n=\frac{n}{(n+1)^2}$

(c) $a_n=\left(1+\frac{2}{n}\right)^n$ (d) $a_n=\left(1-\frac{2}{n}\right)^n$

(e) $a_n=\frac{1}{\sqrt{n}}$ (f) $a_n=\frac{n}{n+1}+\frac{1}{\sqrt{n}}$

(g) $a_n=\frac{n}{n+1}+\sqrt{n}$ (h) $a_n=\frac{4+\sqrt{n}}{n}$

12.3 Compute the following limits:

(a) $\lim_{n\to\infty}\left(1+\frac{1}{n}\right)^{nx}$ (b) $\lim_{n\to\infty}\left(1+\frac{x}{n}\right)^n$ (c) $\lim_{n\to\infty}\left(1+\frac{1}{nx}\right)^n$

— Problems

12.4 Prove the triangle inequality in Lemma 12.15.

HINT: Look at all possible cases where $a\ge 0$ or $a<0$ and $b\ge 0$ and $b<0$.

12.5 Let (a_n) be a convergent sequence. Show by means of the triangle inequality (Lemma 12.15) that its limit is uniquely defined.

HINT: Assume that two limits a and b exist and show that $|a-b|=0$.

HINT: Use inequality $\big||a|-|b|\big|\le|a-b|$.

12.6 Let (a_n) be a convergent sequence with $\lim_{n\to\infty}a_n=a$. Show that

$$\lim_{n\to\infty}|a_n|=|a|\,.$$

State and disprove the converse statement.

12.7 Let (a_n) be a sequence in $\mathbb{R}$ that converge to a and $c\in\mathbb{R}$. Show that

$$\lim_{n\to\infty}c\,a_n=c\,a\,.$$

12.8 Let (a_n) be a sequence in $\mathbb{R}$ that converge to 0 and (c_n) be a bounded sequence. Show that

$$\lim_{n\to\infty}c_n\,a_n=0\,.$$

12.9 Let (a_n) be a convergent sequence with $a_n \geq 0$. Show that

$$\lim_{n\to\infty} a_n \geq 0\,.$$

Disprove that $\lim_{n\to\infty} a_n > 0$ when all elements of this convergent sequence are positive, i.e., $a_n > 0$ for all $n \in \mathbb{N}$.

12.10 When we inspect the second part of the proof of Lemma 12.10 we find that monotonicity of sequence (a_n) is not required. Show that every convergent sequence (a_n) is bounded.

Also disprove the converse claim that every bounded sequence is convergent.

12.11 Compute $\displaystyle\sum_{k=1}^{\infty} q^n$.

12.12 Show that for any $a \in \mathbb{R}$,

$$\lim_{n\to\infty} \frac{a^n}{n!} = 0\,.$$

HINT: There exists an $N > |a|$.

12.13 **Cauchy's covergence criterion.** A sequence (a_n) in $\mathbb{R}$ is called a *Cauchy sequence* if for every $\varepsilon > 0$ there exists a number N such that $|a_n - a_m| < \varepsilon$ for all $n, m > N$.
Show: If a sequence (a_n) converges, then it is a Cauchy sequence.

HINT: Use the triangle inequality.

(Remark: The converse also holds. If (a_n) is a Cauchy sequence, then it converges.)

12.14 Show that $\displaystyle\sum_{n=1}^{\infty} \frac{1}{n!}$ converges.

HINT: Use the ratio test.

12.15 Someone wants to show the (false!) "theorem":

If $\sum_{n=1}^{\infty} a_n$ converges, then $\sum_{n=1}^{\infty} |a_n|$ also converges.

He argues as follows:

Let $\mathcal{P} = \{n \in \mathbb{N}\colon a_n \geq 0\}$ and $\mathcal{N} = \{n \in \mathbb{N}\colon a_n < 0\}$. Then

$$\sum_{n=1}^{\infty} a_n = \sum_{n\in\mathcal{P}} a_n + \sum_{n\in\mathcal{N}} a_n = \sum_{n\in\mathcal{P}} |a_n| - \sum_{n\in\mathcal{N}} |a_n| < \infty$$

and thus both $m_+ = \sum_{n\in\mathcal{P}} |a_n| < \infty$ and $m_- = \sum_{n\in\mathcal{N}} |a_n| < \infty$. Therefore

$$\sum_{n=1}^{\infty} |a_n| = \sum_{n\in\mathcal{P}} |a_n| + \sum_{n\in\mathcal{N}} |a_n| = m_+ + m_- < \infty$$

exists.

13 Topology

We need the concepts of neighborhood *and* boundary.

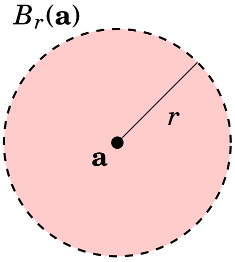
The fundamental idea in analysis can be visualized as *roaming in foggy weather*. We explore a function *locally* around some point by making tiny steps in all directions. However, we then need some conditions that ensure that we do not run against an edge or fall out of our function's world (i.e., its domain). Thus we introduce the concept of an *open neighborhood*.

13.1 Open Neighborhood

Definition 13.1

Interior, exterior and boundary points. Recall that for any point $\mathbf{x} \in \mathbb{R}^n$ the **Euclidean norm** $\|\mathbf{x}\|$ is defined as

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

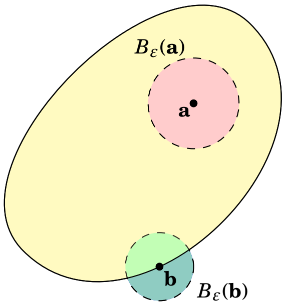


The **Euclidean distance** $d(\mathbf{x},\mathbf{y})$ between any two points $\mathbf{x},\mathbf{y} \in \mathbb{R}^n$ is given as

$$d(\mathbf{x},\mathbf{y}) = \|\mathbf{x}-\mathbf{y}\| = \sqrt{(\mathbf{x}-\mathbf{y})'(\mathbf{x}-\mathbf{y})}\,.$$

These terms allow us to get a notion of points that are "nearby" some point $\mathbf{x}$. The set

$$B_r(\mathbf{a}) = \{\mathbf{x} \in \mathbb{R}^n : d(\mathbf{x},\mathbf{a}) < r\}$$



is called the **open ball** around $\mathbf{a}$ with radius r (> 0). A point $\mathbf{a} \in D$ is called an **interior point** of a set $D \subseteq \mathbb{R}^n$ if there exists an open ball centered at $\mathbf{a}$ which lies inside D, i.e., there exists an $\varepsilon > 0$ such that $B_\varepsilon(\mathbf{a}) \subseteq D$. An immediate consequence of this definition is that we can move away from some interior point $\mathbf{a}$ in any direction without leaving D provided that the step size is sufficiently small. Notice that every set contains all its interior points.

A point $\mathbf{b} \in \mathbb{R}^n$ is called a **boundary point** of a set $D \subseteq \mathbb{R}^n$ if every open ball centered at $\mathbf{b}$ intersects both D and its complement $D^c = \mathbb{R}^n \setminus D$. Notice that a boundary point $\mathbf{b}$ needs not be an element of D.

A point $\mathbf{x} \in \mathbb{R}^n$ is called an **exterior point** of a set $D \subseteq \mathbb{R}^n$ if it is an interior point of its complement $\mathbb{R}^n \setminus D$.

Definition 13.2

A set $D \subseteq \mathbb{R}^n$ is called an **open neighborhood** of $\mathbf{a}$ if $\mathbf{a}$ is an interior point of D, i.e., if D contains some open ball centered at $\mathbf{a}$.

Definition 13.3

A set $D \subseteq \mathbb{R}^n$ is called **open** if all its members are interior points of D, i.e., if for each $\mathbf{a} \in D$, D contains some open ball centered at $\mathbf{a}$ (that is, have an open neighborhood in D). On the real line $\mathbb{R}$, the simplest example of an open set is an open interval $(a,b) = \{x \in \mathbb{R}\colon a < x < b\}$.

A set $D \subseteq \mathbb{R}^n$ is called **closed** if it contains all its boundary points. On the real line $\mathbb{R}$, the simplest example of a closed set is a closed interval $[a,b] = \{x \in \mathbb{R}\colon a \leq x \leq b\}$.

Example 13.4

Show that $H = \{(x,y) \in \mathbb{R}^2\colon x > 0\}$ is an open set.

SOLUTION. Take any point (x_0, y_0) in H and set $\varepsilon = x_0/2$. We claim that $B = B_\varepsilon(x_0,y_0)$ is contained in H. Let $(x,y) \in B$. Then $\varepsilon > \|(x,y)-(x_0,y_0)\| = \sqrt{(x-x_0)^2+(y-y_0)^2} \geq \sqrt{(x-x_0)^2} = |x-x_0|$. Consequently, $x > x_0 - \varepsilon = x_0 - \frac{x_0}{2} = \frac{x_0}{2} > 0$ and thus $(x,y) \in H$ as claimed. □

Lemma 13.5

A set $D \subseteq \mathbb{R}^n$ is closed if and only if its complement D^c is open.

PROOF. See Problem 13.6.

Theorem 13.6

Properties of open sets.

(1) The empty set $\emptyset$ and the whole space $\mathbb{R}^n$ are both open.

(2) Arbitrary unions of open sets are open.

(3) The intersection of finitely many open sets is open.

PROOF IDEA. (1) Every ball of centered at any point is entirely in $\mathbb{R}^n$. Thus $\mathbb{R}^n$ is open. For the empty set observe that it does not contain any element that violates the condition for *"interior point"*.

(2) Every open ball $B_\varepsilon(\mathbf{x})$ remains contained in a set D if we add points to D. Thus interior points of D remain interior points in any superset of D.

(3) If $\mathbf{x}$ is an interior point of open sets $D_1,\dots,D_m$, then there exist open balls $B_i(\mathbf{x}) \subseteq D_i$ centered at $\mathbf{x}$. Since they are only finitely many, there is a smallest one which is thus entirely contained in the intersection of all D_i's.

PROOF. (1) Every ball $B_\varepsilon(\mathbf{a}) \subseteq \mathbb{R}^n$ and thus $\mathbb{R}^n$ is open. All members of the empty set $\emptyset$ are inside balls that are contained entirely in $\emptyset$. Hence $\emptyset$ is open.

(2) Let $\{D_i\}_{i\in I}$ be an arbitrary family of open sets in $\mathbb{R}^n$, and let $D = \bigcup_{i\in I} D_i$ be the union of all these. For each $\mathbf{x} \in D$ there is at least one $i \in I$ such that $\mathbf{x} \in D_i$. Since D_i is open, there exists an open ball $B_\varepsilon(\mathbf{x}) \subseteq D_i \subseteq D$. Hence $\mathbf{x}$ is an interior point of D.

(3) Let $\{D_1, D_2, \dots, D_m\}$ be a finite collection of open sets in $\mathbb{R}^n$, and let $D = \bigcap_{i=1}^m D_i$ be the intersection of all these sets. Let $\mathbf{x}$ be any point in D. Since all D_i are open there exist open balls $B_i = B_{\varepsilon_i}(\mathbf{x}) \subseteq D_i$ with center $\mathbf{x}$. Let ε be the smallest of all radii ε_i. Then $\mathbf{x} \in B_\varepsilon(\mathbf{x}) = \bigcap_{i=1}^m B_i \subseteq \bigcap_{i=1}^m D_i = D$ and thus D is open. □

ε is the minimum of a finite set of numbers.

The intersection of an infinite number of open sets needs not be open, see Problem 13.10.

Similarly by De Morgan's law we find the following properties of closed sets, see Problem 13.11.

Properties of closed sets.

Theorem 13.7

(1) The empty set $\emptyset$ and the whole space $\mathbb{R}^n$ are both closed.

(2) Arbitrary intersections of closed sets are closed.

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

Each $\mathbf{y} \in \mathbb{R}^n$ is either an interior, an exterior or a boundary point of some set $D \subseteq \mathbb{R}^n$. As a consequence there is a corresponding partition of $\mathbb{R}^n$ into three mutually disjoint sets.

Definition 13.8

For a set $D \subseteq \mathbb{R}^n$, the set of all interior points of D is called the **interior** of D. It is denoted by D° or $\mathrm{int}(D)$.

The set of all boundary points of a set D is called the **boundary** of D. It is denoted by ∂D or $\mathrm{bd}(D)$.

The union $D \cup \partial D$ is called the **closure** of D. It is denoted by $\overline{D}$ or $\mathrm{cl}(D)$.

Definition 13.9

A point $\mathbf{a}$ is called an **accumulation point** of a set D if every open neighborhood of $\mathbf{a}$ (i.e., open ball $B_\varepsilon(\mathbf{a})$) has non-empty intersection with D (i.e., $D \cap B_\varepsilon(\mathbf{a}) \neq \emptyset$). Notice that $\mathbf{a}$ need not be an element of D.

Lemma 13.10

A set D is closed if and only if D contains all its accumulation points.

PROOF. See Problem 13.12.

13.2 Convergence

Definition 13.11

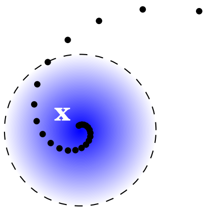
A **sequence** $(\mathbf{x}_k)_{k=1}^\infty$ in $\mathbb{R}^n$ is a *function* that maps the natural numbers into $\mathbb{R}^n$. A point $\mathbf{x}_k$ is called the kth term of the sequence.
Sequences can also be seen as vectors of infinite length.

$\mathbf{x}\colon \mathbb{N} \to \mathbb{R}^n, k \mapsto \mathbf{x}_k$

Recall that a sequence (x_k) in $\mathbb{R}$ converges to a number x if for every $\varepsilon > 0$ there exists an index N such that $|x_k - x| < \varepsilon$ for all $k > N$. This can be easily generalized.

Definition 13.12

Convergence and divergence. A sequence $(\mathbf{x}_k)$ in $\mathbb{R}^n$ **converges** to a point $\mathbf{x}$ if for every $\varepsilon > 0$ there exists an index $N = N(\varepsilon)$ such that $\mathbf{x}_k \in B_\varepsilon(\mathbf{x})$, i.e., $\|\mathbf{x}_k - \mathbf{x}\| < \varepsilon$, for all $k > N$.



Equivalently, $(\mathbf{x}_k)$ converges to $\mathbf{x}$ if $d(\mathbf{x}_k, \mathbf{x}) \to 0$ as $k \to \infty$. The point $\mathbf{x}$ is then called the **limit** of the sequence. We write

$$\mathbf{x}_k \to \mathbf{x} \quad \text{as} \quad k \to \infty, \quad \text{or} \quad \lim_{k\to\infty} \mathbf{x}_k = \mathbf{x}\,.$$

Notice, that the limit of a convergent sequence is uniquely determined. A sequence that is not **convergent** is called **divergent**.

We can look at each of the component sequences in order to determine whether a sequence of points does converge or not. Thus the following theorem allows us to reduce results for convergent sequences in $\mathbb{R}^n$ to corresponding results for convergent sequences of real numbers.

Theorem 13.13

Convergence of each component. A sequence $(\mathbf{x}_k)$ in $\mathbb{R}^n$ converges to the vector $\mathbf{x}$ in $\mathbb{R}^n$ if and only if for each $j = 1, \dots, n$, the real number sequence $\left(x_k^{(j)}\right)_{k=1}^{\infty}$, consisting of the jth component of each vector $\mathbf{x}_k$, converges to $x^{(j)}$, the jth component of $\mathbf{x}$.

PROOF IDEA. For the proof of the necessity of the condition we use the fact that $\max_i |x_i| \leq \|\mathbf{x}\|$. For the sufficiency observe that $\|\mathbf{x}\|^2 \leq n \max_i |x_i|^2$.

PROOF. Assume that $\mathbf{x}_k \to \mathbf{x}$. Then for every $\varepsilon > 0$ there exists an N such that $\|\mathbf{x}_k - \mathbf{x}\| < \varepsilon$ for all $k > N$. Consequently, for each j one has $|x_k^{(j)} - x^{(j)}| \leq \|\mathbf{x}_k - \mathbf{x}\| < \varepsilon$ for all $k > N$, that is, $x_k^{(j)} \to x^{(j)}$.

Now assume that $x_k^{(j)} \to x^{(j)}$ for each j. Then given any $\varepsilon > 0$, for each j there exists a number N_j such that $|x_k^{(j)} - x^{(j)}| \leq \varepsilon/\sqrt{n}$ for all $k > N_j$. It follows that

$$\|\mathbf{x}_k - \mathbf{x}\| = \sqrt{\sum_{i=1}^{n} |x_k^{(i)} - x^{(i)}|^2} < \sqrt{\sum_{i=1}^{n} \varepsilon^2/n} = \sqrt{\varepsilon^2} = \varepsilon$$

for all $k > \max\{N_1, \dots, N_n\}$. Therefore $\mathbf{x}_k \to \mathbf{x}$ as $k \to \infty$. □

We will see in Section 13.3 below that this theorem is just a consequence of the fact that Euclidean norm and supremum norm are equivalent.

The next theorem gives a criterion for convergent sequences. The proof of the necessary condition demonstrates a simple but quite powerful technique.

Definition 13.14

A sequence $(\mathbf{x}_k)$ in $\mathbb{R}^n$ is called a **Cauchy sequence** if for every $\varepsilon > 0$ there exists a number N such that $\|\mathbf{x}_k - \mathbf{x}_m\| < \varepsilon$ for all $k, m > N$.

Theorem 13.15

Cauchy's covergence criterion. A sequence $(\mathbf{x}_k)$ in $\mathbb{R}^n$ is convergent if and only if it is a Cauchy sequence.

PROOF IDEA. For the necessity of the Cauchy sequence we use the trivial equality $\|\mathbf{x}_k - \mathbf{x}_m\| = \|(\mathbf{x}_k - \mathbf{x}) + (\mathbf{x} - \mathbf{x}_m)\|$ and apply the triangle inequality for norms.

For the sufficiency assume that $\|\mathbf{x}_k - \mathbf{x}_m\| \leq \frac{1}{j}$ for all $m > k \geq N_j$ and construct closed balls $\overline{B}_{1/j}(\mathbf{x}_{N_j})$ for all $j \in \mathbb{N}$. Their intersection $\bigcap_{j=1}^{\infty} \overline{B}_{1/j}(\mathbf{x}_{N_j})$ is closed by Theorem 13.7 and is either a single point or the empty set. The latter can be excluded by an axiom of the real numbers.

PROOF. Assume that $(\mathbf{x}_k)$ converges to $\mathbf{x}$. Then there exists a number N such that $\|\mathbf{x}_k - \mathbf{x}\| < \varepsilon/2$ for all $k > N$. Hence by the triangle inequality we find

$$\|\mathbf{x}_k - \mathbf{x}_m\| = \|(\mathbf{x}_k - \mathbf{x}) + (\mathbf{x} - \mathbf{x}_m)\| \leq \|\mathbf{x}_k - \mathbf{x}\| + \|\mathbf{x} - \mathbf{x}_m\| < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon$$

for all $k, m > N$. Thus $(\mathbf{x}_k)$ is a Cauchy sequence.

For the converse assume that for all $\varepsilon = 1/j$ there exists an N_j such that $\|\mathbf{x}_k - \mathbf{x}_m\| \leq \frac{1}{j}$ for all $m > k \geq N_j$, i.e., $\mathbf{x}_m \in \overline{B}_{1/j}(\mathbf{x}_{N_j})$ for all $m > N_j$. Let $D_j = \bigcap_{i=1}^{j} \overline{B}_{1/i}(\mathbf{x}_{N_i})$. Then $\mathbf{x}_m \in D_j$ for all $m > N_j$ and thus $D_j \neq \emptyset$ for all $j \in \mathbb{N}$. Moreover, the diameter of $D_j \leq 2/j \to 0$ for $j \to \infty$. By Theorem 13.7, $D = \bigcap_{i=1}^{\infty} \overline{B}_{1/i}(\mathbf{x}_{N_i})$ is closed. Therefore, either $D = \{\mathbf{a}\}$ consists of a single point or $D = \emptyset$. The latter can be excluded by a fundamental property (i.e., an axiom) of the real numbers. (However, this step is out of the scope of this course.) □

The next theorem is another example of an application of the triangle inequality.

Theorem 13.16

Sum of covergent sequences. Let $(\mathbf{x}_k)$ and $(\mathbf{y}_k)$ be two sequences in $\mathbb{R}^n$ that converge to $\mathbf{x}$ and $\mathbf{y}$, resp. Then

$$\lim_{k \to \infty} (\mathbf{x}_k + \mathbf{y}_k) = \mathbf{x} + \mathbf{y}\,.$$

PROOF IDEA. Use the triangle inequality for each term $\|(\mathbf{x}_k + \mathbf{y}_k) - (\mathbf{x} + \mathbf{y})\| = \|(\mathbf{x}_k - \mathbf{x}) + (\mathbf{y}_k - \mathbf{y})\|$.

PROOF. Let $\varepsilon > 0$ be arbitrary. Since $(\mathbf{x}_k)$ is convergent, there exists a number N_x such that $\|\mathbf{x}_k - \mathbf{x}\| < \varepsilon/2$ for all $k > N_x$. Analogously there exists a number N_y such that $\|\mathbf{y}_k - \mathbf{y}\| < \varepsilon/2$ for all $k > N_y$. Let N be the greater of the two numbers N_x and N_y. Then by the triangle inequality we find for $k > N$,

$$\begin{aligned}\|(\mathbf{x}_k + \mathbf{y}_k) - (\mathbf{x} + \mathbf{y})\| &= \|(\mathbf{x}_k - \mathbf{x}) + (\mathbf{y}_k - \mathbf{y})\| \\ &\le \|\mathbf{x}_k - \mathbf{x}\| + \|\mathbf{y}_k - \mathbf{y}\| < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon\,.\end{aligned}$$

But this means that $(\mathbf{x}_k + \mathbf{y}_k) \to (\mathbf{x} + \mathbf{y})$, as claimed. □

We can use convergent sequences to characterize closed sets.

Theorem 13.17

Closure and convergence. A set $D \subseteq \mathbb{R}^n$ is closed if and only if every convergent sequence of points in D has its limit in D.

PROOF IDEA. For any sequence in D with limit $\mathbf{x}$ every ball $B_\varepsilon(\mathbf{x})$ contains almost all elements of the sequence. Hence it belongs to the closure of D. So if D is closed then $\mathbf{x} \in D$.

Conversely, if $\mathbf{x} \in \text{cl}(D)$ we can select points $\mathbf{x}_k \in B_{1/k}(\mathbf{x}) \cap D$. Then sequence $(\mathbf{x}_k) \to \mathbf{x}$ converges. If we assume that every convergent sequence of points in D has its limit in D it follows that $\mathbf{x} \in D$ and hence D is closed.

PROOF. Assume that D is closed. Let $(\mathbf{x}_k)$ be a convergent sequence with limit $\mathbf{x}$ such that $\mathbf{x}_k \in D$ for all k. Hence for all $\varepsilon > 0$ there exists an N such that $\mathbf{x}_k \in B_\varepsilon(\mathbf{x})$ for all $k > N$. Therefore $B_\varepsilon(\mathbf{x}) \cap D \neq \emptyset$ and $\mathbf{x}$ belongs to the closure of D. Since D is closed, limit $\mathbf{x}$ also belongs to D.

Conversely, assume that every convergent sequence of points in D has its limit in D. Let $\mathbf{x} \in \text{cl}(D)$. Then $B_{1/k}(\mathbf{x}) \cap D \neq \emptyset$ for every $k \in \mathbb{N}$ and we can choose an $\mathbf{x}_k$ in $B_{1/k}(\mathbf{x}) \cap D$. Then $\mathbf{x}_k \to \mathbf{x}$ as $k \to \infty$ by construction. Thus $\mathbf{x} \in D$ by hypothesis. This shows $\text{cl}(D) \subseteq D$, hence D is closed. □

There is also a smaller brother of the limit of a sequence.

Definition 13.18

A point $\mathbf{a}$ is called an **accumulation point** of a sequence $(\mathbf{x}_k)$ if every open ball $B_\varepsilon(\mathbf{a})$ contains infinitely many elements of the sequence.

Example 13.19

The sequence $\left((-1)^k\right)_{k=1}^{\infty} = (-1, 1, -1, 1 \ldots)$ has accumulation points -1 and 1 but neither point is a limit of the sequence. ♢

13.3 Equivalent Norms

Our definition of open sets and convergent sequences is based on the Euclidean norm (or metric) in $\mathbb{R}^n$. However, we have already seen that the concept of *norm* and *metric* can be generalized. Different norms might result in different families of open sets.

Definition 13.20

Two norms $\|\cdot\|$ and $\|\cdot\|'$ are called (topologically) **equivalent** if every open set w.r.t. $\|\cdot\|$ is also an open set w.r.t. $\|\cdot\|'$ and vice versa.

Thus every interior point w.r.t. $\|\cdot\|$ is also an interior point w.r.t. $\|\cdot\|$' and vice versa. That is, there must exist two strictly positive constants c and d such that

$$c\,\|\mathbf{x}\| \leq \|\mathbf{x}\|' \leq d\,\|\mathbf{x}\|$$

for all $\mathbf{x} \in \mathbb{R}^n$.

An immediate consequence is that every sequence that is convergent w.r.t. some norm is also convergent in every equivalent norm.

Theorem 13.21

Euclidean norm $\|\cdot\|_2$, 1-norm $\|\cdot\|_1$, and supremum norm $\|\cdot\|_\infty$ are equivalent in $\mathbb{R}^n$.

PROOF. By a straightforward computation we find

$$\|\mathbf{x}\|_\infty = \max_{i=1,\dots,n} |x_i| \leq \sum_{i=1}^n |x_i| = \|\mathbf{x}\|_1 \leq \sum_{i=1}^n \left(\max_{j=1,\dots,n} |x_j|\right) = n\,\|\mathbf{x}\|_\infty$$

$$\|\mathbf{x}\|_\infty = \max_{i=1,\dots,n} |x_i| = \sqrt{\max_{i=1,\dots,n} |x_i|^2} \leq \sqrt{\sum_{i=1}^n |x_i|^2} = \|\mathbf{x}\|_2$$

$$\|\mathbf{x}\|_2 = \sqrt{\sum_{i=1}^n |x_i|^2} \leq \sqrt{\sum_{i=1}^n \left(\max_{j=1,\dots,n} |x_j|\right)^2} = \sqrt{n}\|\mathbf{x}\|_\infty$$

Equivalence of Euclidean norm and 1-norm can be derived from Minkowski's inequality. Using $\mathbf{x} = \sum_{i=1}^n x_i \mathbf{e}_i$ we find

$$\|\mathbf{x}\|_2 = \left\|\sum_{i=1}^n x_i \mathbf{e}_i\right\|_2 \leq \sum_{i=1}^n \|x_i \mathbf{e}_i\|_2 = \sum_{i=1}^n \sqrt{|x_i|^2} = \|\mathbf{x}\|_1$$

$$\|\mathbf{x}\|_1 \leq \sum_{i=1}^n \sqrt{\sum_{j=1}^n |x_j|^2} = \sum_{i=1}^n \sqrt{\|\mathbf{x}\|_2^2} = \sqrt{n}\|\mathbf{x}\|_2 \quad \square$$

Notice that the equivalence of Euclidean norm and supremum norm immediately implies Theorem 13.13.

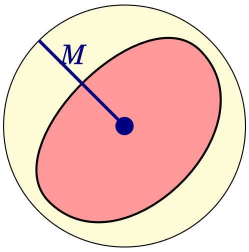
Theorem 13.21 is a corollary of a much stronger result for norms in $\mathbb{R}^n$ which we state without proof.

Theorem 13.22

Finitely generated vector space. All norms in a *finitely generated* vector space are equivalent.

For vector spaces which are not finitely generated this theorem does not hold any more. For example, in probability theory there are different concepts of convergence for sequences of random variates, e.g., convergence in distribution, in probability, almost surely. The corresponding norms or metrices are not equivalent. E.g., a sequence that converges in distribution need not converge almost surely.

13.4 Compact Sets



Bounded set. A set D in $\mathbb{R}^n$ is called **bounded** if there exists a number M such that $\|\mathbf{x}\| \leq M$ for all $x \in D$. A set that is not bounded is called **unbounded**.

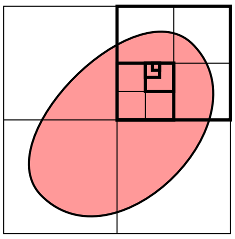
Obviously every convergent sequence is bounded (see Problem 13.15). However, the converse is not true. A sequence in a bounded set need not be convergent. But it always contains an accumulation point and a convergent subsequence.

Definition 13.23

Subsequence. Let $(\mathbf{x}_k)_{k=1}^{\infty}$ be a sequence in $\mathbb{R}^n$. Consider a strictly increasing sequence $k_1 < k_2 < k_3 < k_4 < \dots$ of natural numbers, and let $\mathbf{y}_j = \mathbf{x}_{k_j}$, for $j \in \mathbb{N}$. Then the sequence $(\mathbf{y}_j)_{j=1}^{\infty}$ is called a **subsequence** of $(\mathbf{x}_k)$. It is often denoted by $(\mathbf{x}_{k_j})_{j=1}^{\infty}$.

Example 13.24

Let $(x_k)_{k=1}^{\infty} = \left((-1)^k \frac{1}{k}\right)_{k=1}^{\infty} = \left(-1, \frac{1}{2}, -\frac{1}{3}, \frac{1}{4}, -\frac{1}{5}, \frac{1}{6}, -\frac{1}{7}, \dots\right)$. Then $(y_k)_{k=1}^{\infty} = \left(\frac{1}{2k}\right)_{k=1}^{\infty} = \left(\frac{1}{2}, \frac{1}{4}, \frac{1}{6}, \frac{1}{8}, \dots\right)$ and $(z_k)_{k=1}^{\infty} = \left(-\frac{1}{2k-1}\right)_{k=1}^{\infty} = \left(-1, -\frac{1}{3}, -\frac{1}{5}, -\frac{1}{7}, \dots\right)$ are two subsequences of (x_k). ◊



Now let $(\mathbf{x}_k)$ be a sequence in a bounded subset $D \subseteq \mathbb{R}^2$. Since D is bounded there exists a bounding square $K_0 \supseteq D$ of edge length L. Divide K_0 into four equal squares, each of which has sides of length $L/2$. At least one of these squares, say K_1, must contain infinitely many elements $\mathbf{x}_k$ of this sequence. Pick one of these, say $\mathbf{x}_{k_1}$. Next divide K_1 into four squares of edge length $L/4$. Again in at least one of them, say K_2, there will still be an infinite number of terms from sequence $(\mathbf{x}_k)$. Take one of these, $\mathbf{x}_{k_2}$, with $k_2 > k_1$.

Repeating this procedure ad infinitum we eventually obtain a subsequence $(\mathbf{x}_{k_j})$ of the original sequence that converges by Cauchy's criterion. It is quite obvious that this approach also works in any $\mathbb{R}^n$ where n may not equal to 2. Then we start with a bounding n-cube which is recursively divided into 2^n subcubes.

We summarize our observations in the following theorem (without giving a stringent formal proof).

Theorem 13.25

Bolzano-Weierstrass. A subset D of $\mathbb{R}^n$ is bounded if and only if every sequence of points in D has a convergent subsequence.

Corollary 13.26

A subset D of $\mathbb{R}^n$ is bounded if and only if every sequence has an accumulation point.

We now have seen that convergent sequences can be used to characterize *closed sets* (Theorem 13.17) and *bounded sets* (Theorem 13.25).

Definition 13.27

Compact set. A set D in $\mathbb{R}^n$ is called **compact** if it is closed and bounded.

Compactness is a central concept in mathematical analysis, see, e.g., Theorems 13.36 and 13.37 below. When we combine the results of Theorems 13.17 and 13.25 we get the following characterization.

Theorem 13.28

Bolzano-Weierstrass. A subset D of $\mathbb{R}^n$ is compact if and only if every sequence of points in D has a subsequence that converges to a point in D.

13.5 Continuous Functions

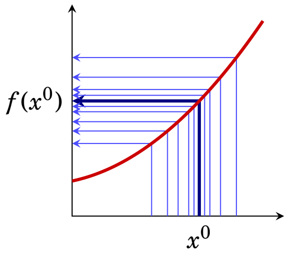
Recall that a univariate function $f\colon \mathbb{R} \to \mathbb{R}$ is called continuous if (roughly spoken) small changes in the argument cause small changes of the function value. One of the formal definitions reads: f is continuous at a point $x^0 \in \mathbb{R}$ if $f(x_k) \to f(x^0)$ for every sequence (x_k) of points that converge to x^0. By our concept of open neighborhood this can easily be generalized for vector-valued functions.

Definition 13.29

Continuous functions. A function $\mathbf{f} = (f_1, \dots, f_m)\colon D \subseteq \mathbb{R}^n \to \mathbb{R}^m$ is said to be **continuous** at a point $\mathbf{x}^0$ if $\mathbf{f}(\mathbf{x}_k) \to \mathbf{f}(\mathbf{x}^0)$ for every sequence $(\mathbf{x}_k)$ of points in D that converges to $\mathbf{x}^0$. We then have

$$\lim_{k\to\infty} \mathbf{f}(\mathbf{x}_k) = \mathbf{f}(\lim_{k\to\infty} \mathbf{x}_k)\,.$$

If $\mathbf{f}$ is continuous at every point $\mathbf{x}^0 \in D$, we say that $\mathbf{f}$ is continuous on D.



The easiest way to show that a vector-valued function is continuous, is by looking at each of its components. We get the following result by means of Theorem 13.13.

Theorem 13.30

Continuity of each component. A function $\mathbf{f} = (f_1, \dots, f_m)\colon D \subseteq \mathbb{R}^n \to \mathbb{R}^m$ is continuous at a point $\mathbf{x}^0$ if and only if each component function $f_j\colon D \subseteq \mathbb{R}^n \to \mathbb{R}$ is continuous at $\mathbf{x}^0$.

There exist equivalent characterizations of continuity which are also used for alternative definitions of continuous functions in the literature. The first one uses open balls.

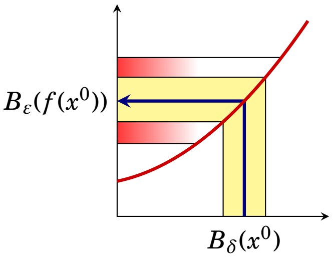
Theorem 13.31

Continuity and images of balls. A function $\mathbf{f}\colon D \subseteq \mathbb{R}^n \to \mathbb{R}^m$ is continuous at a point $\mathbf{x}^0$ in D if and only if for every $\varepsilon > 0$ there exists a $\delta > 0$ such that

$$\left\|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{x}^0)\right\| < \varepsilon \quad \text{for all } \mathbf{x} \in D \text{ with } \quad \left\|\mathbf{x} - \mathbf{x}^0\right\| < \delta$$

or equivalently,

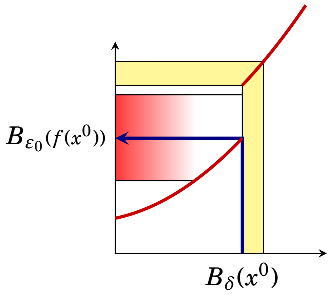
$$\mathbf{f}\bigl(B_\delta(\mathbf{x}^0) \cap D\bigr) \subseteq B_\varepsilon\bigl(\mathbf{f}(\mathbf{x}^0)\bigr)\,.$$



PROOF IDEA. Assume that the condition holds and let $(\mathbf{x}_k)$ be a convergent sequence with limit $\mathbf{x}^0$. Then for every $\varepsilon > 0$ we can find an N such that $\|\mathbf{f}(\mathbf{x}_k) - \mathbf{f}(\mathbf{x}^0)\| < \varepsilon$ for all $k > N$, i.e., $\mathbf{f}(\mathbf{x}_k) \to \mathbf{f}(\mathbf{x}^0)$, which means that $\mathbf{f}$ is continuous at $\mathbf{x}^0$.

Now suppose that there exists an $\varepsilon_0 > 0$ where the condition is violated. Then there exists an $\mathbf{x}_k \in B_\delta(\mathbf{x}^0)$ with $\mathbf{f}(\mathbf{x}_k) \in \mathbf{f}\big(B_\delta(\mathbf{x}^0)\big) \setminus B_{\varepsilon_0}\big(\mathbf{f}(\mathbf{x}^0)\big)$ for every $\delta = \frac{1}{k}$, $k \in \mathbb{N}$. By construction $\mathbf{x}_k \to \mathbf{x}^0$ but $\mathbf{f}(\mathbf{x}_k) \not\to \mathbf{f}(\mathbf{x}^0)$. Thus $\mathbf{f}$ is not continuous at $\mathbf{x}^0$.

PROOF. Suppose that the condition holds. Let $\varepsilon > 0$ be given. Then there exists a $\delta > 0$ such that $\|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{x}^0)\| < \varepsilon$ whenever $\|\mathbf{x} - \mathbf{x}^0\| < \delta$. Now let $(\mathbf{x}_k)$ be a sequence in D that converges to $\mathbf{x}^0$. Thus for every $\delta > 0$ there exists a number N such that $\|\mathbf{x}_k - \mathbf{x}^0\| < \delta$ for all $k > N$. But then $\|\mathbf{f}(\mathbf{x}_k) - \mathbf{f}(\mathbf{x}^0)\| < \varepsilon$ for all $k > N$, and consequently $\mathbf{f}(\mathbf{x}_k) \to \mathbf{f}(\mathbf{x}^0)$ for $k \to \infty$, which implies that $\mathbf{f}$ is continuous at $\mathbf{x}^0$.



Conversely, assume that $\mathbf{f}$ is continuous at $\mathbf{x}^0$ but the condition does not hold, that is, there exists an $\varepsilon_0 > 0$ such that for all $\delta = 1/k$, $k \in \mathbb{N}$, there is an $\mathbf{x} \in D$ with $\|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{x}^0)\| \geq \varepsilon_0$ albeit $\|\mathbf{x} - \mathbf{x}^0\| < 1/k$. Now pick a point $\mathbf{x}_k$ in D with this property for all $k \in \mathbb{N}$. Then sequence $(\mathbf{x}_k)$ converges to $\mathbf{x}^0$ by construction but $\mathbf{f}(\mathbf{x}_k) \notin B_{\varepsilon_0}\big(\mathbf{f}(\mathbf{x}^0)\big)$. This means, however, that $(\mathbf{f}(\mathbf{x}_k))$ does not converge to $\mathbf{f}(\mathbf{x}^0)$, a contradiction to our assumption that $\mathbf{f}$ is continuous. □

Continuous functions $\mathbf{f}\colon \mathbb{R}^n \to \mathbb{R}^m$ can also be characterized by their preimages. While the image $\mathbf{f}(D)$ of some open set $D \subseteq \mathbb{R}^n$ need not necessarily be an open set (see Problem 13.18) this always holds for the **preimage** of some open set $U \subseteq \mathbb{R}^m$,

$$\mathbf{f}^{-1}(U) = \{\mathbf{x}\colon \mathbf{f}(\mathbf{x}) \in U\}\,.$$

For the statement of the general result where the domain of $\mathbf{f}$ is not necessarily open we need the notion of relative open sets.

Definition 13.32 Let D be a subset in $\mathbb{R}^n$. Then

(a) A is **relatively open** in D if $A = U \cap D$ for some open set U in $\mathbb{R}^n$.

(b) A is **relatively closed** in D if $A = F \cap D$ for some closed set F in $\mathbb{R}^n$.

Obviously every open subset of an open set $D \subseteq \mathbb{R}^n$ is relatively open. The usefulness of the concept can be demonstrated by the following example.

Example 13.33 Let $D = [0,1] \subseteq \mathbb{R}$ be the domain of some function f. Then $A = (1/2, 1]$ obviously is not an open set in $\mathbb{R}$. However, A is relatively open in D as $A = (1/2, \infty) \cap [0,1] = (1/2, \infty) \cap D$. ◊

Theorem 13.34 **Characterization of continuity.** A function $\mathbf{f}\colon D \subseteq \mathbb{R}^n \to \mathbb{R}^m$ is continuous if and only if either of the following equivalent conditions is satisfied:

(a) $\mathbf{f}^{-1}(U)$ is relatively open for each open set U in $\mathbb{R}^m$.

(b) $\mathbf{f}^{-1}(F)$ is relatively closed for each closed set F in $\mathbb{R}^m$.

PROOF IDEA. If $U \subseteq \mathbb{R}^m$ is open, then for all $\mathbf{x} \in \mathbf{f}^{-1}(U)$ there exists an $\varepsilon > 0$ such that $B_\varepsilon(\mathbf{f}(\mathbf{x})) \subseteq U$. If in addition $\mathbf{f}$ is continuous, then $B_\delta(\mathbf{x}) \subseteq \mathbf{f}^{-1}(B_\varepsilon(\mathbf{f}(\mathbf{x}))) \subseteq \mathbf{f}^{-1}(U)$ by Theorem 13.31 and hence $\mathbf{f}^{-1}(U)$ is open.

Conversely, if $\mathbf{f}^{-1}(B_\varepsilon(\mathbf{f}(\mathbf{x})))$ is open for all $\mathbf{x} \in D$ and all $\varepsilon > 0$, then there exists a $\delta > 0$ such that $B_\delta(\mathbf{x}) \subseteq \mathbf{f}^{-1}(B_\varepsilon(\mathbf{f}(\mathbf{x})))$ and thus $\mathbf{f}(B_\delta(\mathbf{x})) \subseteq B_\varepsilon(\mathbf{f}(\mathbf{x}))$, i.e., $\mathbf{f}$ is continuous at $\mathbf{x}$ by Theorem 13.31.

PROOF. For simplicity we only prove the case where $D = \mathbb{R}^n$.

(a) Suppose $\mathbf{f}$ is continuous and U is an open set in $\mathbb{R}^m$. Let $\mathbf{x}$ be any point in $\mathbf{f}^{-1}(U)$. Then $\mathbf{f}(\mathbf{x}) \in U$. As U is open there exists an $\varepsilon > 0$ such that $B_\varepsilon(\mathbf{f}(\mathbf{x})) \subseteq U$. By Theorem 13.31 there exists a $\delta > 0$ such that $\mathbf{f}(B_\delta(\mathbf{x})) \subseteq B_\varepsilon(\mathbf{f}(\mathbf{x})) \subseteq U$. Thus $B_\delta(\mathbf{x})$ belongs to the preimage of U. Therefore $\mathbf{x}$ is an interior point of $\mathbf{f}^{-1}(U)$ which means that $\mathbf{f}^{-1}(U)$ is an open set.

Conversely, assume that $\mathbf{f}^{-1}(U)$ is open for each open set $U \subseteq \mathbb{R}^m$. Let $\mathbf{x}$ be any point in D. Let $\varepsilon > 0$ be arbitrary. Then $U = B_\varepsilon(\mathbf{f}(\mathbf{x}))$ is an open set and by hypothesis the preimage $\mathbf{f}^{-1}(U)$ is open in D. Thus there exists a $\delta > 0$ such that $B_\delta(\mathbf{x}) \subseteq \mathbf{f}^{-1}(U) = \mathbf{f}^{-1}(B_\varepsilon(\mathbf{f}(\mathbf{x})))$ and hence $\mathbf{f}(B_\delta(\mathbf{x})) \subseteq U = B_\varepsilon(\mathbf{f}(\mathbf{x}))$. Consequently, $\mathbf{f}$ is continuous at $\mathbf{x}$ by Theorem 13.31. This completes the proof.

(b) This follows immediately from (a) and Lemma 13.5. □

Example 13.35

Let $U(\mathbf{x}) = U(x_1,\dots,x_n)$ be a household's real-valued utility function, where $\mathbf{x}$ denotes its commodity vector and U is defined on the whole of $\mathbb{R}^n$. Then for a number a the upper level set $\Gamma_a = \{\mathbf{x} \in \mathbb{R}^n : U(\mathbf{x}) \geq a\}$ consists of all vectors where the household values are at least as much as a. Let F be the closed interval $[a,\infty)$. Then

$$\Gamma_a = \{\mathbf{x} \in \mathbb{R}^n : U(\mathbf{x}) \geq a\} = \{\mathbf{x} \in \mathbb{R}^n : U(\mathbf{x}) \in F\} = U^{-1}(F)\,.$$

According to Theorem 13.34, if U is continuous, then Γ_a is closed for each value of a. Hence, *continuous functions generate close upper level sets*. They also generate closed lower level sets. ♢

Let $\mathbf{f}$ be a continuous function. As already noted the image $\mathbf{f}(D)$ of some open set needs not be open. Similarly neither the image of a closed set is necessarily closed, nor needs the image of a bounded set be bounded (see Problem 13.18). However, there is a remarkable exception.

Theorem 13.36

Continuous functions preserve compactness. Let $\mathbf{f}\colon D \subseteq \mathbb{R}^n \to \mathbb{R}^m$ be continuous. Then the image $\mathbf{f}(K) = \{\mathbf{f}(\mathbf{x}) : \mathbf{x} \in K\}$ of every compact subset K of D is compact.

PROOF IDEA. Take any sequence $(\mathbf{y}_k)$ in $\mathbf{f}(K)$ and a sequence $(\mathbf{x}_k)$ of its preimages in K, i.e., $\mathbf{y}_k = \mathbf{f}(\mathbf{x}_k)$. We now apply the Bolzano-Weierstrass Theorem twice: $(\mathbf{x}_k)$ has a subsequence $(\mathbf{x}_{k_j})$ that converges to some point $\mathbf{x}^0 \in K$. By continuity $\mathbf{y}_{k_j} = \mathbf{f}(\mathbf{x}_{k_j})$ converges to $\mathbf{f}(\mathbf{x}^0) \in \mathbf{f}(K)$. Hence $\mathbf{f}(K)$ is compact by the Bolzano-Weierstrass Theorem.

PROOF. Let $(\mathbf{y}_k)$ be any sequence in $\mathbf{f}(K)$. By definition, for each k there is a point $\mathbf{x}_k \in K$ such that $\mathbf{y}_k = \mathbf{f}(\mathbf{x}_k)$. By Theorem 13.28 (Bolzano-Weierstrass Theorem), there exists a subsequence $(\mathbf{x}_{k_j})$ that converges to a point $\mathbf{x}^0 \in K$. Because $\mathbf{f}$ is continuous, $\mathbf{f}(\mathbf{x}_{k_j}) \to \mathbf{f}(\mathbf{x}^0)$ as $j \to \infty$ where $\mathbf{f}(\mathbf{x}^0) \in \mathbf{f}(K)$. But then $(\mathbf{y}_{k_j})$ is a subsequence of $(\mathbf{y}_k)$ that converges to $\mathbf{f}(\mathbf{x}^0) \in \mathbf{f}(K)$. Thus $\mathbf{f}(K)$ is compact by Theorem 13.28, as claimed. □

We close this section with an important result in optimization theory.

Theorem 13.37

Extreme-value theorem. Let $\mathbf{f}\colon K \subseteq \mathbb{R}^n \to \mathbb{R}$ be a continuous function on a compact set K. Then $\mathbf{f}$ has both a maximum point and a minimum point in K.

PROOF IDEA. By Theorem 13.36, $\mathbf{f}(K)$ is compact. Thus $\mathbf{f}(K)$ is bounded and closed, that is, $\mathbf{f}(K) = [a,b]$ for $a,b \in \mathbb{R}$ and $\mathbf{f}$ attains its minimum and maximum in respective points $\mathbf{x}_m, \mathbf{x}_M \in K$.

PROOF. By Theorem 13.36, $f(K)$ is compact. In particular, $f(K)$ is bounded, and so $-\infty < a = \inf_{\mathbf{x}\in K} f(\mathbf{x})$ and $b = \sup_{\mathbf{x}\in K} f(\mathbf{x}) < \infty$. Clearly a and b are boundary points of $f(K)$ which belong to $f(K)$, as $f(K)$ is closed. Hence there must exist points $\mathbf{x}_m$ and $\mathbf{x}_M$ such that $f(\mathbf{x}_m) = a$ and $f(\mathbf{x}_M) = b$. Obviously $\mathbf{x}_m$ and $\mathbf{x}_M$ are minimum point and a maximum point of K, respectively. □

— Problems

13.1 Is $Q = \{(x, y) \in \mathbb{R}^2 : x > 0, y \geq 0\}$ open, closed, or neither?

13.2 Is $H = \{(x, y) \in \mathbb{R}^2 : x > 0, y \geq 1/x\}$ open, closed, or neither?

HINT: Sketch set H.

13.3 Let $F = \{(1/k, 0) \in \mathbb{R}^2 : k \in \mathbb{N}\}$. Is F open, closed, or neither?

HINT: Is $(0,0) \in F$?

13.4 Show that the open ball $D = B_r(\mathbf{a})$ is an open set.

HINT: Take any point $\mathbf{x} \in B_r(\mathbf{a})$ and an open ball $B_\varepsilon(\mathbf{x})$ of sufficiently small radius ε. (How small is "sufficiently small"?) Show that $B_\varepsilon(\mathbf{x}) \subseteq D$ by means of the triangle inequality.

13.5 Give respective examples for non-empty sets $D \subseteq \mathbb{R}^2$ which are

(a) neither open nor closed, or
(b) both open and closed, or
(c) closed and have empty interior, or
(d) not closed and have empty interior.

13.6 Show that a set $D \subseteq \mathbb{R}^n$ is closed if and only if its complement $D^c = \mathbb{R}^n \setminus D$ is open (Lemma 13.5).

HINT: Look at boundary points of D.

13.7 Show that a set $D \subseteq \mathbb{R}^n$ is open if and only if its complement D^c is closed.

HINT: Use Lemma 13.5.

13.8 Show that closure $\text{cl}(D)$ and boundary ∂D are closed for any $D \subseteq \mathbb{R}^n$.

HINT: Suppose that there is a boundary point of ∂D that is not a boundary point of D.

13.9 Let D and F be subsets of $\mathbb{R}^n$ such that $D \subseteq F$. Show that

$$\text{int}(D) \subseteq \text{int}(F) \quad \text{and} \quad \text{cl}(D) \subseteq \text{cl}(F).$$

13.10 Recall the proof of Theorem 13.6.

(a) Where exactly do you need the assumption that there is an intersection of *finitely* many open sets in statement (3)?

(b) Let D be the intersection of the infinite family $B_{1/k}(\mathbf{0})$, $k = 1, 2, \ldots$, of open balls centered at $\mathbf{0}$. Is D open or closed?

HINT: Is there any point in D other than $\mathbf{0}$?

13.11 Prove Theorem 13.7.

HINT: Use Theorem 13.6 and De Morgan's law.

13.12 Prove Theorem 13.10.

13.13 Show that the limit of a convergent sequence is uniquely defined.

HINT: Suppose that two limits exist.

13.14 Show that for any points $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ and every $j = 1, \ldots, n$,

$$|x_j - y_j| \leq \|\mathbf{x} - \mathbf{y}\|_2 \leq \sqrt{n} \max_{i=1,\ldots,n} |x_i - y_i| .$$

13.15 Show that every convergent sequence $(\mathbf{x}_k)$ in $\mathbb{R}^n$ is bounded.

13.16 Give an example for a bounded sequence that is not convergent.

HINT: Use Theorem 13.31 and inequality $\big|\|\mathbf{x}\| - \|\mathbf{y}\|\big| \le \|\mathbf{x}-\mathbf{y}\|$.

13.17 For fixed $\mathbf{a} \in \mathbb{R}^n$, show that the function $f : \mathbb{R}^n \to \mathbb{R}$ defined by $f(\mathbf{x}) = \|\mathbf{x}-\mathbf{a}\|$ is continuous.

13.18 Give examples of non-empty subsets D of $\mathbb{R}^n$ and continuous functions $f : \mathbb{R} \to \mathbb{R}$ such that

(a) D is closed, but $f(D)$ is not closed.
(b) D is open, but $f(D)$ is not open.
(c) D is bounded, but $f(D)$ is not bounded.

13.19 Prove that the set

$$D = \{\mathbf{x} \in \mathbb{R}^n : g_j(\mathbf{x}) \le 0, j = 1, \dots, m\}$$

is closed if the functions g_j are all continuous.

14

Derivatives

We want to have the best linear approximation *of a function.*

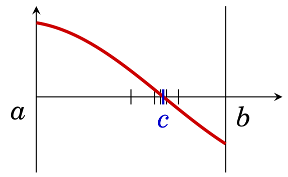
Derivatives are an extremely powerful tool for investigating properties of functions. For univariate functions it allows to check for monotonicity or concavity, or to find candidates for extremal points and verify its optimality. Therefore we want to generalize this tool for multivariate functions.

14.1 Roots of Univariate Functions

The following theorem seems to be trivial. However, it is of great importance as is assures the existence of a root of a continuous function.

Theorem 14.1

Intermediate value theorem (Bolzano). Let $f\colon [a,b] \subseteq \mathbb{R} \to \mathbb{R}$ be a continuous function and assume that $f(a) > 0$ and $f(b) < 0$. Then there exists a point $c \in (a,b)$ such that $f(c) = 0$.



PROOF IDEA. We use a technique called *interval bisectioning*: Start with interval $[a_0,b_0] = [a,b]$, split the interval at $c_1 = (a_1+b_1)/2$ and continue with the subinterval where f changes sign. By iterating this procedure we obtain a sequence of intervals $[a_n,b_n]$ of lengths $(b-a)/2^n \to 0$. By Cauchy's convergence criterion sequence (c_n) converges to some point c with $0 \leq \lim_{n\to\infty} f(c_n) \leq 0$. As f is continuous, we find $f(c) = \lim_{n\to\infty} f(c_n) = 0$.

PROOF. We construct a sequence of intervals $[a_n,b_n]$ by a method called **interval bisectioning**. Let $[a_0,b_0] = [a,b]$. Define $c_n = \frac{a_n+b_n}{2}$ and

$$[a_{n+1},b_{n+1}] = \begin{cases} [c_n,b_n] & \text{if } f(c_n) \geq 0, \\ [a_n,c_n] & \text{if } f(c_n) < 0, \end{cases} \qquad \text{for } n = 1,2,\ldots$$

Notice that $|a_k - a_n| < 2^{-N}(b-a)$ and $|b_k - b_n| < 2^{-N}(b-a)$ for all $k,n \geq N$. Hence (a_i) and (b_i) are Cauchy sequences and thus converge to respective points c_+ and c_- in $[a,b]$ by Cauchy's convergence criterion. Moreover, for every $\varepsilon > 0$, $|c_+ - c_-| \leq |a_k - c_+| + |b_k - c_-| < \varepsilon$ for sufficiently large

k and thus $c_+ = c_- = c$. By construction $f(a_k) \geq 0$ and $f(b_k) \leq 0$ for all k. By assumption f is continuous and thus $f(c) = \lim_{k\to\infty} f(a_k) \geq 0$ and $f(c) = \lim_{k\to\infty} f(b_k) \leq 0$, i.e., $f(c) = 0$ as claimed. □

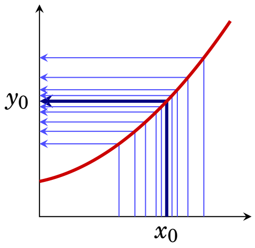
Interval bisectioning is a brute force method for finding a root of some function f. It is sometimes used as a last resort. Notice, however, that this is a rather slow method. *Newton's method*, *secant method* or *regula falsi* are much faster algorithms.

14.2 Limits of a Function

For the definition of derivative we need the concept of *limit* of a function.

Definition 14.2

Limit. Let $\mathbf{f}\colon D \subseteq \mathbb{R} \to \mathbb{R}$ be some function. Then the **limit** of f as x approaches x_0 is y_0 if for every convergent sequence of arguments $x_k \to x_0$, the sequences of images converges to y_0, i.e., $f(x_k) \to y_0$ as $k \to \infty$. We write



$$\lim_{x\to x_0} f(x) = y_0, \quad \text{or} \quad f(x) \to y_0 \quad \text{as} \quad x \to x_0\,.$$

Notice that x_0 need not be an element of domain D and (in abuse of language) may also be ∞ or $-\infty$.

Thus results for limits of sequences (Theorem 12.14) translates immediately into results on limits of functions.

Theorem 14.3

Rules for limits. Let $f\colon \mathbb{R} \to \mathbb{R}$ and $g\colon \mathbb{R} \to \mathbb{R}$ be two functions where both $\lim\limits_{x\to x_0} f(x)$ and $\lim\limits_{x\to x_0} g(x)$ exist. Then

(1) $\lim\limits_{x\to x_0} \big(\alpha f(x) + \beta g(x)\big) = \alpha \lim\limits_{x\to x_0} f(x) + \beta \lim\limits_{x\to x_0} g(x) \qquad \text{for all } \alpha, \beta \in \mathbb{R}$

(2) $\lim\limits_{x\to x_0} \big(f(x) \cdot g(x)\big) = \lim\limits_{x\to x_0} f(x) \cdot \lim\limits_{x\to x_0} g(x)$

(3) $\lim\limits_{x\to x_0} \dfrac{f(x)}{g(x)} = \dfrac{\lim_{x\to x_0} f(x)}{\lim_{x\to x_0} g(x)} \qquad (\text{if } \lim_{x\to x_0} g(x) \neq 0)$

(4) $\lim\limits_{x\to x_0} \big(f(x)\big)^\alpha = \big(\lim\limits_{x\to x_0} f(x)\big)^\alpha \qquad (\text{for } \alpha \in \mathbb{R}, \text{ if } \big(\lim_{x\to x_0} f(x)\big)^\alpha \text{ is defined})$

The notion of *limit* can be easily generalized for arbitrary transformations.

Definition 14.4

Let $\mathbf{f}\colon D \subseteq \mathbb{R}^n \to \mathbb{R}^m$ be some function. Then the **limit** of $\mathbf{f}$ as $\mathbf{x}$ approaches $\mathbf{x}_0$ is $\mathbf{y}_0$ if for every convergent sequence of arguments $\mathbf{x}_k \to \mathbf{x}_0$, the sequences of images converges to $\mathbf{y}_0$, i.e., $\mathbf{f}(\mathbf{x}_0) \to \mathbf{y}_0$. We write

$$\lim_{\mathbf{x}\to\mathbf{x}_0} \mathbf{f}(\mathbf{x}) = \mathbf{y}_0, \quad \text{or} \quad \mathbf{f}(\mathbf{x}) \to \mathbf{y}_0 \quad \text{as} \quad \mathbf{x} \to \mathbf{x}_0\,.$$

The point $\mathbf{x}_0$ need not be an element of domain D.

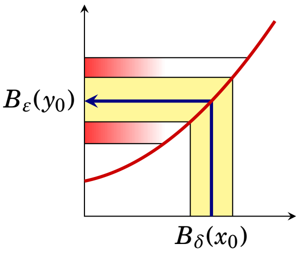
Writing $\lim_{\mathbf{x}\to\mathbf{x}_0} \mathbf{f}(\mathbf{x}) = \mathbf{y}_0$ means that we can make $\mathbf{f}(\mathbf{x})$ as close to $\mathbf{y}_0$ as we want when we put $\mathbf{x}$ sufficiently close to $\mathbf{x}_0$. Notice that a limit at some point $\mathbf{x}_0$ may not exist.

Similarly to our results in Section 13.5 we get the following equivalent characterization of the limit of a function. It is often used as an alternative definition of the term *limit*.

Theorem 14.5

Let $\mathbf{f}\colon D \subseteq \mathbb{R}^n \to \mathbb{R}^m$ be a function. Then $\lim_{\mathbf{x}\to\mathbf{x}_0} \mathbf{f}(\mathbf{x}) = \mathbf{y}_0$ if and only if for every $\varepsilon > 0$ there exists a $\delta > 0$ such that

$$\mathbf{f}(B_\delta(\mathbf{x}_0) \cap D) \subseteq B_\varepsilon(\mathbf{y}_0))\ .$$



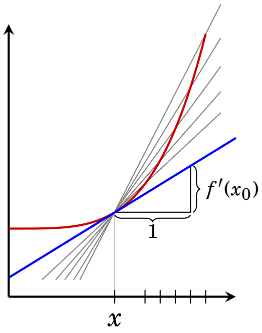
14.3 Derivatives of Univariate Functions

Definition 14.6

Recall that the **derivative** of a function $f: D \subseteq \mathbb{R} \to \mathbb{R}$ at some point x is defined as the limit

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

If this limit exists we say that f is **differentiable** at x. If f is differentiable at every point $x \in D$, we say that f is differentiable on D.



Notice that the term *derivative* is a bit ambiguous. The *derivative at point x* is a *number*, namely the limit of the **difference quotient** of f at point x, that is

$$f'(x) = \frac{d}{dx} f(x) = \lim_{\Delta x\to 0} \frac{\Delta f(x)}{\Delta x} = \lim_{\Delta x\to 0} \frac{f(x+\Delta x) - f(x)}{\Delta x}\ .$$

This number is sometimes called **differential coefficient**. The *differential notation* $\frac{df}{dx}$ is an alternative notation for the derivative which is due to Leibniz. It is very important to remind that *differentiability* is a *local* property of a function.

On the other hand, the **derivative** of f is a *function* that assigns every point x the derivative $\frac{df}{dx}$ at x. Its domain is the set of all points where f is differentiable. Thus $\frac{d}{dx}$ is called the **differential operator** which maps a given function f to its derivative f'. Notice that the differential operator is a linear map, that is

$$\frac{d}{dx}\Big(\alpha f(x) + \beta g(x)\Big) = \alpha \frac{d}{dx} f(x) + \beta \frac{d}{dx} g(x)$$

for all $\alpha, \beta \in \mathbb{R}$, see rules (1) and (2) in Table 14.9.

Differentiability is a stronger property than continuity. Observe that the numerator $f(x+h) - f(x)$ of the difference quotient must coverge to 0 for $h \to 0$ if f is differentiable in x since otherwise the differential quotient would not exist. Thus $\lim_{h\to 0} f(x+h) = f(x)$ and we find:

$f(x)$	$f'(x)$	
c	0	
x^α	$\alpha \cdot x^{\alpha-1}$	(Power rule)
e^x	e^x	
$\ln(x)$	$\dfrac{1}{x}$	
$\sin(x)$	$\cos(x)$	
$\cos(x)$	$-\sin(x)$	

Table 14.8

Derivatives of elementary fun

Lemma 14.7

If $f : D \subseteq \mathbb{R} \to \mathbb{R}$ is differentiable at x, then f is also continuous at x.

See Problem 14.14 for a special case.

Computing limits is a hard job. Therefore, we just list derivatives of some elementary functions in Table 14.8 without proof.

In addition, there exist a couple of rules to reduce the derivative of a given expression to those of elementary functions. Table 14.9 summarizes these rules. Their proofs are straightforward and we given some of these below. See Problem 14.20 for the summation rule and Problem 14.21 for the quotient rule.

PROOF OF RULE (3). Let $F(x) = f(x) \cdot g(x)$. Then we find by Theorem 14.3

$$\begin{aligned} F'(x) &= \lim_{h\to 0} \frac{F(x+h)-F(x)}{h} \\ &= \lim_{h\to 0} \frac{[f(x+h)\cdot g(x+h)] - [f(x)\cdot g(x)]}{h} \\ &= \lim_{h\to 0} \frac{f(x+h)g(x+h) - f(x)g(x+h) + f(x)g(x+h) - f(x)g(x)}{h} \\ &= \lim_{h\to 0} \frac{f(x+h)-f(x)}{h} g(x+h) + \lim_{h\to 0} f(x) \frac{g(x+h)-g(x)}{h} \\ &= \lim_{h\to 0} \frac{f(x+h)-f(x)}{h} \left[\frac{g(x+h)-g(x)}{h} h + g(x) \right] + \lim_{h\to 0} f(x) \frac{g(x+h)-g(x)}{h} \\ &= f'(x)g(x) + f(x)g'(x) \end{aligned}$$

as proposed. □

PROOF OF RULE (4). Let $F(x) = (f \circ g)(x) = f(g(x))$. Then

$$F'(x) = \lim_{h\to 0} \frac{F(x+h)-F(x)}{h} = \lim_{h\to 0} \frac{f(g(x+h)) - f(g(x))}{h}$$

The change from x to $x+h$ causes the value of g change by the amount $k = g(x+h) - g(x)$. As $\lim_{h\to 0} k = \lim_{h\to 0} \left[\frac{g(x+h)-g(x)}{h}\right] \cdot h = g'(x) \cdot 0 = 0$ we find by

Let g be differentiable at x and f be differentiable at x and $g(x)$. Then sum $f+g$, product $f \cdot g$, composition $f \circ g$, and quotient f/g (for $g(x) \neq 0$) are differentiable at x, and

(1) $(c \cdot f(x))' = c \cdot f'(x)$

(2) $(f(x)+g(x))' = f'(x)+g'(x)$ (Summation rule)

(3) $(f(x) \cdot g(x))' = f'(x) \cdot g(x) + f(x) \cdot g'(x)$ (Product rule)

(4) $(f(g(x)))' = f'(g(x)) \cdot g'(x)$ (Chain rule)

(5) $\left(\frac{f(x)}{g(x)}\right)' = \frac{f'(x) \cdot g(x) - f(x) \cdot g'(x)}{(g(x))^2}$ (Quotient rule)

Table 14.9

Rules for differentiation.

Theorem 14.3

$$\begin{aligned} F'(x) &= \lim_{h \to 0} \frac{f(g(x)+k)) - f(g(x))}{k} \cdot \frac{k}{h} \\ &= \lim_{h \to 0} \frac{f(g(x)+k)) - f(g(x))}{k} \cdot \frac{g(x+h)-g(x)}{h} \\ &= f'(g(x)) \cdot g'(x) \end{aligned}$$

as claimed. □

The chain rule can be stated in a quite convenient form by means of differential notation. Let y be function of u, i.e. $y = y(u)$, and u itself is a function of x, i.e., $u = u(x)$, then we find for the derivative of $y(u(x))$,

$$\frac{dy}{dx} = \frac{dy}{du} \cdot \frac{du}{dx} .$$

An important application of the chain rule is in the computation of derivatives when variables are changed. Problem 14.24 discusses the case when linear scale is replaces by logarithmic scale.

14.4 Higher Order Derivatives

We have seen that the derivative f' of a function f is again a function. This function may again be differentiable and we then can compute the derivative of derivative f'. It is called the **second derivative** of f and denoted by f''. Recursively, we can compute the third, forth, fifth, ... derivatives denote by f''', f^{iv}, f^{v}, ...

The nth order derivative is denoted by $f^{(n)}$ and we have

$$f^{(n)} = \frac{d}{dx}\left(f^{(n-1)}\right) \quad \text{with} \quad f^{(0)} = f .$$

14.5 The Mean Value Theorem

Our definition of the derivative of a function,

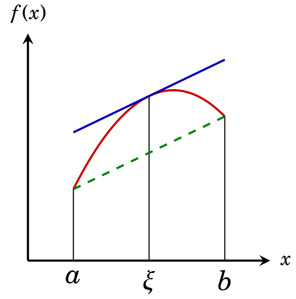
$$f'(x) = \lim_{\Delta x \to 0} \frac{f(x+\Delta x) - f(x)}{\Delta x} ,$$

implies for small values of Δx

$$f(x+\Delta x) \approx f(x) + f'(x)\,\Delta x .$$

The deviation of this *linear* approximation of function f at $x+\Delta x$ becomes small for small values of $|\Delta x|$. We even may improve this approximation.

Theorem 14.10



Mean value theorem. Let f be continuous in the closed bounded interval $[a,b]$ and differentiable in (a,b). Then there exists a point $\xi \in (a,b)$ such that

$$f'(\xi) = \frac{f(b) - f(a)}{b-a} .$$

In particular we find

$$f(b) = f(a) + f'(\xi)(b-a) .$$

PROOF IDEA. We first consider the special case where $f(a) = f(b)$. Then by Theorem 13.37 (and w.l.o.g.) there exists a maximum $\xi \in (a,b)$ of f. We then estimate the limit of the differential quotient when x approaches ξ from the left hand side and from the right hand side, respectively. For the first case we find that $f'(\xi) \geq 0$. The second case implies $f'(\xi) \leq 0$ and hence $f'(\xi) = 0$.

ξ exists by Theorem 13.37.

PROOF. Assume first that $f(a) = f(b)$. If f is constant, then we trivially have $f'(x) = 0 = \frac{f(b)-f(a)}{b-a}$ for all $x \in (a,b)$. Otherwise there exists an x with $f(x) \neq f(a)$. Without loss of generality, $f(x) > f(a)$. (Otherwise we consider $-f$.) Let ξ be a maximum of f, i.e., $f(\xi) \geq f(x)$ for all $x \in [a,b]$. By our assumptions, $\xi \in (a,b)$. Now construct sequences $x_k \to \xi$ as $k \to \infty$ with $x_k \in [a,\xi)$ and $y_k \to \xi$ as $k \to \infty$ with $y_k \in (\xi,b]$. Then we find

$$0 \leq \lim_{k\to\infty} \underbrace{\frac{f(y_k) - f(\xi)}{y_k - \xi}}_{\geq 0} = f'(\xi) = \lim_{k\to\infty} \underbrace{\frac{f(x_k) - f(\xi)}{x_k - \xi}}_{\leq 0} \leq 0 .$$

Consequently, $f'(\xi) = 0$ as claimed.

For the general case consider the function

$$g(x) = f(x) - \frac{f(b) - f(a)}{b-a}(x-a) .$$

Then $g(a) = g(b)$ and there exists a point $\xi \in (a,b)$ such that $g'(\xi) = 0$, i.e., $f'(\xi) - \frac{f(a)-f(b)}{b-a} = 0$. Thus the proposition follows. □

The special case where $f(a) = f(b)$ is also known as **Rolle's theorem**.

14.6 Gradient and Directional Derivatives

Definition 14.11

The **partial derivative** of a multivariate function $f(\mathbf{x}) = f(x_1,\dots,x_n)$ with respect to variable x_i is given as

$$\frac{\partial f}{\partial x_i} = \lim_{h\to 0} \frac{f(\dots,x_i+h,\dots) - f(\dots,x_i,\dots)}{h}$$

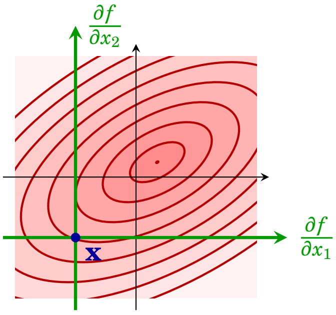
that is, the derivative of f when all variables x_j with $j \neq i$ are held constant.

In the literature there exist several symbols for the partial derivative of f:

$\frac{\partial f}{\partial x_i}$	… derivative w.r.t. x_i
$f_{x_i}(\mathbf{x})$	… derivative w.r.t. variable x_i
$f_i(\mathbf{x})$	… derivative w.r.t. the ith variable
$f'_i(\mathbf{x})$	… ith component of the gradient f'

Notice that the notion of partial derivative is equivalent to the derivative of the univariate function $g(t) = f(\mathbf{x} + t\mathbf{e}_i)$ at $t = 0$, where $\mathbf{e}_i$ denotes the ith unit vector,

$$f_{x_i}(\mathbf{x}) = \frac{\partial f}{\partial x_i} = \left.\frac{dg}{dt}\right|_{t=0} = \left.\frac{d}{dt} f(\mathbf{x} + t\cdot\mathbf{e}_i)\right|_{t=0}$$


$\frac{\partial f}{\partial x_2}$
$\frac{\partial f}{\partial x_1}$
$\mathbf{x}$

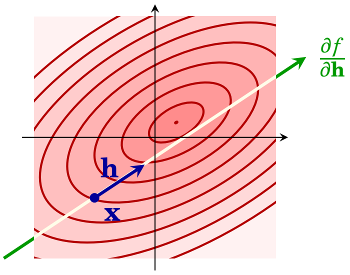
We can, however, replace the unit vectors by arbitrary normalized vectors $\mathbf{h}$ (i.e., $\|\mathbf{h}\| = 1$). Thus we obtain the derivative of f when we move along a straight line through $\mathbf{x}$ in direction $\mathbf{h}$.

Definition 14.12

The **directional derivative** of $f(\mathbf{x}) = f(x_1,\dots,x_n)$ at $\mathbf{x}$ with respect to $\mathbf{h}$ is given by

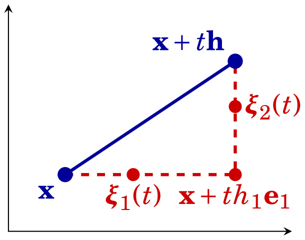
$$f_{\mathbf{h}} = \frac{\partial f}{\partial \mathbf{h}} = \left.\frac{dg}{dt}\right|_{t=0} = \left.\frac{d}{dt} f(\mathbf{x} + t\cdot\mathbf{h})\right|_{t=0}$$

Partial derivatives are special cases of directional derivatives.



The directional derivative can be computed by means of the partial derivatives of f. For the bivariate case ($n = 2$) we find

$$\begin{aligned}
\frac{\partial f}{\partial \mathbf{h}} &= \lim_{t\to 0} \frac{f(\mathbf{x}+t\mathbf{h}) - f(\mathbf{x})}{t} \\
&= \lim_{t\to 0} \frac{\big(f(\mathbf{x}+t\mathbf{h}) - f(\mathbf{x}+th_1\mathbf{e}_1))\big) + \big(f(\mathbf{x}+th_1\mathbf{e}_1) - f(\mathbf{x})\big)}{t} \\
&= \lim_{t\to 0} \frac{f(\mathbf{x}+t\mathbf{h}) - f(\mathbf{x}+th_1\mathbf{e}_1)}{t} + \lim_{t\to 0} \frac{f(\mathbf{x}+th_1\mathbf{e}_1) - f(\mathbf{x})}{t}
\end{aligned}$$



Notice that $t\mathbf{h} = th_1\mathbf{e}_1 + th_2\mathbf{e}_2$. By the mean value theorem there exists a point $\boldsymbol{\xi}_1(t) \in \{\mathbf{x} + \theta h_1\mathbf{e}_1 : \theta \in (0,t)\}$ such that

$$f(\mathbf{x} + th_1\mathbf{e}_1) - f(\mathbf{x}) = f_{x_1}(\boldsymbol{\xi}_1(t)) \cdot th_1$$

and a point $\boldsymbol{\xi}_2(t) \in \{\mathbf{x} + th_1\mathbf{e}_1 + \theta h_2\mathbf{e}_2 : \theta \in (0,t)\}$ such that

$$f(\mathbf{x} + th_1\mathbf{e}_1 + th_2\mathbf{e}_2) - f(\mathbf{x} + th_1\mathbf{e}_1) = f_{x_2}(\boldsymbol{\xi}_2(t)) \cdot th_2 \,.$$

Consequently,

$$\begin{aligned}\frac{\partial f}{\partial \mathbf{h}} &= \lim_{t\to 0} \frac{f_{x_2}(\boldsymbol{\xi}_2(t)) \cdot th_2}{t} + \lim_{t\to 0} \frac{f_{x_1}(\boldsymbol{\xi}_1(t)) \cdot th_1}{t} \\ &= \lim_{t\to 0} f_{x_2}(\boldsymbol{\xi}_2(t))h_2 + \lim_{t\to 0} f_{x_1}(\boldsymbol{\xi}_1(t))h_1 \\ &= f_{x_2}(\mathbf{x})h_2 + f_{x_1}(\mathbf{x})h_1\end{aligned}$$

The last equality holds if the partial derivatives f_{x_1} and f_{x_2} are continuous functions of $\mathbf{x}$.

The continuity of the partial derivatives is crucial for our deduction. Thus we define the class of **continuously differentiable functions**, denoted by $\mathscr{C}^1$.

Definition 14.13

A function $f : D \subseteq \mathbb{R}^n \to \mathbb{R}$ belongs to class $\mathscr{C}^m$ if all its partial derivatives of order m or smaller are continuous. The function belongs to class $\mathscr{C}^\infty$ if partial derivatives of all orders exist.

It also seems appropriate to collect all first partial derivatives in a row vector.

Definition 14.14

Gradient. Let $f : D \subseteq \mathbb{R}^n \to \mathbb{R}$ be a $\mathscr{C}^1$ function. Then the **gradient** of f at $\mathbf{x}$ is the row vector

$$f'(\mathbf{x}) = \nabla f(\mathbf{x}) = \big(f_{x_1}(\mathbf{x}), \ldots, f_{x_n}(\mathbf{x})\big) \qquad \text{[called “}\textit{nabla f}\text{”.]}$$

We can summarize our observations in the following theorem.

Theorem 14.15

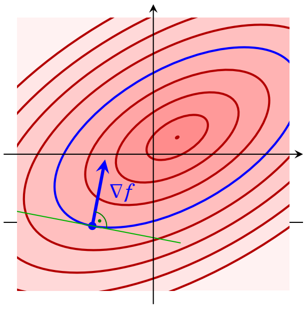
The **directional derivative** of a $\mathscr{C}^1$ function $f(\mathbf{x}) = f(x_1, \ldots, x_n)$ at $\mathbf{x}$ with respect to direction $\mathbf{h}$ with $\|\mathbf{h}\| = 1$ is given by

$$\frac{\partial f}{\partial \mathbf{h}}(\mathbf{x}) = f_{x_1}(\mathbf{x}) \cdot h_1 + \cdots + f_{x_n}(\mathbf{x}) \cdot h_n = \nabla f(\mathbf{x}) \cdot \mathbf{h} \,.$$

This theorem implies some nice properties of the gradient.

Theorem 14.16

Properties of the gradient. Let $f : D \subseteq \mathbb{R}^n \to \mathbb{R}$ be a $\mathscr{C}^1$ function. Then we find



(1) $\nabla f(\mathbf{x})$ points into the direction of the steepest directional derivative at $\mathbf{x}$.

(2) $\|\nabla f(\mathbf{x})\|$ is the maximum among all directional derivatives at $\mathbf{x}$.

(3) $\nabla f(\mathbf{x})$ is orthogonal to the level set through $\mathbf{x}$.

PROOF. By the Cauchy-Schwarz inequality we have

$$\nabla f(\mathbf{x})\mathbf{h} \le |\nabla f(\mathbf{x})\mathbf{h}| \le \|\nabla f(\mathbf{x})\| \cdot \underbrace{\|\mathbf{h}\|}_{=1} = \nabla f(\mathbf{x}) \frac{\nabla f(\mathbf{x})}{\|\nabla f(\mathbf{x})\|}$$

where equality holds if and only if $\mathbf{h} = \nabla f(\mathbf{x})/\|\nabla f(\mathbf{x})\|$. Thus (1) and (2) follow. For the proof of (3) we need the concepts of level sets and implicit functions. Thus we skip the proof. □

14.7 Higher Order Partial Derivatives

Definition 14.17

The functions f_{x_i} are called **first-order partial derivatives**. Provided that these functions are again differentiable, we can generate new functions by taking their partial derivatives. Thus we obtain **second-order partial derivatives**. They are represented as

$$\frac{\partial}{\partial x_j}\left(\frac{\partial f}{\partial x_i}\right) = \frac{\partial^2 f}{\partial x_j \partial x_i} \qquad \text{and} \qquad \frac{\partial}{\partial x_i}\left(\frac{\partial f}{\partial x_i}\right) = \frac{\partial^2 f}{\partial x_i^2}\,.$$

Alternative notations are

$$f_{x_i x_j} \quad \text{and} \quad f_{x_i x_i} \qquad \text{or} \qquad f''_{ij} \quad \text{and} \quad f''_{ii}\,.$$

There are n^2 many second-order derivatives for a function $f(x_1,\dots,x_n)$. Fortunately, for essentially all our functions we need not take care about the succession of particular derivatives. The next theorem provides a sufficient condition. Notice that we again need that all the requested partial derivatives are continuous.

Theorem 14.18

Young's theorem, Schwarz' theorem. Let $f\colon D \subseteq \mathbb{R}^n \to \mathbb{R}$ be a $\mathscr{C}^m$ function, that is, all the mth order partial derivatives of $f(x_1,\dots,x_n)$ exist and are continuous. If any two of them involve differentiating w.r.t. each of the variables the same number of times, then they are necessarily equal. In particular we find for every $\mathscr{C}^2$ function f,

$$\frac{\partial^2 f}{\partial x_i \partial x_j} = \frac{\partial^2 f}{\partial x_j \partial x_i}\,.$$

A proof of this theorem is given in most advanced calculus books.

Definition 14.19

Hessian matrix. Let $f\colon D \subseteq \mathbb{R}^n \to \mathbb{R}$ be a two times differentiable function. Then the $n \times n$ matrix

$$f''(\mathbf{x}) = \mathbf{H}_f(\mathbf{x}) = \begin{pmatrix} f''_{11} & \cdots & f''_{1n} \\ \vdots & \ddots & \vdots \\ f''_{n1} & \cdots & f''_{nn} \end{pmatrix}$$

is called the **Hessian** of f.

By Young's theorem the Hessian is symmetric for $\mathscr{C}^2$ functions.

14.8 Derivatives of Multivariate Functions

We want to generalize the notion of derivative to multivariate functions and transformations. Our starting point is the following observation for univariate functions.

Theorem 14.20

Linear approximation. A function $f: D \subseteq \mathbb{R} \to \mathbb{R}$ is differentiable at an interior point $x_0 \in D$ if and only if there exists a linear function ℓ such that

$$\lim_{h \to 0} \frac{|(f(x_0+h) - f(x_0)) - \ell(h)|}{|h|} = 0 .$$

We have $\ell(h) = f'(x_0) \cdot h$ (i.e., the differential of f at x_0).

PROOF. Assume that f is differentiable in x_0. Then

$$\begin{aligned} \lim_{h \to 0} \frac{(f(x_0+h) - f(x_0)) - f'(x_0)h}{h} &= \lim_{h \to 0} \frac{f(x_0+h) - f(x_0)}{h} - f'(x_0) \\ &= f'(x_0) - f'(x_0) = 0 . \end{aligned}$$

Since the absolute value is a continuous function of its argument, the proposition follows.

Conversely, assume that a linear function $\ell(h) = ah$ exists such that

$$\lim_{h \to 0} \frac{|(f(x_0+h) - f(x_0)) - \ell(h)|}{|h|} = 0 .$$

Then we find

$$\begin{aligned} 0 &= \lim_{h \to 0} \frac{|(f(x_0+h) - f(x_0)) - ah|}{|h|} = \lim_{h \to 0} \frac{(f(x_0+h) - f(x_0)) - ah}{h} \\ &= \lim_{h \to 0} \frac{f(x_0+h) - f(x_0)}{h} - a \end{aligned}$$

and consequently

$$\lim_{h \to 0} \frac{f(x_0+h) - f(x_0)}{h} = a .$$

But then the limit of the difference quotient exists and f is differentiable at x_0. □

An immediate consequence of Theorem 14.20 is that we can use the existence of such a linear function for the definition of the term *differentiable* and the linear function ℓ for the definition of *derivative*. With the notion of *norm* we can easily extend such a definition to transformations.

Definition 14.21

A function $\mathbf{f}: D \subseteq \mathbb{R}^n \to \mathbb{R}^m$ is **differentiable** at an interior point $\mathbf{x}_0 \in D$ if there exists a linear function $\boldsymbol{\ell}$ such that

$$\lim_{\mathbf{h} \to 0} \frac{\|(\mathbf{f}(\mathbf{x}_0+\mathbf{h}) - \mathbf{f}(\mathbf{x}_0)) - \boldsymbol{\ell}(\mathbf{h})\|}{\|\mathbf{h}\|} = 0 .$$

The linear function (if it exists) is then given by an $m \times n$ matrix $\mathbf{A}$, i.e., $\boldsymbol{\ell}(\mathbf{h}) = \mathbf{A}\mathbf{h}$. This matrix is called the **(total) derivative** of $\mathbf{f}$ and denoted by $\mathbf{f}'(\mathbf{x}_0)$ or $D\mathbf{f}(\mathbf{x}_0)$.

Lemma 14.22

A function $\mathbf{f} = (f_1(\mathbf{x}),\dots,f_m(\mathbf{x}))' : D \subseteq \mathbb{R}^n \to \mathbb{R}^m$ is differentiable at an interior point $\mathbf{x}_0$ of D if and only if each component function $f_i : D \to \mathbb{R}$ is differentiable.

PROOF. Let $\mathbf{A}$ be an $m \times n$ matrix and $\mathbf{R}(\mathbf{h}) = (\mathbf{f}(\mathbf{x}_0+\mathbf{h})-\mathbf{f}(\mathbf{x}_0))-\mathbf{A}\mathbf{h}$. Then we find for each $j = 1,\dots,m$,

$$0 \le |R_j(\mathbf{h})| \le \|\mathbf{R}(\mathbf{h})\|_2 \le \|\mathbf{R}(\mathbf{h})\|_1 = \sum_{i=1}^{m} |R_i(\mathbf{h})|\,.$$

Therefore, $\lim_{\mathbf{h}\to 0} \frac{\|\mathbf{R}(\mathbf{h})\|}{\|\mathbf{h}\|} = 0$ if and only if $\lim_{\mathbf{h}\to 0} \frac{|R_j(\mathbf{h})|}{\|\mathbf{h}\|} = 0$ for all $j = 1,\dots,m$. □

The derivative can be computed by means of the partial derivatives of all the components of $\mathbf{f}$.

Theorem 14.23

Computation of derivative. Let $\mathbf{f} = (f_1(\mathbf{x}),\dots,f_m(\mathbf{x}))' : D \subseteq \mathbb{R}^n \to \mathbb{R}^m$ be differentiable at $\mathbf{x}_0$. Then

$$D\mathbf{f}(\mathbf{x}_0) = \begin{pmatrix} \frac{\partial f_1}{x_1}(\mathbf{x}_0) & \dots & \frac{\partial f_1}{x_n}(\mathbf{x}_0) \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{x_1}(\mathbf{x}_0) & \dots & \frac{\partial f_m}{x_n}(\mathbf{x}_0) \end{pmatrix} = \begin{pmatrix} \nabla f_1(\mathbf{x}_0) \\ \vdots \\ \nabla f_m(\mathbf{x}_0) \end{pmatrix}$$

This matrix is called the **Jacobian matrix** of $\mathbf{f}$ at $\mathbf{x}_0$.

PROOF IDEA. In order to compute the components of $\mathbf{f}'(\mathbf{x}_0)$ we estimate the change of f_j as function of the kth variable.

PROOF. Let $\mathbf{A} = (\mathbf{a}_1,\dots,\mathbf{a}_m)'$ denote the derivative of $\mathbf{f}$ at $\mathbf{x}_0$ where $\mathbf{a}'_j$ is the jth row vector of $\mathbf{A}$. By Lemma 14.22 each component function f_j is differentiable at $\mathbf{x}_0$ and thus

$$\lim_{\mathbf{h}\to 0} \frac{|(f_j(\mathbf{x}_0+\mathbf{h}) - f_j(\mathbf{x}_0)) - \mathbf{a}'_j\mathbf{h}|}{\|\mathbf{h}\|} = 0\,.$$

Now set $\mathbf{h} = t\,\mathbf{e}_k$ where $\mathbf{e}_k$ denotes the kth unit vector in $\mathbb{R}^n$. Then

$$\begin{aligned} 0 &= \lim_{t\to 0} \frac{|(f_j(\mathbf{x}_0+t\mathbf{e}_k) - f_j(\mathbf{x}_0)) - t\mathbf{a}'_j\mathbf{e}_k|}{|t|} \\ &= \lim_{t\to 0} \frac{f_j(\mathbf{x}_0+t\mathbf{e}_k) - f_j(\mathbf{x}_0)}{t} - \mathbf{a}'_j\mathbf{e}_k \\ &= \frac{\partial f_j}{\partial x_k}(\mathbf{x}_0) - a_{jk}\,. \end{aligned}$$

That is, $a_{jk} = \frac{\partial f_j}{\partial x_k}(\mathbf{x}_0)$, as proposed. □

Notice that an immediate consequence of Theorem 14.23 is that the derivative $\mathbf{f}'(\mathbf{x}_0)$ is uniquely defined (if it exists).

If $f: D \subseteq \mathbb{R}^n \to \mathbb{R}$, then the Jacobian matrix reduces to a row vector and we find $f'(\mathbf{x}) = \nabla f(\mathbf{x})$, i.e., the gradient of f.

The computation by means of the Jacobian matrix suggests that the derivative of a function exists whenever all its partial derivatives exist. However, this need not be the case. Problem 14.25 shows a counterexample. Nevertheless, there exists a simple condition for the existence of the derivative of a multivariate function.

Theorem 14.24

Existence of derivatives. If $\mathbf{f}$ is a $\mathscr{C}^1$ function from an open set $D \subseteq \mathbb{R}^n$ into $\mathbb{R}^m$, then $\mathbf{f}$ is differentiable at every point $\mathbf{x} \in D$.

SKETCH OF PROOF. Similar to the proof of Theorem 14.15 on page 142. □

Differentiability is a stronger property than continuity as the following result shows.

Theorem 14.25

If $\mathbf{f}: D \subseteq \mathbb{R}^n \to \mathbb{R}^m$ is differentiable at an interior point $\mathbf{x}_0 \in D$, then $\mathbf{f}$ is also continuous at $\mathbf{x}_0$.

PROOF. Let $\mathbf{A}$ denote the derivative at $\mathbf{x}_0$. Then we find

$$\begin{aligned}\|\mathbf{f}(\mathbf{x}_0+\mathbf{h})-\mathbf{f}(\mathbf{x}_0)\| &= \|\mathbf{f}(\mathbf{x}_0+\mathbf{h})-\mathbf{f}(\mathbf{x}_0)-\mathbf{A}\mathbf{h}+\mathbf{A}\mathbf{h}\| \\ &\le \underbrace{\|\mathbf{h}\|}_{\to 0}\cdot\underbrace{\frac{\|\mathbf{f}(\mathbf{x}_0+\mathbf{h})-\mathbf{f}(\mathbf{x}_0)-\mathbf{A}\mathbf{h}\|}{\|\mathbf{h}\|}}_{\to 0}+\underbrace{\|\mathbf{A}\mathbf{h}\|}_{\to 0}\to 0 \quad \text{as } \mathbf{h}\to 0.\end{aligned}$$

The ratio tends to 0 since $\mathbf{f}$ is differentiable. Thus $\mathbf{f}$ is continuous at $\mathbf{x}_0$, as claimed. □

Theorem 14.26

Chain rule. Let $\mathbf{f}: D \subseteq \mathbb{R}^n \to \mathbb{R}^m$ and $\mathbf{g}: B \subseteq \mathbb{R}^m \to \mathbb{R}^p$ with $\mathbf{f}(D) \subseteq B$. Suppose $\mathbf{f}$ and $\mathbf{g}$ are differentiable at $\mathbf{x}$ and $\mathbf{f}(\mathbf{x})$, respectively. Then the composite function $\mathbf{g}\circ\mathbf{f}: D \to \mathbb{R}^p$ defined by $(\mathbf{g}\circ\mathbf{f})(\mathbf{x}) = \mathbf{g}(\mathbf{f}(\mathbf{x}))$ is differentiable at $\mathbf{x}$, and

$$(\mathbf{g}\circ\mathbf{f})'(\mathbf{x}) = \mathbf{g}'\big(\mathbf{f}(\mathbf{x})\big)\cdot\mathbf{f}'(\mathbf{x}))\,.$$

PROOF IDEA. A heuristic derivation for the chain rule using linear approximation is obtained in the following way:

$$\begin{aligned}(\mathbf{g}\circ\mathbf{f})'(\mathbf{x})\mathbf{h} &\approx (\mathbf{g}\circ\mathbf{f})(\mathbf{x}+\mathbf{h})-(\mathbf{g}\circ\mathbf{f})(\mathbf{x}) \\ &= \mathbf{g}\big(\mathbf{f}(\mathbf{x}+\mathbf{h})\big)-\mathbf{g}\big(\mathbf{f}(\mathbf{x})\big) \\ &\approx \mathbf{g}'\big(\mathbf{f}(\mathbf{x})\big)\big[\mathbf{f}(\mathbf{x}+\mathbf{h})-\mathbf{f}(\mathbf{x})\big] \\ &\approx \mathbf{g}'\big(\mathbf{f}(\mathbf{x})\big)\,\mathbf{f}'(\mathbf{x})\,\mathbf{h}\end{aligned}$$

for “sufficiently short” vectors $\mathbf{h}$.

PROOF. Let $\mathbf{R}_f(\mathbf{h}) = \mathbf{f}(\mathbf{x}+\mathbf{h})-\mathbf{f}(\mathbf{x})-\mathbf{f}'(\mathbf{x})\mathbf{h}$ and $\mathbf{R}_g(\mathbf{k}) = \mathbf{g}\big(\mathbf{f}(\mathbf{x})+\mathbf{k}\big)-\mathbf{g}\big(\mathbf{f}(\mathbf{x})\big)-\mathbf{g}'\big(\mathbf{f}(\mathbf{x})\big)\mathbf{k}$. As both $\mathbf{f}$ and $\mathbf{g}$ are differentiable at $\mathbf{x}$ and $\mathbf{f}(\mathbf{x})$, respectively, $\lim\limits_{\mathbf{h}\to 0} \|\mathbf{R}_f(\mathbf{h})\|/\|\mathbf{h}\| = 0$ and $\lim\limits_{\mathbf{k}\to 0} \|\mathbf{R}_g(\mathbf{k})\|/\|\mathbf{k}\| = 0$. Define $\mathbf{k}(\mathbf{h}) = \mathbf{f}(\mathbf{x}+\mathbf{h})-\mathbf{f}(\mathbf{x})$. Then we find

$$\begin{aligned}
\mathbf{R}(\mathbf{h}) &= \mathbf{g}\big(\mathbf{f}(\mathbf{x}+\mathbf{h})\big) - \mathbf{g}\big(\mathbf{f}(\mathbf{x})\big) - \mathbf{g}'\big(\mathbf{f}(\mathbf{x})\big)\,\mathbf{f}'(\mathbf{x})\mathbf{h} \\
&= \mathbf{g}\big(\mathbf{f}(\mathbf{x}) + \mathbf{k}(\mathbf{h})\big) - \mathbf{g}\big(\mathbf{f}(\mathbf{x})\big) - \mathbf{g}'\big(\mathbf{f}(\mathbf{x})\big)\,\mathbf{f}'(\mathbf{x})\mathbf{h} \\
&= \mathbf{g}'\big(\mathbf{f}(\mathbf{x})\big)\,\mathbf{k}(\mathbf{h}) + \mathbf{R}_g\big(\mathbf{k}(\mathbf{h})\big) - \mathbf{g}'\big(\mathbf{f}(\mathbf{x})\big)\,\mathbf{f}'(\mathbf{x})\mathbf{h} \\
&= \mathbf{g}'\big(\mathbf{f}(\mathbf{x})\big)\,\big[\mathbf{k}(\mathbf{h}) - \mathbf{f}'(\mathbf{x})\mathbf{h}\big] + \mathbf{R}_g\big(\mathbf{k}(\mathbf{h})\big) \\
&= \mathbf{g}'\big(\mathbf{f}(\mathbf{x})\big)\,\big[\mathbf{f}(\mathbf{x}+\mathbf{h}) - \mathbf{f}(\mathbf{x}) - \mathbf{f}'(\mathbf{x})\mathbf{h}\big] + \mathbf{R}_g\big(\mathbf{k}(\mathbf{h})\big) \\
&= \mathbf{g}'\big(\mathbf{f}(\mathbf{x})\big)\,\mathbf{R}_f(\mathbf{h}) + \mathbf{R}_g\big(\mathbf{k}(\mathbf{h})\big)\,.
\end{aligned}$$

Thus by the triangle inequality we have

$$\frac{\|\mathbf{R}(\mathbf{h})\|}{\|\mathbf{h}\|} \le \frac{\|\mathbf{g}'(\mathbf{f}(\mathbf{x}))\mathbf{R}_f(\mathbf{h})\|}{\|\mathbf{h}\|} + \frac{\|\mathbf{R}_g(\mathbf{k}(\mathbf{h}))\|}{\|\mathbf{h}\|}\,.$$

The right hand side converges to zero as $\mathbf{h} \to 0$ and hence proposition follows[1]. □

Notice that the derivatives in the chain rule are *matrices*. Thus the derivative of a composite function is the composite of linear functions.

Example 14.27

Let $\mathbf{f}(x,y) = \begin{pmatrix} x^2+y^2 \\ x^2-y^2 \end{pmatrix}$ and $\mathbf{g}(x,y) = \begin{pmatrix} e^x \\ e^y \end{pmatrix}$ be two differentiable functions defined on $\mathbb{R}^2$. Compute the derivative of $\mathbf{g}\circ\mathbf{f}$ at $\mathbf{x}$ by means of the chain rule.

SOLUTION. Since $\mathbf{f}'(\mathbf{x}) = \begin{pmatrix} 2x & 2y \\ 2x & -2y \end{pmatrix}$ and $\mathbf{g}'(\mathbf{x}) = \begin{pmatrix} e^x & 0 \\ 0 & e^y \end{pmatrix}$, we have

$$\begin{aligned}
(\mathbf{g}\circ\mathbf{f})'(\mathbf{x}) = \mathbf{g}'(\mathbf{f}(\mathbf{x}))\mathbf{f}'(\mathbf{x}) &= \begin{pmatrix} e^{x^2+y^2} & 0 \\ 0 & e^{x^2-y^2} \end{pmatrix} \cdot \begin{pmatrix} 2x & 2y \\ 2x & -2y \end{pmatrix} \\
&= \begin{pmatrix} 2x\,e^{x^2+y^2} & 2y\,e^{x^2+y^2} \\ 2x\,e^{x^2-y^2} & -2y\,e^{x^2-y^2} \end{pmatrix}
\end{aligned}$$

♢

Example 14.28

Derive the formula for the directional derivative from Theorem 14.15 by means of the chain rule.

SOLUTION. Let $f\colon D \subseteq \mathbb{R}^n \to \mathbb{R}$ some differentiable function and $\mathbf{h}$ a fixed direction (with $\|\mathbf{h}\| = 1$). Then $\mathbf{s}\colon \mathbb{R} \to D \subseteq \mathbb{R}^n, t \mapsto \mathbf{x}_0 + t\mathbf{h}$ is a path in $\mathbb{R}^n$ and we find

$$f'(\mathbf{s}(0)) = f'(\mathbf{x}_0) = \nabla f(\mathbf{x}_0) \qquad \text{and} \qquad \mathbf{s}'(0) = \mathbf{h}$$

[1] At this point we need some tools from advanced calculus which we do not have available. Thus we unfortunately still have an heuristic approach albeit on some higher level.

and therefore

$$\frac{\partial f}{\partial \mathbf{h}}(\mathbf{x}_0) = (f \circ \mathbf{s})'(0) = f'(\mathbf{s}(0)) \cdot \mathbf{s}'(0) = \nabla f(\mathbf{x}_0) \cdot \mathbf{h}$$

as claimed. ♢

Example 14.29

Let $f(x_1, x_2, t)$ be a differentiable function defined on $\mathbb{R}^3$. Suppose that both $x_1(t)$ and $x_2(t)$ are themselves functions of t. Compute the total derivative of $z(t) = f\big(x_1(t), x_2(t), t\big)$.

SOLUTION. Let $\mathbf{x}\colon \mathbb{R} \to \mathbb{R}^3$, $t \mapsto \begin{pmatrix} x_1(t) \\ x_2(t) \\ t \end{pmatrix}$. Then $z(t) = (f \circ \mathbf{x})(t)$ and we have

$$\begin{aligned}
\frac{dz}{dt} &= (f \circ \mathbf{x})'(t) = f'\big(\mathbf{x}(t)\big) \cdot \mathbf{x}'(t) \\
&= \nabla f\big(\mathbf{x}(t)\big) \cdot \begin{pmatrix} x_1'(t) \\ x_2'(t) \\ 1 \end{pmatrix} = \Big(f_{x_1}\big(\mathbf{x}(t)\big), f_{x_2}\big(\mathbf{x}(t)\big), f_t\big(\mathbf{x}(t)\big)\Big) \cdot \begin{pmatrix} x_1'(t) \\ x_2'(t) \\ 1 \end{pmatrix} \\
&= f_{x_1}\big(\mathbf{x}(t)\big) \cdot x_1'(t) + f_{x_2}\big(\mathbf{x}(t)\big) \cdot x_2'(t) + f_t\big(\mathbf{x}(t)\big) \\
&= f_{x_1}(x_1, x_2, t) \cdot x_1'(t) + f_{x_2}(x_1, x_2, t) \cdot x_2'(t) + f_t(x_1, x_2, t) . \quad \diamondsuit
\end{aligned}$$

— Exercises

14.1 Estimate the following limits:

(a) $\lim\limits_{x\to\infty} \frac{1}{x+1}$ (b) $\lim\limits_{x\to 0} x^2$ (c) $\lim\limits_{x\to\infty} \ln(x)$

(d) $\lim\limits_{x\to 0} \ln|x|$ (e) $\lim\limits_{x\to\infty} \frac{x+1}{x-1}$

14.2 Sketch the following functions.
Which of these are continuous functions?
In which points are these functions not continuous?

(a) $D=\mathbb{R}, f(x)=x$ (b) $D=\mathbb{R}, f(x)=3x+1$

(c) $D=\mathbb{R}, f(x)=e^{-x}-1$ (d) $D=\mathbb{R}, f(x)=|x|$

(e) $D=\mathbb{R}^+, f(x)=\ln(x)$ (f) $D=\mathbb{R}, f(x)=[x]$

(g) $D=\mathbb{R},\ f(x)=\begin{cases} 1 & \text{for } x\le 0\\ x+1 & \text{for } 0<x\le 2\\ x^2 & \text{for } x>2\end{cases}$

HINT: Let $x=p+y$ with $p\in\mathbb{Z}$ and $y\in[0,1)$. Then $[x]=p$.

14.3 Differentiate:

(a) $3x^2+5\cos(x)+1$ (b) $(2x+1)x^2$

(c) $x\ln(x)$ (d) $(2x+1)x^{-2}$

(e) $\frac{3x^2-1}{x+1}$ (f) $\ln(\exp(x))$

(g) $(3x-1)^2$ (h) $\sin(3x^2)$

(i) 2^x (j) $\frac{(2x+1)(x^2-1)}{x+1}$

(k) $2e^{3x+1}(5x^2+1)^2+\frac{(x+1)^3}{x-1}-2x$

14.4 Compute the second and third derivatives of the following functions:

(a) $f(x)=e^{-\frac{x^2}{2}}$ (b) $f(x)=\dfrac{x+1}{x-1}$

(c) $f(x)=(x-2)(x^2+3)$

14.5 Compute all first and second order partial derivatives of the following functions at $(1,1)$:

(a) $f(x,y)=x+y$ (b) $f(x,y)=xy$

(c) $f(x,y)=x^2+y^2$ (d) $f(x,y)=x^2y^2$

(e) $f(x,y)=x^\alpha y^\beta,\quad \alpha,\beta>0$ (f) $f(x,y)=\sqrt{x^2+y^2}$

14.6 Compute gradient and Hessian matrix of the functions in Exercise 14.5 at $(1,1)$.

14.7 Let $f(\mathbf{x}) = \sum_{i=1}^{n} x_i^2$. Compute the directional derivative of f into direction $\mathbf{a}$ using

(a) function $g(t) = f(\mathbf{x}+t\mathbf{a})$;

(b) the gradient ∇f;

(c) the chain rule.

14.8 Let $f(x,y)$ be a differentiable function. Suppose its directional derivative in $(0,0)$ in maximal in direction $\mathbf{a} = (1,3)$ with $\frac{\partial f}{\partial \mathbf{a}} = 4$. Compute the gradient of f in $(0,0)$.

14.9 Let $f(x,y) = x^2 + y^2$ and $\mathbf{g}(t) = \begin{pmatrix} g_1(t) \\ g_2(t) \end{pmatrix} = \begin{pmatrix} t \\ t^2 \end{pmatrix}$. Compute the derivatives of the compound functions $f \circ g$ and $g \circ f$ by means of the chain rule.

14.10 Let $\mathbf{f}(\mathbf{x}) = (x_1^3 - x_2, x_1 - x_2^3)'$ and $\mathbf{g}(\mathbf{x}) = (x_2^2, x_1)'$. Compute the derivatives of the compound functions $\mathbf{f}\circ\mathbf{g}$ and $\mathbf{g}\circ\mathbf{f}$ by means of the chain rule.

HINT: Use Cramer's rule.

14.11 Let $\mathbf{A}$ be a regular $n \times n$ matrix, $\mathbf{b} \in \mathbb{R}^n$ and $\mathbf{x}$ the solution of the linear equation $\mathbf{A}\mathbf{x} = \mathbf{b}$. Compute $\frac{\partial x_i}{\partial b_i}$. Also give the Jacobian matrix of $\mathbf{x}$ as a function of $\mathbf{b}$.

14.12 Let $F(K,L,t)$ be a production function where $L = L(t)$ and $K = K(t)$ are also functions of time t. Compute $\frac{dF}{dt}$.

— Problems

HINT: See proof of Theorem 13.31.

14.13 Prove Theorem 14.5.

14.14 Let $f(x) = x^n$ for some $n \in \mathbb{N}$. Show that $f'(x) = n\,x^{n-1}$ by computing the limit of the difference quotient.

Say "n choose k".

HINT: Use the binomial theorem

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k \cdot b^{n-k}$$

14.15 Show that $f(x) = |x|$ is not differentiable on $\mathbb{R}$.

HINT: Recall that a function is differentiable on D if it is differentiable on every $x \in D$.

14.16 Show that

$$f(x) = \begin{cases} \sqrt{x}, & \text{for } x \geq 0, \\ -\sqrt{-x}, & \text{for } x < 0, \end{cases}$$

is not differentiable on $\mathbb{R}$.

14.17 Construct a function that is differentiable but not twice differentiable.

HINT: Recall that a function is differentiable on D if it is differentiable on every $x \in D$.

14.18 Show that the function

$$f(x) = \begin{cases} x^2 \sin\left(\frac{1}{x}\right), & \text{for } x \neq 0, \\ 0, & \text{for } x = 0, \end{cases}$$

is differentiable in $x = 0$ but not continuously differentiable.

14.19 Compute the derivative of $f(x) = a^x$ $(a > 0)$.

HINT: $a^x = e^{\ln(a)x}$

14.20 Prove the summation rule. (Rule (2) in Table 14.9).

HINT: Let $F(x) = f(x) + g(x)$ and apply Theorem 14.3 for the limit.

14.21 Prove the quotient rule. (Rule (5) in Table 14.9).

HINT: Use chain rule, product rule and power rule.

14.22 Verify the *Square Root Rule*:

$$\left(\sqrt{x}\right)' = \frac{1}{2\sqrt{x}}$$

HINT: Use the rules from Tabs. 14.8 and 14.9.

14.23 Let $f : \mathbb{R} \to (0,\infty)$ be a differentiable function. Show that

$$(\ln(f(x)))' = \frac{f'(x)}{f(x)}$$

14.24 Let $f : (0,\infty) \to (0,\infty)$ be a differentiable function. Then the term

$$\varepsilon_f(x) = x \cdot \frac{f'(x)}{f(x)}$$

is called the **elasticity** of f at x. It describes relative changes of f w.r.t. relative changes of its variable x. We can, however, derive the elasticity by changing from a linear scale to a logarithmic scale. Thus we replace variable x by its logarithm $v = \ln(x)$ and differentiate the logarithm of f w.r.t. v and find

$$\varepsilon_f(x) = \frac{d(\ln(f(x)))}{d(\ln(x))}$$

Derive this formula by means of the chain rule.

HINT: Differentiate $y(v) = \ln(f(e^v))$ and substitute $v = \ln(x)$.

14.25 Let

$$f(x,y) = \begin{cases} \dfrac{xy^2}{x^2 + y^4}, & \text{for } (x,y) \neq 0, \\ 0, & \text{for } (x,y) = 0. \end{cases}$$

(a) Plot the graph of f (by means of the computer program of your choice).

(b) Compute all first partial derivatives for $(x, y) \neq 0$.

(c) Compute all first partial derivatives for $(x, y) = 0$ by computing the respective limits

$$f_x(0,0) = \lim_{t \to 0} \frac{f(t,0) - f(0,0)}{t}$$

$$f_y(0,0) = \lim_{t \to 0} \frac{f(0,t) - f(0,0)}{t}$$

(d) Compute the directional derivative at 0 into some direction $\mathbf{h}' = (h_1, h_2)$,

$$f_{\mathbf{h}}(0,0) = \lim_{t \to 0} \frac{f(th_1, th_2) - f(0,0)}{t}$$

What do you expect if f were differentiable at 0?

14.26 Let $\mathbf{f}\colon \mathbb{R}^n \to \mathbb{R}^m$ be a linear function with $\mathbf{f}(\mathbf{x}) = \mathbf{A}\mathbf{x}$ for some matrix $\mathbf{A}$.

(a) What are the dimensions of matrix $\mathbf{A}$ (number of rows and columns)?

(b) Compute the Jacobian matrix of $\mathbf{f}$.

14.27 Let $\mathbf{A}$ be a symmetric $n \times n$ matrix. Compute the Jacobian matrix of the corresponding quadratic form $\mathbf{q}(\mathbf{x}) = \mathbf{x}'\mathbf{A}\mathbf{x}$.

14.28 A function $f(\mathbf{x})$ is called **homogeneous** of degree k, if

$$f(\alpha\mathbf{x}) = \alpha^k f(\mathbf{x}) \quad \text{for all } \alpha \in \mathbb{R}.$$

(a) Give an example for a homogeneous function of degree 2 and draw level lines of this function.

(b) Show that all first order partial derivatives of a differentiable homogeneous function of degree k ($k \geq 1$) are homogeneous of degree $k - 1$.

(c) Show that the level lines are parallel along each ray from the origin. (A **ray** from the origin in direction $\mathbf{r} \neq 0$ is the halfline $\{\mathbf{x} = \alpha\mathbf{r}\colon \alpha \geq 0\}$.)

HINT: Differentiate both sides of equation $f(\alpha\mathbf{x}) = \alpha^k f(\mathbf{x})$ w.r.t. x_i.

14.29 Let f and g be two n times differentiable functions. Show by induction that

$$(f \cdot g)^{(n)}(x) = \sum_{k=0}^{n} \binom{n}{k} f^{(k)}(x) \cdot g^{(n-k)}(x).$$

HINT: Use the recursion $\binom{n+1}{k+1} = \binom{n}{k} + \binom{n}{k+1}$ for $k = 0, \ldots, n-1$.

14.30 Let

$$f(x) = \begin{cases} \exp\left(-\frac{1}{x^2}\right), & \text{for } x \neq 0, \\ 0, & \text{for } x = 0. \end{cases}$$

(a) Show that f is differentiable in $x = 0$.

(b) Show that $f'(x) = \begin{cases} -x^{-3} f(x), & \text{for } x \neq 0, \\ 0, & \text{for } x = 0. \end{cases}$

(c) Show that f is continuously differentiable in $x = 0$.

(d) Argue why all derivatives of f vanish in $x = 0$, i.e., $f^{(n)}(0) = 0$ for all $n \in \mathbb{N}$.

HINT: For (a) use $\lim_{x\to 0} f(x) = \lim_{x\to\infty} f\left(\frac{1}{x}\right)$; for (b) use the formula from Problem 14.29.

15

Taylor Series

We need a local approximation *of a function that is as simple as possible, but not simpler.*

15.1 Taylor Polynomial

The derivative of a function can be used to find the best linear approximation of a univariate function f, i.e.,

$$f(x) \approx f(x_0) + f'(x_0)(x - x_0) .$$

Notice that we evaluate both f and its derivative f' at x_0. By the mean value theorem (Theorem 14.10) we have

$$f(x) = f(x_0) + f'(\xi)(x - x_0)$$

for some appropriate point $\xi \in [x, x_0]$. When we need to improve this *first-order approximation*, then we have to use a polynomial p_n of degree n. We thus select the coefficients of this polynomial such that its first n derivatives at some point x_0 coincides with the first k derivatives of f at x_0, i.e.,

$$p_n^{(k)}(x_0) = f^{(k)}(x_0) , \quad \text{for } k = 0, \dots, n.$$

We then find

$$f(x) = \sum_{k=0}^{n} \frac{f^{(k)}(x_0)}{k!} (x - x_0)^k + R_n(x) .$$

The term R_n is called the **remainder** and is the error when we approximate function f by this so called Taylor polynomial of degree n.

Definition 15.1

Let f be an n times differentiable function. Then the polynomial

$$T_{f,x_0,n}(x) = \sum_{k=0}^{n} \frac{f^{(k)}(x_0)}{k!} (x - x_0)^k$$

is called the nth-order **Taylor polynomial** of f around $x = x_0$. The term $f^{(0)}$ refers to the “0-th derivative”, i.e., function f itself.

The special case with $x_0 = 0$ is called the **Maclaurin polynomial**.

If we expand the summation symbol we can write the Maclaurin polynomial as

$$T_{f,0,n}(x) = f(0) + f'(0)x + \frac{f''(0)}{2!}x^2 + \frac{f'''(0)}{3!}x^3 + \cdots + \frac{f^{(n)}(0)}{n!}x^n \,.$$

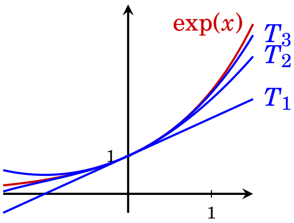
Example 15.2

Exponential function. The derivatives of $f(x) = e^x$ at $x_0 = 0$ are given by

$$f^{(n)}(x) = e^x \quad \text{hence} \quad f^{(n)}(0) = 1 \quad \text{for all } n \geq 0.$$

Therefore we find for the nth order Maclaurin polynomial

$$T_{f,0,n}(x) = \sum_{k=0}^{n} \frac{f^{(k)}(0)}{k!}x^k = \sum_{k=0}^{n} \frac{x^k}{k!} \,. \qquad \diamondsuit$$



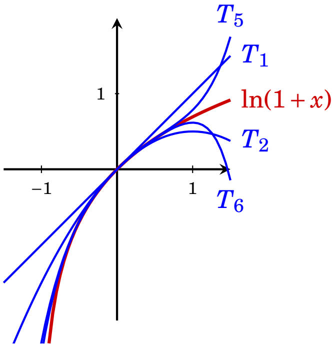
Example 15.3

Logarithm. The derivatives of $f(x) = \ln(1+x)$ at $x_0 = 0$ are given by

$$f^{(n)}(x) = (-1)^{n+1}(n-1)!\,(1+x)^{-n}$$

hence $f^{(n)}(0) = (-1)^{n+1}(n-1)!$ for all $n \geq 1$. As $f(0) = \ln(1) = 0$ we find for the nth order Maclaurin polynomial

$$T_{f,0,n}(x) = \sum_{k=0}^{n} \frac{f^{(k)}(0)}{k!}x^k = \sum_{k=1}^{n} \frac{(-1)^{k+1}(k-1)!}{k!}x^k = \sum_{k=1}^{n} (-1)^{k+1}\frac{x^k}{k} \,. \qquad \diamondsuit$$



Obviously, the approximation of a function f by its Taylor polynomial is only useful if the remainder $R_n(x)$ is small. Indeed, the error will go to 0 faster than $(x-x_0)^n$ as x tends to x_0.

Theorem 15.4

Taylor's theorem. Let function $f\colon \mathbb{R} \to \mathbb{R}$ be n times differentiable at the point $x_0 \in \mathbb{R}$. Then there exists a function $h_n\colon \mathbb{R} \to \mathbb{R}$ such that

$$f(x) = T_{f,x_0,n}(x) + h_n(x)(x-x_0)^n \qquad \text{and} \qquad \lim_{x\to x_0} h_n(x) = 0\,.$$

There are even stronger results. The error term can be estimated more precisely. The following theorem gives one such result. Observe that Theorem 15.4 is then just a corollary when the assumptions of Theorem 15.5 are met.

Theorem 15.5

Lagrange's form of the remainder. Suppose f is $n+1$ times differentiable in the interval $[x, x_0]$. Then the remainder for $T_{f,x_0,n}$ can be written as

$$R_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!}(x-x_0)^{n+1}$$

for some point $\xi \in (x, x_0)$.

PROOF IDEA. We construct a function

$$g(t) = R_n(t) - \frac{(t-x_0)^{n+1}}{(x-x_0)^{n+1}} R_n(x)$$

and show that all derivatives $g^{(k)}(x_0) = 0$ vanish for all $k = 0, \ldots, n$. Moreover, $g(\xi_0) = 0$ for $\xi_0 = x$ and thus Rolle's Theorem implies that there exists a $\xi_1 \in (\xi_0, x_0)$ such that $g'(\xi_1) = 0$. Repeating this argument recursively we eventually obtain a $\xi = \xi_{n+1} \in (\xi_k, x_0) \subseteq (x, x_0)$ with $g^{(n+1)}(\xi) = f^{(n+1)}(\xi) - \frac{(n+1)!}{(x-x_0)^{n+1}} R_n(x) = 0$ and thus the result follows.

PROOF. Let $R_n(x) = f(x) - T_{f,x_0,n}(x)$ and

$$g(t) = R_n(t) - \frac{(t-x_0)^{n+1}}{(x-x_0)^{n+1}} R_n(x) .$$

We then find $g(x) = 0$. Moreover, $g(x_0) = 0$ and $g^{(k)}(x_0) = 0$ for all $k = 0, \ldots, n$ since the first n derivatives of f and $T_{f,x_0,n}$ coincide at x_0 by construction (Problem 15.9). Thus $g(x) = g(x_0)$ and the mean value theorem (Rolle's Theorem, Theorem 14.10) implies that there exists a $\xi_1 \in (x, x_0)$ such that $g'(\xi_1) = 0$ and thus $g'(\xi_1) = g'(x_0) = 0$. Again the mean value theorem implies that there exists a $\xi_2 \in (\xi_1, x_0) \subseteq (x, x_0)$ such that $g''(\xi_2) = 0$. Repeating this argument we find $\xi_1, \xi_2, \ldots, \xi_{n+1} \in (x, x_0)$ such that $g^{(k)}(\xi_k) = 0$ for all $k = 1, \ldots, n+1$. In particular, for $\xi = \xi_{n+1}$ we then have

$$0 = g^{(n+1)}(\xi) = f^{(n+1)}(\xi) - \frac{(n+1)!}{(x-x_0)^{n+1}} R_n(x)$$

and thus the formula for R_n follows. □

Lagrange's form of the remainder can be seen as a generalization of the mean value theorem for higher order derivatives.

15.2 Taylor Series

Taylor series expansion. The series Definition 15.6

$$\sum_{n=0}^{\infty} \frac{f^{(n)}(x_0)}{n!} (x-x_0)^n$$

is called the **Taylor series** of f at x_0. We say that we **expand** f into a *Taylor series* around x_0.

If the remainder $R_n(x) \to 0$ as $n \to \infty$, then the Taylor series converges to $f(x)$, i.e., we them have

$$f(x) = \sum_{n=0}^{\infty} \frac{f^{(n)}(x_0)}{n!} (x-x_0)^n .$$

Table 15.7 lists Maclaurin series of some important functions. The meaning of ρ is explained in Section 15.4 below.

In some cases it is quite straightforward to show the convergence of the Taylor series.

$f(x)$	Maclaurin series		ρ
$\exp(x)$	$=\sum_{n=0}^{\infty}\frac{x^n}{n!}$	$=1+x+\frac{x^2}{2!}+\frac{x^3}{3!}+\frac{x^4}{4!}+\cdots$	∞
$\ln(1+x)$	$=\sum_{n=1}^{\infty}(-1)^{n+1}\frac{x^n}{n}$	$=x-\frac{x^2}{2}+\frac{x^3}{3}-\frac{x^4}{4}+\cdots$	1
$\sin(x)$	$=\sum_{n=0}^{\infty}(-1)^n\frac{x^{2n+1}}{(2n+1)!}$	$=x-\frac{x^3}{3!}+\frac{x^5}{5!}-\frac{x^7}{7!}+\cdots$	∞
$\cos(x)$	$=\sum_{n=0}^{\infty}(-1)^n\frac{x^{2n}}{(2n)!}$	$=1-\frac{x^2}{2!}+\frac{x^4}{4!}-\frac{x^6}{6!}+\cdots$	∞
$\frac{1}{1-x}$	$=\sum_{n=0}^{\infty}x^n$	$=1+x+x^2+x^3+x^4+\cdots$	1

Table 15.7

Maclaurin seri
some elementa
functions.

Theorem 15.8

Convergence of remainder. Assume that *all* derivatives of f are bounded in the interval (x,x_0) by some number M, i.e., $|f^{(k)}(\xi)|\le M$ for all $\xi\in(x,x_0)$ and all $k\in\mathbb{N}$. Then

$$|R_n(x)|\le M\frac{|x-x_0|^{n+1}}{(n+1)!}\qquad\text{for all } n\in\mathbb{N}$$

and thus $\lim_{n\to\infty}R_n(x)=0$ as $n\to\infty$.

PROOF. Immediately by Theorem 15.5 and hypothesis of the theorem. □

Example 15.9

We have seen in Example 15.2 that $f^{(n)}(x)=e^x$ for all $n\in\mathbb{N}$. Thus $|f^{(n)}(\xi)|\le M=\max\{|e^x|,|e^{x_0}|\}$ for all $\xi\in(x,x_0)$ and all $k\in\mathbb{N}$. Then by Theorem 15.8, $e^x=\sum_{n=0}^{\infty}\frac{f^{(n)}(0)}{n!}x^n$ for all $x\in\mathbb{R}$. ◊

The required order of the Taylor polynomial for the approximation of a function f of course depends on the particular task. A first-order Taylor polynomial may be used to linearize a given function near some point of interest. This also may be sufficient if one needs to investigate local monotonicity of some function. When local convexity or concavity of the function are of interest we need at least a second-order Taylor polynomial.

15.3 Landau Symbols

If all derivatives of f are bounded in the interval (x,x_0), then Lagrange's form of the remainder $R_n(x)$ is expressed as a multiple of the nth power of the distance between the point x of interest and the expansion point x_0, that is, $C|x-x_0|^{n+1}$ for some positive constant C. The constant itself is often hard to compute and thus it is usually not specified. However, in many cases this is not necessary at all.

Suppose we have two terms $C_1|x-x_0|^k$ and $C_2|x-x_0|^{k+1}$ with $C_1, C_2 > 0$, then for values of x *sufficiently close* to x_0 the second term becomes negligible small compared to the first one as

$$\frac{C_2|x-x_0|^{k+1}}{C_1|x-x_0|^k} = \frac{C_2}{C_1} \cdot |x-x_0| \to 0 \quad \text{as } x \to x_0.$$

More precisely, this ratio can be made as small as desired provided that x is in some sufficiently small open ball around x_0. This observation remains true independent of the particular values of C_1 and C_2. Only the diameter of this *"sufficiently small open ball"* may vary.

Such a situation where we want to describe local or asymptotic behavior of some function up to some non-specified constant is quite common in mathematics. For this purpose the so called *Landau symbol* is used.

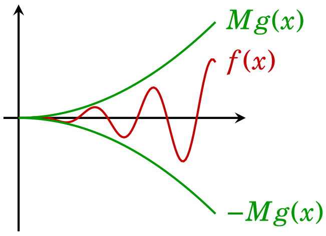
Definition 15.10

Landau symbol. Let $f(x)$ and $g(x)$ be two functions defined on some subset of $\mathbb{R}$. We write

$$f(x) = O\big(g(x)\big) \quad \text{as } x \to x_0 \qquad \text{(say "}f(x)\text{ is big O of }g\text{")}$$

if there exist positive numbers M and δ such that

$$|f(x)| \le M\,|g(x)| \quad \text{for all } x \text{ with } |x-x_0| < \delta.$$



By means of this notation we can write Taylor's formula with the Lagrange form of the remainder as

$$f(x) = \sum_{k=0}^{n} \frac{f^{(k)}(x_0)}{k!}(x-x_0)^k + O(|x-x_0|^{n+1})$$

(provided that f is $n+1$ times differentiable at x_0).

Observe that $f(x) = O\big(g(x)\big)$ implies that there exist positive numbers M and δ such that

$$\left|\frac{f(x)}{g(x)}\right| \le M \quad \text{for all } x \text{ with } |x-x_0| < \delta.$$

We also may have situations where we know that this fraction even converges to 0. Formally, we then write

$$f(x) = o\big(g(x)\big) \quad \text{as } x \to x_0 \qquad \text{(say "}f(x)\text{ is small O of }g\text{")}$$

if for every $\varepsilon > 0$ there exists a positive δ such that

$$|f(x)| \le \varepsilon|g(x)| \quad \text{for all } x \text{ with } |x-x_0| < \delta.$$

Using this notation we can write Taylor's Theorem 15.4 as

$$f(x) = \sum_{k=0}^{n} \frac{f^{(k)}(x_0)}{k!}(x-x_0)^k + o(|x-x_0|^n)$$

The symbols $O(\cdot)$ and $o(\cdot)$ are called **Landau symbols**.

The notation "$f(x) = O\big(g(x)\big)$" is a slight abuse of language as it merely indicates that f belongs to a family of functions that locally behaves similar to $g(x)$. Thus this is sometimes also expressed as

$$f(x) \in O\big(g(x)\big) .$$

15.4 Power Series and Analytic Functions

Taylor series are a special case of so called **power series**

$$p(x) = \sum_{n=1}^{\infty} a_n (x - x_0)^n .$$

Suppose that $\lim_{n\to\infty}\left|\frac{a_{n+1}}{a_n}\right|$ exists. Then the ratio test (Lemma 12.28) implies that the power series converges if

$$\lim_{n\to\infty}\left|\frac{a_{n+1}(x-x_0)^{n+1}}{a_n(x-x_0)^n}\right| = \lim_{n\to\infty}\left|\frac{a_{n+1}}{a_n}\right| |x-x_0| < 1$$

that is, if

$$|x-x_0| < \lim_{n\to\infty}\left|\frac{a_n}{a_{n+1}}\right| .$$

Similarly we find that the series diverges if

$$|x-x_0| > \lim_{n\to\infty}\left|\frac{a_n}{a_{n+1}}\right| .$$

Example 15.11 For the exponential function in Example 15.2 we find $a_n = 1/n!$. Thus

$$\lim_{n\to\infty}\left|\frac{a_n}{a_{n+1}}\right| = \lim_{n\to\infty}\frac{(n+1)!}{n!} = \lim_{n\to\infty} n+1 = \infty .$$

Hence the Taylor series converges for all $x \in \mathbb{R}$. ♢

Example 15.12 For function $f(x) = \ln(1+x)$ the situation is different. Recall that for function $f(x) = \ln(1+x)$, we find $a_n = (-1)^{n+1}/n$ (see Example 15.3).

$$\lim_{n\to\infty}\left|\frac{a_n}{a_{n+1}}\right| = \lim_{n\to\infty}\frac{n+1}{n} = 1 .$$

Hence the Taylor series converges for all $x \in (-1,1)$; and diverges for $x > 1$ or $x < -1$.

For $x = -1$ we get the divergent harmonic series, see Lemma 12.21. For $x = 1$ we get the convergent alternating harmonic series, see Lemma 12.25. However, a proof requires more sophisticated methods. ♢

Example 15.12 demonstrates that a Taylor series need not converge for all $x \in \mathbb{R}$. Instead there is a maximal distance ρ such that the series converges for all $x \in B_\rho(x_0)$ but diverges for all x with $|x - x_0| > \rho$. The value ρ is called the **radius of convergence** of the power series. Table 15.7 also lists this radius for the given Maclaurin series. $\rho = \infty$ means that the series converges for all $x \in \mathbb{R}$.

There is, however, a subtle difference between Examples 15.9 and 15.11. In the first example we have show that $\sum_{n=0}^\infty \frac{f^{(n)}(0)}{n!} x^n = e^x$ for all $x \in \mathbb{R}$ while in the latter we have just shown that $\sum_{n=0}^\infty \frac{f^{(n)}(0)}{n!} x^n$ converges. Similarly, we have shown in Example 15.12 that the Taylor series which we have computed in Example 15.3 converges, but we have not given a proof that $\sum_{k=1}^n (-1)^{k+1} \frac{x^k}{k} = \ln(1 + x)$.

Indeed functions f exist where the Taylor series converge but do not coincide with $f(x)$.

Example 15.13

The function

$$f(x) = \begin{cases} \exp\left(-\frac{1}{x^2}\right), & \text{for } x \neq 0, \\ 0, & \text{for } x = 0. \end{cases}$$

is infinitely differentiable in $x = 0$ and $f^{(n)}(0) = 0$ for all $n \in \mathbb{N}$ (see Problem 14.30). Consequently, we find for all Maclaurin polynomials $T_{f,n,0}(x) = 0$ for all $x \in \mathbb{R}$. Thus the Maclaurin series converges to 0 for all $x \in \mathbb{R}$. However, $f(x) > 0$ for all $x \neq 0$, i.e., albeit the series converges we find $\sum_{n=0}^\infty \frac{f^{(n)}(0)}{n!} x^n \neq f(x)$. ♢

Definition 15.14

Analytic function. An infinitely differentiable function f is called **analytic** in an open interval $B_r(x_0)$ if its Taylor series around x_0 converges and

$$f(x) = \sum_{n=0}^\infty \frac{f^{(n)}(x_0)}{n!} (x - x_0)^n \quad \text{for all } x \in B_r(x_0).$$

15.5 Defining Functions

Computations with power series are quite straightforward. Power series can be

- added or subtracted termwise,
- multiplied,
- divided,
- differentiated and integrated termwise.

Example 15.15

We get the Maclaurin series of the exponential function by differentiating the Maclaurin series of e^x:

$$\big(\exp(x)\big)' = \left(\sum_{n=0}^{\infty} \frac{1}{n!} x^n\right)' = \sum_{n=0}^{\infty} \frac{1}{n!} \big(x^n\big)' = \sum_{n=1}^{\infty} \frac{n}{n!} x^{n-1} = \sum_{n=1}^{\infty} \frac{1}{(n-1)!} x^{n-1}$$
$$= \sum_{n=0}^{\infty} \frac{1}{n!} x^n = \exp(x)\,.$$

Example 15.16

We get the Maclaurin series of $f(x) = x^2{\cdot}\sin(x)$ by multiplying the Maclaurin series of $\sin(x)$ by x^2:

$$x^2 \cdot \sin(x) = x^2 \cdot \sum_{n=0}^{\infty} (-1)^n \frac{x^{2n+1}}{(2n+1)!} = \sum_{n=0}^{\infty} (-1)^n x^2 \frac{x^{2n+1}}{(2n+1)!}$$
$$= \sum_{n=0}^{\infty} (-1)^n \frac{x^{2n+3}}{(2n+1)!}\,.$$

We also can substitute x in the Maclaurin series from Table 15.7 by some polynomial.

Example 15.17

We obtain the Maclaurin series of $\exp(-x^2)$ by substituting $-x^2$ into the Maclaurin series of the exponential function.

$$\exp(-x^2) = \sum_{n=0}^{\infty} \frac{1}{n!} \big(-x^2\big)^n = \sum_{n=0}^{\infty} \frac{(-1)^n}{n!} x^{2n}\,.$$

For that reason it is quite convenient to define analytic functions by its Taylor series.

$$\exp(x) := \sum_{n=0}^{\infty} \frac{1}{n!} x^n$$

15.6 Taylor's Formula for Multivariate Functions

Taylor polynomials can also be established for multivariate functions. We then construct a polynomial where all its kth order partial derivatives coincide with the corresponding partial derivatives of f at some given expansion point $\mathbf{x}_0$.

In opposition to the univariate case the number of coefficients of a polynomial in two or more variables increases exponentially in the degree of the polynomial. Thus we restrict our interest to the 2nd order Taylor polynomials which can be written as

$$p_2(x_1,\dots,x_n) = a_0 + \sum_{i=1}^{n} a_i x^i + \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x^i x^j$$

or, using vectors and quadratic forms,

$$p_2(\mathbf{x}) = a_0 + \mathbf{a}'\mathbf{x} + \mathbf{x}'\mathbf{A}\mathbf{x}$$

where $\mathbf{A}$ is an $n \times n$ matrix with $[\mathbf{A}]_{ij} = a_{ij}$ and $\mathbf{a}' = (a_1,\dots,a_n)$.

If we choose the coefficients a_i and a_{ij} such that all first and second order partial derivatives of p_2 at $\mathbf{x}_0 = 0$ coincides with the corresponding derivatives of f we find,

$$p_2(\mathbf{x}) = f(0) + f'(0)\mathbf{x} + \frac{1}{2}\mathbf{x}' f''(0)\mathbf{x}$$

For a general expand point $\mathbf{x}_0$ we get the following analog to Taylor's Theorem 15.4 which we state without proof.

Theorem 15.18

Taylor's formula for multivariate functions. Suppose that f is a $\mathscr{C}^3$ function in an open set containing the line segment $[\mathbf{x}^0, \mathbf{x}^0 + \mathbf{h}]$. Then

$$f(\mathbf{x}_0 + \mathbf{h}) = f(\mathbf{x}_0) + f'(\mathbf{x}_0) \cdot \mathbf{h} + \frac{1}{2}\mathbf{h}' \cdot f''(\mathbf{x}_0) \cdot \mathbf{h} + O(\|\mathbf{h}\|^3)\,.$$

Example 15.19

Let $f(x,y) = e^{x^2-y^2} + x$. Then gradient and Hessian matrix are given by

$$f'(x,y) = \left(2x\,e^{x^2-y^2} + 1,\; -2y\,e^{x^2-y^2}\right)$$

$$f''(x,y) = \begin{pmatrix} (2+4x^2)\,e^{x^2-y^2} & -4xy\,e^{x^2-y^2} \\ -4xy\,e^{x^2-y^2} & (-2+4x^2)\,e^{x^2-y^2} \end{pmatrix}$$

and thus we get for the 2nd order Taylor polynomial around $\mathbf{x}_0 = 0$

$$\begin{aligned} f(x,y) &= f(0,0) + f'(0,0)(x,y)' + \frac{1}{2}(x,y)f''(0,0)(x,y)' + O(\|(x,y)\|^3) \\ &= 1 + (1,0)(x,y)' + \frac{1}{2}(x,y)\begin{pmatrix} 2 & 0 \\ 0 & -2 \end{pmatrix}(x,y)' + O(\|(x,y)\|^3) \\ &= 1 + x + x^2 - y^2 + O(\|(x,y)\|^3)\,. \end{aligned}$$

— Exercises

15.1 Expand $f(x) = \frac{1}{2-x}$ into a Maclaurin polynomial of

(a) first order;

(b) second order.

Draw the graph of $f(x)$ and of these two Maclaurin polynomials in the interval $[-3,5]$.
Give an estimate for the radius of convergence.

15.2 Expand $f(x) = (x+1)^{1/2}$ into the 3rd order Taylor polynomial around $x_0 = 0$.

15.3 Expand $f(x) = \sin(x^{10})$ into a Maclaurin polynomial of degree 30.

15.4 Expand $f(x) = \sin(x^2 - 5)$ into a Maclaurin polynomial of degree 4.

15.5 Expand $f(x) = 1/(1 + x^2)$ into a Maclaurin series. Compute its radius of convergence.

15.6 Expand the density of the standard normal distribution $f(x) = \exp\left(-\frac{x^2}{2}\right)$ into a Maclaurin series. Compute its radius of convergence.

15.7 Expand $f(x,y) = e^{x^2+y^2}$ into a 2nd order Taylor series around $\mathbf{x}_0 = (0,0)$.

— Problems

15.8 Expand the exponential function $\exp(x)$ into a Taylor series about $x_0 = 0$. Give an upper bound of the remainder $R_n(1)$ as a function of order n. When is this bound less than 10^{-16}?

15.9 Assume that f is n times differentiable in x_0. Show that for the first n derivative of f and of its n-order Taylor polynomial coincide in x_0, i.e.,

$$\left(T_{f,x_0,n}\right)^{(k)}(x_0) = f^{(k)}(x_0)\,, \quad \text{for all } k = 0,\dots,n.$$

15.10 Verify the Maclaurin series from Table 15.7.

15.11 Show by means of the Maclaurin series from Table 15.7 that

(a) $(\sin(x))' = \cos(x)$ (b) $\left(\ln(1+x)\right)' = \frac{1}{1+x}$

16

Inverse and Implicit Functions

Can we invert the action of some function?

16.1 Inverse Functions

Inverse function. Let $\mathbf{f}\colon D_f \subseteq \mathbb{R}^n \to W_f \subseteq \mathbb{R}^m, \mathbf{x} \mapsto \mathbf{y} = \mathbf{f}(\mathbf{x})$ be some function. Suppose that there exists a function $\mathbf{f}^{-1}\colon W_f \to D_f, \mathbf{y} \mapsto \mathbf{x} = \mathbf{f}^{-1}(\mathbf{y})$ such that Definition 16.1

$$\mathbf{f}^{-1} \circ \mathbf{f} = \mathbf{f} \circ \mathbf{f}^{-1} = \mathbf{id}$$

that is, $\mathbf{f}^{-1}(\mathbf{f}(\mathbf{x})) = \mathbf{f}^{-1}(\mathbf{y}) = \mathbf{x}$ for all $\mathbf{x} \in D_f$, and $\mathbf{f}(\mathbf{f}^{-1}(\mathbf{y})) = \mathbf{f}(\mathbf{x}) = \mathbf{y}$ for all $\mathbf{y} \in W_f$. Then $\mathbf{f}^{-1}$ is called the **inverse function** of $\mathbf{f}$.

Obviously, the inverse function exists if and only if $\mathbf{f}$ is a *bijection*. Lemma 16.2

We get the function term of the inverse function by solving equation $\mathbf{y} = \mathbf{f}(\mathbf{x})$ w.r.t. to $\mathbf{x}$.

Affine function. Suppose that $\mathbf{f}\colon \mathbb{R}^n \to \mathbb{R}^m, \mathbf{x} \mapsto \mathbf{y} = \mathbf{f}(\mathbf{x}) = \mathbf{A}\mathbf{x} + \mathbf{b}$ where $\mathbf{A}$ is an $m \times n$ matrix and $\mathbf{b} \in \mathbb{R}^m$. Then we find Example 16.3

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{b} \quad \Leftrightarrow \quad \mathbf{x} = \mathbf{A}^{-1}\mathbf{y} - \mathbf{A}^{-1}\mathbf{b} = \mathbf{A}^{-1}(\mathbf{y} - \mathbf{b})$$

provided that $\mathbf{A}$ is invertible. In particular we must have $n = m$. Thus we have $\mathbf{f}^{-1}(\mathbf{y}) = \mathbf{A}^{-1}(\mathbf{y} - \mathbf{b})$. Observe that

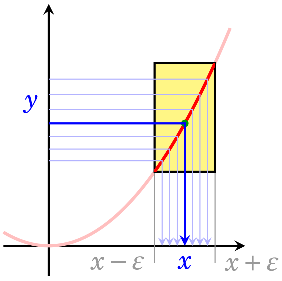
$$D\mathbf{f}^{-1}(\mathbf{y}) = \mathbf{A}^{-1} = (D\mathbf{f}(\mathbf{x}))^{-1} . \qquad \diamondsuit$$

For an arbitrary function the inverse need not exist. E.g., the function $f : \mathbb{R} \to \mathbb{R}, x \mapsto x^2$ is not invertible. However, if we restrict the domain of our function to some (sufficiently small) open interval $D = B_\varepsilon(x_0) \subset$

$(0,\infty)$ then the inverse exists. Motivated by Example 16.3 above we expect that this always works whenever $f'(x_0) \neq 0$, i.e., when $\frac{1}{f'(x_0)}$ exists. Moreover, it is possible to compute the derivative $\left(f^{-1}\right)'$ of its inverse in $y_0 = f(x_0)$ as $\left(f^{-1}\right)'(y_0) = \frac{1}{f'(x_0)}$ without having an explicit expression for f^{-1}.

This useful fact is stated in the inverse function theorem.

Theorem 16.4



Inverse function theorem. Let $\mathbf{f}\colon D \subseteq \mathbb{R}^n \to \mathbb{R}^n$ be a $\mathscr{C}^k$ function in some open set D containing $\mathbf{x}^0$. Suppose that the **Jacobian determinant** of $\mathbf{f}$ at $\mathbf{x}^0$ is nonzero, i.e.,

$$\frac{\partial(f_1,\dots,f_n)}{\partial(x_1,\dots,x_n)} = \left|\mathbf{f}'(\mathbf{x}^0)\right| \neq 0 \quad \text{for } \mathbf{x} = \mathbf{x}^0.$$

Then there exists an open set U around $\mathbf{x}^0$ such that $\mathbf{f}$ maps U one-to-one onto an open set V around $\mathbf{y}^0 = \mathbf{f}(\mathbf{x}^0)$. Thus there exists an inverse mapping $\mathbf{f}^{-1}\colon V \to U$ which is also in $\mathscr{C}^k$. Moreover, for all $\mathbf{y} \in V$, we have

$$(\mathbf{f}^{-1})'(\mathbf{y}^0) = (\mathbf{f}'(\mathbf{x}^0))^{-1}.$$

In other words, a $\mathscr{C}^k$ function $\mathbf{f}$ with a nonzero Jacobian determinant at $\mathbf{x}^0$ has a *local inverse* around $\mathbf{f}(\mathbf{x}^0)$ which is again $\mathscr{C}^k$.

This theorem is an immediate corollary of the Implicit Function Theorem 16.11 below, see Problem 16.10. The idea behind the proof is that we can locally replace function $\mathbf{f}$ by its differential in order to get its local inverse.

For the case $n = 1$, that is, a function $f\colon \mathbb{R} \to \mathbb{R}$, we find

$$(f^{-1})'(y_0) = \frac{1}{f'(x_0)} \qquad \text{where } y_0 = f(x_0).$$

Example 16.5

Let $f\colon \mathbb{R} \to \mathbb{R}, x \mapsto y = f(x) = x^2$ and $x_0 = 3$. Then $f'(x_0) = 6 \neq 0$ thus f^{-1} exists in open ball around $y_0 = f(x_0) = 9$. Moreover

$$(f^{-1})'(9) = \frac{1}{f'(3)} = \frac{1}{6}.$$

We remark here that Theorem 16.4 does *not* imply that function f^{-1} does not exist in any open ball around $f(0)$. As $f'(0) = 0$ we simply cannot apply the theorem in this case. ♢

Example 16.6

Let $\mathbf{f}\colon \mathbb{R}^2 \to \mathbb{R}^2, \mathbf{x} \mapsto \mathbf{f}(\mathbf{x}) = \begin{pmatrix} x_1^2 - x_2^2 \\ x_1 x_2 \end{pmatrix}$. Then we find $D\mathbf{f}(\mathbf{x}) = \begin{pmatrix} 2x_1 & -2x_2 \\ x_2 & x_1 \end{pmatrix}$ and thus

$$\frac{\partial(f_1,f_2)}{\partial(x_1,x_2)} = |D\mathbf{f}(\mathbf{x})| = \begin{vmatrix} 2x_1 & -2x_2 \\ x_2 & x_1 \end{vmatrix} = 2x_1^2 + 2x_2^2 \neq 0$$

for all $\mathbf{x} \neq 0$. Consequently, $\mathbf{f}^{-1}$ exists around all $\mathbf{y} = \mathbf{f}(\mathbf{x})$ where $\mathbf{x} \neq 0$. The derivative at $\mathbf{y} = \mathbf{f}(1,1)$ is given by

$$D(\mathbf{f}^{-1})(\mathbf{y}) = (D\mathbf{f}(1,1))^{-1} = \begin{pmatrix} 2 & -2 \\ 1 & 1 \end{pmatrix}^{-1} = \begin{pmatrix} \frac{1}{4} & \frac{2}{4} \\ -\frac{1}{4} & \frac{2}{4} \end{pmatrix}.$$ ♢

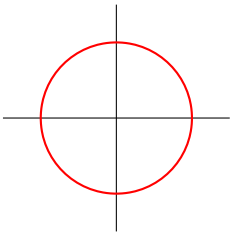
16.2 Implicit Functions

Suppose we are given some function $F(x,y)$. Then equation $F(x,y)=0$ describes a relation between the two variables x and y. Then if we fix x then y is implicitly given. Thus we call this an **implicit function**. One may ask the question whether it is possible to express y as an *explicit function* of x.

Example 16.7

Linear function. Let $F(x,y)=ax+by=0$ for $a,b\in\mathbb{R}$. Then we easily find $y=f(x)=-\frac{a}{b}x$ provided that $b\neq 0$. Observe that $F_x=a$ and $F_y=b$. Thus we find

$$y=-\frac{F_x}{F_y}x \quad\text{and}\quad \frac{dy}{dx}=-\frac{F_x}{F_y} \qquad \text{provided that } F_y\neq 0. \qquad \diamondsuit$$



For non-linear functions this need not work. E.g., for

$$F(x,y)=x^2+y^2-1=0$$

it is not possible to globally express y as a function of x. Nevertheless, we may try to find such an explicit expression that works locally, i.e., within an open rectangle around a given point (x_0,y_0) that satisfies this equation. Thus we replace F locally by its total derivative

$$dF=F_x dx+F_y dy=d0=0$$

and obtain formally the derivative

$$\frac{dy}{dx}=-\frac{F_x}{F_y}\,.$$

Obviously this only works when $F_y(x_0,y_0)\neq 0$.

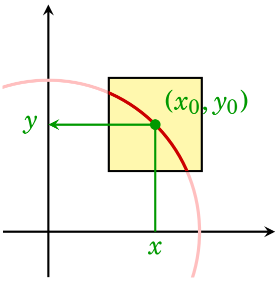
Theorem 16.8

Implicit function theorem. Let $F\colon\subseteq\mathbb{R}^2\to\mathbb{R}$ be a differentiable function in some open set D. Consider an interior point $(x_0,y_0)\in D$ where

$$F(x_0,y_0)=0 \qquad\text{and}\qquad F_y(x_0,y_0)\neq 0\,.$$

Then there exists an open rectangle R around (x_0,y_0), such that

- $F(x,y)=0$ has a unique solution $y=f(x)$ in R, and
- $\dfrac{dy}{dx}=-\dfrac{F_x}{F_y}\,.$



Example 16.9

Let $F(x,y)=x^2+y^2-8=0$ and $(x_0,y_0)=(2,2)$. Since $F(x_0,y_0)=0$ and $F_y(x_0,y_0)=2y_0=4\neq 0$, there exists a rectangle R around $(2,2)$ such that y can be expressed as an explicit function of x and we find

$$\frac{dy}{dx}(x_0)=-\frac{F_x(x_0,y_0)}{F_y(x_0,y_0)}=-\frac{2x_0}{2y_0}=-\frac{4}{4}=-1\,.$$

Observe that we cannot apply Theorem 16.8 for the point $(\sqrt{8}, 0)$ as then $F_y(\sqrt{8}, 0) = 0$. Thus the hypothesis of the theorem is violated. Notice, however, that this does *not* necessarily imply that the requested local explicit function does not exist at all. ◊

We can generalize Theorem 16.8 to the functions with arbitrary numbers of arguments. Thus we first need a generalization of the partial derivative.

Definition 16.10

Jacobian matrix. Let $\mathbf{F}\colon \mathbb{R}^{n+m} \to \mathbb{R}^m$ be a differentiable function with

$$(\mathbf{x}, \mathbf{y}) \mapsto \mathbf{F}(\mathbf{x}, \mathbf{y}) = \begin{pmatrix} F_1(x_1, \dots, x_n, y_1, \dots, y_m) \\ \vdots \\ F_m(x_1, \dots, x_n, y_1, \dots, y_m) \end{pmatrix}$$

Then the matrix

$$\frac{\partial \mathbf{F}(\mathbf{x}, \mathbf{y})}{\partial \mathbf{y}} = \begin{pmatrix} \frac{\partial F_1}{\partial y_1} & \cdots & \frac{\partial F_1}{\partial y_m} \\ \vdots & \ddots & \vdots \\ \frac{\partial F_m}{\partial y_1} & \cdots & \frac{\partial F_m}{\partial y_m} \end{pmatrix}$$

is called the **Jacobian matrix** of $\mathbf{F}(\mathbf{x}, \mathbf{y})$ w.r.t. $\mathbf{y}$.

Theorem 16.11

Implicit function theorem. Let $\mathbf{F}\colon D \subseteq \mathbb{R}^{n+m} \to \mathbb{R}^m$ be $\mathscr{C}^k$ in some open set D. Consider an interior point $(\mathbf{x}^0, \mathbf{y}^0) \in D$ where

$$\mathbf{F}(\mathbf{x}_0, \mathbf{y}_0) = 0 \qquad \text{and} \qquad \left| \frac{\partial \mathbf{F}(\mathbf{x}, \mathbf{y})}{\partial \mathbf{y}} \right| \neq 0 \quad \text{for } (\mathbf{x}, \mathbf{y}) = (\mathbf{x}_0, \mathbf{y}_0).$$

Then there exist open balls $B(\mathbf{x}^0) \subseteq \mathbb{R}^n$ and $B(\mathbf{y}^0) \subseteq \mathbb{R}^m$ around $\mathbf{x}^0$ and $\mathbf{y}^0$, respectively, with $B(\mathbf{x}^0) \times B(\mathbf{y}^0) \subseteq D$ such that for every $\mathbf{x} \in B(\mathbf{x}^0)$ there exists a unique $\mathbf{y} \in B(\mathbf{y}^0)$ with $\mathbf{F}(\mathbf{x}, \mathbf{y}) = 0$. In this way we obtain a $\mathscr{C}^k$ function $\mathbf{f}\colon B(\mathbf{x}^0) \subseteq \mathbb{R}^n \to B(\mathbf{y}^0) \subseteq \mathbb{R}^m$ with $\mathbf{f}(\mathbf{x}) = \mathbf{y}$. Moreover,

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = -\left(\frac{\partial \mathbf{F}}{\partial \mathbf{y}} \right)^{-1} \cdot \left(\frac{\partial \mathbf{F}}{\partial \mathbf{x}} \right)$$

The proof of this theorem requires tools from advanced calculus which are beyond the scope of this course. Nevertheless, the rule for the derivative for the local inverse function (if it exists) can be easily derived by means of the chain rule, see Problem 16.12.

Obviously, Theorem 16.8 is just a special case of Theorem 16.11. For the special case where $F\colon \mathbb{R}^{n+1} \to \mathbb{R}$, $(\mathbf{x}, y) \mapsto F(\mathbf{x}, y) = F(x_1, \dots, x_n, y)$, and some point $(\mathbf{x}_0, y)$ with $F(\mathbf{x}_0, y_0) = 0$ and $F_y(\mathbf{x}_0, y_0) \neq 0$ we then find that there exists an open rectangle around $(\mathbf{x}_0, y)$ such that $y = f(\mathbf{x})$ and

$$\frac{\partial y}{\partial x_i} = -\frac{F_{x_i}}{F_y}.$$

Example 16.12

Let

$$F(x_1,x_2,x_3,x_4) = x_1^2 + x_2 x_3 + x_3^2 - x_3 x_4 - 1 = 0\,.$$

We are given a point $(x_1,x_2,x_3,x_4) = (1,0,1,1)$. We find $F(1,0,1,1) = 0$ and $F_{x_2}(1,0,1,1) = 1 \neq 0$. Thus there exists an open rectangle where x_2 can be expressed locally by an explicit function of the remaining variables, $x_2 = f(x_1,x_3,x_4)$, and we find for the partial derivative w.r.t. in $(x_1,x_3,x_4) = (1,1,1)$,

$$\frac{\partial x_2}{\partial x_3} = -\frac{F_{x_3}}{F_{x_2}} = -\frac{x_2 + 2x_3 - x_4}{x_3} = -1\,.$$

Notice that we cannot apply the Implicit Function Theorem neither at $(1,1,1,1)$ nor at $(1,1,0,1)$ as $F(1,1,1,1) \neq 0$ and $F_{x_2}(1,1,0,1) = 0$, respectively. ♢

Example 16.13

Let

$$\mathbf{F}(\mathbf{x},\mathbf{y}) = \begin{pmatrix} F_1(x_1,x_2,y_1,y_2) \\ F_2(x_1,x_2,y_1,y_2) \end{pmatrix} = \begin{pmatrix} x_1^2 + x_2^2 - y_1^2 - y_2^2 + 3 \\ x_1^3 + x_2^3 + y_1^3 + y_2^3 - 11 \end{pmatrix}$$

and some point $(\mathbf{x}_0,\mathbf{y}_0) = (1,1,1,2)$.

$$\frac{\partial \mathbf{F}}{\partial \mathbf{x}} = \begin{pmatrix} \frac{\partial F_1}{\partial x_1} & \frac{\partial F_1}{\partial x_2} \\ \frac{\partial F_2}{\partial x_1} & \frac{\partial F_2}{\partial x_2} \end{pmatrix} = \begin{pmatrix} 2x_1 & 2x_2 \\ 3x_1^2 & 3x_2^2 \end{pmatrix} \quad \text{and} \quad \frac{\partial \mathbf{F}}{\partial \mathbf{x}}(1,1,1,2) = \begin{pmatrix} 2 & 2 \\ 3 & 3 \end{pmatrix}$$

$$\frac{\partial \mathbf{F}}{\partial \mathbf{y}} = \begin{pmatrix} \frac{\partial F_1}{\partial y_1} & \frac{\partial F_1}{\partial y_2} \\ \frac{\partial F_2}{\partial y_1} & \frac{\partial F_2}{\partial y_2} \end{pmatrix} = \begin{pmatrix} -2y_1 & -2y_2 \\ 3y_1^2 & 3y_2^2 \end{pmatrix} \quad \text{and} \quad \frac{\partial \mathbf{F}}{\partial \mathbf{y}}(1,1,1,2) = \begin{pmatrix} -2 & -4 \\ 3 & 12 \end{pmatrix}$$

Since $\mathbf{F}(1,1,1,2) = 0$ and $\left|\frac{\partial \mathbf{F}(\mathbf{x},\mathbf{y})}{\partial \mathbf{y}}\right| = -12 \neq 0$ we can apply the Implicit Function Theorem and get

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = -\left(\frac{\partial \mathbf{F}}{\partial \mathbf{y}}\right)^{-1} \cdot \left(\frac{\partial \mathbf{F}}{\partial \mathbf{x}}\right) = -\frac{1}{-12}\begin{pmatrix} 12 & 4 \\ -3 & -2 \end{pmatrix} \cdot \begin{pmatrix} 2 & 2 \\ 3 & 3 \end{pmatrix} = \begin{pmatrix} 3 & 3 \\ 0 & 0 \end{pmatrix}.$$

♢

— Exercises

16.1 Let $\mathbf{f}\colon \mathbb{R}^2 \to \mathbb{R}^2$ be a function with

$$\mathbf{x} \mapsto \mathbf{f}(\mathbf{x}) = \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} x_1 - x_1 x_2 \\ x_1 x_2 \end{pmatrix}$$

(a) Compute the Jacobian matrix and determinant of **f**.

(b) Around which points is it possible to find a local inverse of **f**?

(c) Compute the Jacobian matrix for the inverse function.

(d) Compute the inverse function (where it exists).

16.2 Let $T\colon \mathbb{R}^2 \to \mathbb{R}^2$ be a function with

$$(x,y) \mapsto (u,v) = (ax+by, cx+dy)$$

where a, b, c, and d are non-zero constants.

Show: If the Jacobian determinant of T equals 0, then the image of T is a straight line through the origin.

16.3 Give a sufficient condition for f and g such that the equations

$$u = f(x,y), \quad v = g(x,y)$$

can be solved w.r.t. x and y.

Suppose we have the solutions $x = F(u,v)$ and $y = G(u,v)$. Compute $\frac{\partial F}{\partial u}$ and $\frac{\partial G}{\partial u}$.

16.4 Show that the following equations define y as a function of x in an interval around x_0. Compute $y'(x_0)$.

(a) $y^3 + y - x^3 = 0, \quad x_0 = 0$

(b) $x^2 + y + \sin(xy) = 0, \quad x_0 = 0$

16.5 Compute $\frac{dy}{dx}$ from the implicit function $x^2 + y^3 = 0$.
For which values of x does an explicit function $y = f(x)$ exist locally?
For which values of y does an explicit function $x = g(y)$ exist locally?

16.6 Which of the given implicit functions can be expressed as $z = g(x,y)$ in a neighborhood of the given point (x_0, y_0, z_0).
Compute $\frac{\partial g}{\partial x}$ and $\frac{\partial g}{\partial y}$.

(a) $x^3 + y^3 + z^3 - xyz - 1 = 0, \quad (x_0, y_0, z_0) = (0,0,1)$

(b) $\exp(z) - z^2 - x^2 - y^2 = 0, \quad (x_0, y_0, z_0) = (1,0,0)$

16.7 Compute the marginal rate of substitution of K for L for the following isoquant of the given production function, that is $\frac{dK}{dL}$:

$$F(K,L) = AK^\alpha L^\beta = F_0$$

16.8 Compute the derivative $\frac{dx_i}{dx_j}$ of the indifference curve of the utility function:

(a) $u(x_1,x_2) = \left(x_1^{\frac{1}{2}} + x_2^{\frac{1}{2}}\right)^2$

(b) $u(x_1,\dots,x_n) = \left(\sum_{i=1}^{n} x_i^{\frac{\theta-1}{\theta}}\right)^{\frac{\theta}{\theta-1}}$ $\qquad (\theta > 1)$

— Problems

16.9 Prove Lemma 16.2.

16.10 Derive Theorem 16.4 from Theorem 16.11.

HINT: Consider function $\mathbf{F}(\mathbf{x},\mathbf{y}) = \mathbf{f}(\mathbf{x}) - \mathbf{y} = 0$.

16.11 Does the inverse function theorem (Theorem 16.4) provide a necessary or a sufficient condition for the existence of a local inverse function or is the condition both necessary and sufficient?

If the condition is not necessary, give a counterexample.

If the condition is not sufficient, give a counterexample.

HINT: Use a function $f: \mathbb{R} \to \mathbb{R}$.

16.12 Let $\mathbf{f}: \mathbb{R}^n \to \mathbb{R}^n$ be a $\mathscr{C}^1$ function that has a local inverse function $\mathbf{f}^{-1}$ around some point $\mathbf{x}^0$. Show by means of the chain rule that

$$\left(\mathbf{f}^{-1}\right)'(\mathbf{y}^0) = \left(\mathbf{f}'(\mathbf{x}^0)\right)^{-1} \qquad \text{where } \mathbf{y}^0 = \mathbf{f}(\mathbf{x}^0).$$

HINT: Notice that $\mathbf{f}^{-1} \circ \mathbf{f} = \mathrm{id}$, where id denotes the identity function, i.e, $\mathrm{id}(\mathbf{x}) = \mathbf{x}$. Compute the derivatives on either side of the equation. Use the chain rule for the left hand side. What is the derivative of id?

17

Convex Functions

Is there a panoramic view over our entire function?

17.1 Convex Sets

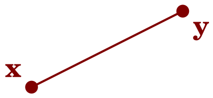
Definition 17.1

Convex set. A set $D \subseteq \mathbb{R}^n$ is called **convex**, if each pair of points $\mathbf{x}, \mathbf{y} \in D$ can be joints by a line segment lying entirely in D, i.e., if

$$(1-t)\mathbf{x} + t\mathbf{y} \in D \quad \text{for all } \mathbf{x}, \mathbf{y} \in D \text{ and all } t \in [0,1].$$

The line segment between $\mathbf{x}$ and $\mathbf{y}$ is the set

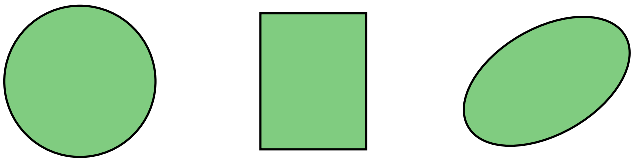
$$[\mathbf{x}, \mathbf{y}] = \left\{\mathbf{z} = (1-t)\mathbf{x} + t\mathbf{y} \colon t \in [0,1]\right\}.$$



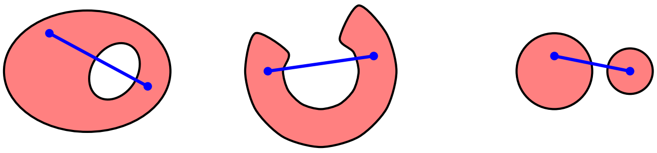
whose elements are so called **convex combinations** of $\mathbf{x}$ and $\mathbf{y}$. Hence $[\mathbf{x}, \mathbf{y}]$ is also called the **convex hull** of these points.

Example 17.2

The following sets are convex:



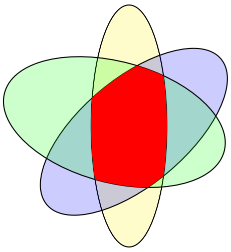
The following sets are not convex:



Theorem 17.3

Intersection. The intersection of convex sets is convex.

Proof. See Problem 17.7. □



Notice that the union of convex need not be convex.

Example 17.4

Half spaces. Let $\mathbf{p} \in \mathbb{R}^n$, $\mathbf{p} \neq 0$, and $m \in \mathbb{R}$. Then the set

$$H = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{p}' \cdot \mathbf{x} = m\}$$

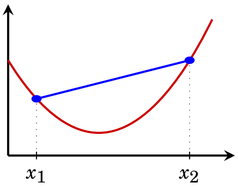
is called a **hyperplane** in $\mathbb{R}^n$. It divides $\mathbb{R}^n$ into two so called **half spaces**

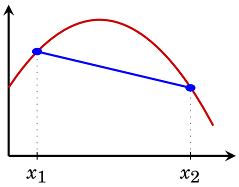
$$H_+ = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{p}' \cdot \mathbf{x} \geq m\} \quad \text{and} \quad H_- = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{p}' \cdot \mathbf{x} \leq m\} .$$

All sets H, H^+, and H^- are convex, see Problem 17.8. ♢

17.2 Convex and Concave Functions

Definition 17.5





Convex and concave function. A function $f : D \subseteq \mathbb{R}^n \to \mathbb{R}$ is **convex** if D is convex and

$$f\big((1-t)\mathbf{x}_1 + t\mathbf{x}_2\big) \leq (1-t)f(\mathbf{x}_1) + t f(\mathbf{x}_2)$$

for all $\mathbf{x}_1, \mathbf{x}_2 \in D$ and all $t \in [0,1]$. This is equivalent to the property that the set $\big\{(\mathbf{x}, y) \in \mathbb{R}^{n+1} : f(\mathbf{x}) \geq y\big\}$ is convex.

Function f is **concave** if D is convex and

$$f\big((1-t)\mathbf{x}_1 + t\mathbf{x}_2\big) \geq (1-t)f(\mathbf{x}_1) + t f(\mathbf{x}_2)$$

for all $\mathbf{x}_1, \mathbf{x}_2 \in D$ and all $t \in [0,1]$.

Notice that a function f is concave if and only if $-f$ is convex, see Problem 17.9.

Definition 17.6

Strictly convex function. A function $f : D \subseteq \mathbb{R}^n \to \mathbb{R}$ is **strictly convex** if D is convex and

$$f\big((1-t)\mathbf{x}_1 + t\mathbf{x}_2\big) < (1-t)f(\mathbf{x}_1) + t f(\mathbf{x}_2)$$

for all $\mathbf{x}_1, \mathbf{x}_2 \in D$ with $\mathbf{x}_1 \neq \mathbf{x}_2$ and all $t \in (0,1)$. Function f is **strictly concave** if this equation holds with "<" replaced by ">".

Example 17.7

Linear function. Let $\mathbf{a} \in \mathbb{R}^n$ be constant. Then $f(\mathbf{x}) = \mathbf{a}' \cdot \mathbf{x}$ is both convex and concave:

$$\begin{aligned} f\big((1-t)\mathbf{x}_1 + t\mathbf{x}_2\big) &= \mathbf{a}' \cdot \big((1-t)\mathbf{x}_1 + t\mathbf{x}_2\big) = (1-t)\mathbf{a}' \cdot \mathbf{x}_1 + t\,\mathbf{a}' \cdot \mathbf{x}_2 \\ &= (1-t)f(\mathbf{x}_1) + t f(\mathbf{x}_2) \end{aligned}$$

However, it is neither strictly convex nor strictly concave. ♢

Example 17.8

Quadratic function. Function $f(x) = x^2$ is strictly convex:

$$\begin{aligned} &f((1-t)x + ty) - \big[(1-t)f(x) + t f(y)\big] \\ &\qquad = \big((1-t)x + ty\big)^2 - \big[(1-t)x^2 + ty^2\big] \\ &\qquad = (1-t)^2x^2 + 2(1-t)txy + t^2y^2 - (1-t)x^2 - ty^2 \\ &\qquad = -t(1-t)x^2 + 2(1-t)txy - t(1-t)y^2 \\ &\qquad = -t(1-t)(x-y)^2 < 0 \end{aligned}$$

for $x \neq y$ and $0 < t < 1$. ♢

Convex sum. Let $\alpha_1,\dots,\alpha_k > 0$. If $f_1(\mathbf{x}),\dots,f_k(\mathbf{x})$ are convex (concave) functions, then Theorem 17.9

$$g(\mathbf{x}) = \sum_{i=1}^{k} \alpha_i f_i(\mathbf{x})$$

is convex (concave). Function $g(\mathbf{x})$ is strictly convex (strictly concave) if at least one of the functions $f_i(\mathbf{x})$ is strictly convex (strictly concave).

PROOF. See Problem 17.12. □

An immediate consequence of this theorem and Example 17.8 is that a quadratic function $f(x) = ax^2 + bx + c$ is strictly convex if $a > 0$, strictly concave if $a < 0$ and both convex and concave if $a = 0$.

Quadratic form. Let $\mathbf{A}$ be a symmetric $n \times n$ matrix. Then quadratic form $q(\mathbf{x}) = \mathbf{x}'\mathbf{A}\mathbf{x}$ is strictly convex if and only if $\mathbf{A}$ is positive definite. It is convex if and only if $\mathbf{A}$ is positive semidefinite. Theorem 17.10

Similarly, q is strictly concave if and only if $\mathbf{A}$ is negative definite. It is concave if and only if $\mathbf{A}$ is negative semidefinite.

PROOF IDEA. We first show by a straightforward computation that the univariate function $g(t) = q\big((1-t)\mathbf{x}_1 + t\mathbf{x}_2\big)$ is strictly convex for all $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^n$ if and only if $\mathbf{A}$ is positive definite.

PROOF. Let $\mathbf{x}_1$ and $\mathbf{x}_2$ be two distinct points in $\mathbb{R}^n$. Then

$$\begin{aligned} g(t) &= q\big((1-t)\mathbf{x}_1 + t\mathbf{x}_2\big) = q\big(\mathbf{x}_1 + t(\mathbf{x}_2 - \mathbf{x}_1)\big) \\ &= \big(\mathbf{x}_1 + t(\mathbf{x}_2 - \mathbf{x}_1)\big)'\mathbf{A}\big(\mathbf{x}_1 + t(\mathbf{x}_2 - \mathbf{x}_1)\big) \\ &= t^2(\mathbf{x}_2 - \mathbf{x}_1)'\mathbf{A}(\mathbf{x}_2 - \mathbf{x}_1) + 2t(\mathbf{x}_1'\mathbf{A}\mathbf{x}_2 - \mathbf{x}_1'\mathbf{A}\mathbf{x}_1) + \mathbf{x}_1'\mathbf{A}\mathbf{x}_1 \\ &= q(\mathbf{x}_1 - \mathbf{x}_2)\,t^2 + 2(\mathbf{x}_1'\mathbf{A}\mathbf{x}_2 - q(\mathbf{x}_1))\,t + q(\mathbf{x}_1) \end{aligned}$$

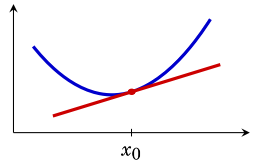
is a quadratic function in t which is strictly convex if and only if $q(\mathbf{x}_1 - \mathbf{x}_2) > 0$. This is the case for each pair of points $\mathbf{x}_1$ and $\mathbf{x}_2$ if and only if $\mathbf{A}$ is positive definite. We then find

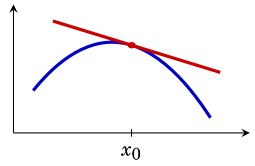
$$\begin{aligned} q\big((1-t)\mathbf{x}_1 + t\mathbf{x}_2\big) &= g(t) = g\big((1-t)\,0 + t\,1\big) \\ &> (1-t)g(0) + tg(1) = (1-t)q(\mathbf{x}_1) + tq(\mathbf{x}_2) \end{aligned}$$

for all $t \in (0,1)$ and hence q is strictly convex as well. The cases where q is convex and (strictly) concave follow analogously. □

Recall from Linear Algebra that we can determine the definiteness of a symmetric matrix $\mathbf{A}$ by means of the signs of its eigenvalues or by the signs of (leading) principle minors.

Theorem 17.11





Tangents of convex functions. A $\mathscr{C}^1$ function f is convex in an open, convex set D if and only if

$$f(\mathbf{x}) - f(\mathbf{x}_0) \geq \nabla f(\mathbf{x}_0) \cdot (\mathbf{x} - \mathbf{x}_0) \tag{17.1}$$

for all $\mathbf{x}$ and $\mathbf{x}_0$ in D, i.e., the tangent is always below the function. Function f is strictly convex if and only if inequality (17.1) is strict for $\mathbf{x} \neq \mathbf{x}_0$.

A $\mathscr{C}^1$ function f is concave in an open, convex set D if and only if

$$f(\mathbf{x}) - f(\mathbf{x}_0) \leq \nabla f(\mathbf{x}_0) \cdot (\mathbf{x} - \mathbf{x}_0)$$

for all $\mathbf{x}$ and $\mathbf{x}_0$ in D, i.e., the tangent is always above the function.

PROOF IDEA. For the necessity of condition (17.1) we transform the inequality for convexity (see Definition 17.5) into an inequality about difference quotients and apply the Mean Value Theorem. Using continuity of the gradient of f yields inequality (17.1).

We note here that for the case of strict convexity we need some technical trick to obtain the requested strict inequality.

For sufficiency we split an interval $[\mathbf{x}_0, \mathbf{x}]$ into two subintervals $[\mathbf{x}_0, \mathbf{z}]$ and $[\mathbf{z}, \mathbf{x}]$ and apply inequality (17.1) on each.

PROOF. Assume that f is convex, and let $\mathbf{x}_0, \mathbf{x} \in D$. Then we have by definition

$$f\big((1-t)\mathbf{x}_0 + t\mathbf{x}\big) \leq (1-t) f(\mathbf{x}_0) + t f(\mathbf{x})$$

and thus

$$f(\mathbf{x}) - f(\mathbf{x}_0) \geq \frac{f\big(\mathbf{x}_0 + t(\mathbf{x} - \mathbf{x}_0)\big) - f(\mathbf{x}_0)}{t} = \nabla f(\xi(t)) \cdot (\mathbf{x} - \mathbf{x}_0)$$

by the mean value theorem (Theorem 14.10) where $\xi(t) \in [\mathbf{x}_0, \mathbf{x}_0 + t(\mathbf{x} - \mathbf{x}_0)]$. (Notice that the central term is the difference quotient corresponding to the directional derivative.) Since f is a $\mathscr{C}^1$ function we find

$$f(\mathbf{x}) - f(\mathbf{x}_0) \geq \lim_{t \to 0} \nabla f(\xi(t)) \cdot (\mathbf{x} - \mathbf{x}_0) = \nabla f(\mathbf{x}_0) \cdot (\mathbf{x} - \mathbf{x}_0)$$

as claimed.

Conversely assume that (17.1) holds for all $\mathbf{x}_0, \mathbf{x} \in D$. Let $t \in [0,1]$ and $\mathbf{z} = (1-t)\mathbf{x}_0 + t\mathbf{x}$. Then $\mathbf{z} \in D$ and by (17.1) we find

$$\begin{aligned} (1-t)\big(f(\mathbf{x}_0) - f(\mathbf{z})\big) &+ t\big(f(\mathbf{x}) - f(\mathbf{z})\big) \\ &\geq (1-t)\,\nabla f(\mathbf{z})(\mathbf{x}_0 - \mathbf{z}) + t\,\nabla f(\mathbf{z})(\mathbf{x} - \mathbf{z}) \\ &= \nabla f(\mathbf{z})\big((1-t)\mathbf{x}_0 + t\mathbf{x} - \mathbf{z}\big) = \nabla f(\mathbf{z})\,0 = 0\,. \end{aligned}$$

Consequently,

$$(1-t)f(\mathbf{x}_0) - tf(\mathbf{x}) \geq f(\mathbf{z}) = f\big((1-t)\mathbf{x}_0 + t\mathbf{x}\big)$$

and thus f is convex.

The proof for the case where f is strictly convex is analogous. However, in the first part of the proof $f(\mathbf{x})-f(\mathbf{x}_0)>\nabla f(\xi(t))\cdot(\mathbf{x}-\mathbf{x}_0)$ does not imply strict inequality in

$$f(\mathbf{x})-f(\mathbf{x}_0)\geq\lim_{t\to 0}\nabla f(\xi(t))\cdot(\mathbf{x}-\mathbf{x}_0)\,.$$

So we need a technical trick. Assume $\mathbf{x}\neq\mathbf{x}_0$ and let $\mathbf{x}_1=(\mathbf{x}+\mathbf{x}_0)/2$. By strict convexity of $\mathbf{f}$ we have $f(\mathbf{x}_1)<\frac{1}{2}(f(\mathbf{x})+f(\mathbf{x}_0))$. Hence we find

$$2(\mathbf{x}_1-\mathbf{x}_0)=\mathbf{x}-\mathbf{x}_0 \quad\text{and}\quad 2(f(\mathbf{x}_1)-f(\mathbf{x}_0))<f(\mathbf{x})-f(\mathbf{x}_0)$$

and thus

$$f(\mathbf{x})-f(\mathbf{x}_0)>2(f(\mathbf{x}_1)-f(\mathbf{x}_0))\geq 2\nabla f(\mathbf{x}_0)\cdot(\mathbf{x}_1-\mathbf{x}_0)=\nabla f(\mathbf{x}_0)\cdot(\mathbf{x}-\mathbf{x}_0)$$

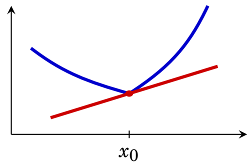
as claimed. □

There also exists a version for functions that are not necessarily differentiable.

Theorem 17.12

Subgradient and supergradient. Let f be a convex function on a convex set $D\subseteq\mathbb{R}^n$, and let $\mathbf{x}_0$ be an interior point of D. If f is convex, then there exists a vector $\mathbf{p}$ such that

$$f(\mathbf{x})-f(\mathbf{x}_0)\geq\mathbf{p}'\cdot(\mathbf{x}-\mathbf{x}_0) \quad\text{for all } \mathbf{x}\in D.$$



If f is a concave function on D, then there exists a vector $\mathbf{q}$ such that

$$f(\mathbf{x})-f(\mathbf{x}_0)\leq\mathbf{q}'\cdot(\mathbf{x}-\mathbf{x}_0) \quad\text{for all } \mathbf{x}\in D.$$

The vectors $\mathbf{p}$ and $\mathbf{q}$ are called **subgradient** and **supergradient**, resp., of f at $\mathbf{x}_0$.

We omit the proof and refer the interested reader to [3, Sect. 2.4].

Theorem 17.13

Jensen's inequality, discrete version. A function f on a convex domain $D\subseteq\mathbb{R}^n$ is concave if and only if

$$f\left(\sum_{i=1}^{k}\alpha_i\mathbf{x}_i\right)\geq\sum_{i=1}^{k}\alpha_i f(\mathbf{x}_i)$$

for all $\mathbf{x}_i\in D$ and $\alpha_i\geq 0$ with $\sum_{i=1}^{k}\alpha_i=1$.

PROOF. See Problem 17.13.

We finish with a quite obvious proposition.

Definition 17.14

Restriction of a function. Let $f\colon D\subseteq\mathbb{R}^n\to\mathbb{R}^m$ be some function and $S\subset D$. Then the function $f\big|_S\colon S\to\mathbb{R}^m$ defined by $f\big|_S(\mathbf{x})=f(\mathbf{x})$ for all $\mathbf{x}\in S$ is called the **restriction** of f to S.

Lemma 17.15

Let f be a (strictly) convex function on a convex set $D \subset \mathbb{R}^n$ and $S \subset D$ a convex subset. Then $f\big|_S$ is (strictly) convex.

We close this section with a few useful results.

Lemma 17.16

Function $f(\mathbf{x})$ is convex if and only if $\{(\mathbf{x}, y) \colon \mathbf{x} \in D_f, f(\mathbf{x}) \le y\}$ is convex. Function $f(\mathbf{x})$ is concave if and only if $\{(\mathbf{x}, y) \colon \mathbf{x} \in D_f, f(\mathbf{x}) \ge y\}$ is convex.

PROOF. Observe that $\{(\mathbf{x}, y) \colon \mathbf{x} \in D_f, f(\mathbf{x}) \le y\}$ is the region above the graph of f. Thus the result follows from Definition 17.5. □

Lemma 17.17

Minimum and maximum of two convex functions.

(a) If $f(\mathbf{x})$ and $g(\mathbf{x})$ are concave, then $\min\{f(\mathbf{x}), g(\mathbf{x})\}$ is concave.

(b) If $f(\mathbf{x})$ and $g(\mathbf{x})$ are convex, then $\max\{f(\mathbf{x}), g(\mathbf{x})\}$ is convex.

PROOF. See Problem 17.14.

Theorem 17.18

Composite functions. Suppose that $f \colon D_f \subseteq \mathbb{R}^n \to \mathbb{R}$ and $F \colon D_F \subseteq \mathbb{R} \to \mathbb{R}$ are two functions such that $f(D_f) \subseteq D_F$. Then the following holds:

(a) If $f(\mathbf{x})$ is concave and $F(u)$ is concave and increasing, then $G(\mathbf{x}) = F(f(\mathbf{x}))$ is concave.

(b) If $f(\mathbf{x})$ is convex and $F(u)$ is convex and increasing, then $G(\mathbf{x}) = F(f(\mathbf{x}))$ is convex.

(c) If $f(\mathbf{x})$ is concave and $F(u)$ is convex and decreasing, then $G(\mathbf{x}) = F(f(\mathbf{x}))$ is convex.

(d) If $f(\mathbf{x})$ is convex and $F(u)$ is concave and decreasing, then $G(\mathbf{x}) = F(f(\mathbf{x}))$ is concave.

PROOF. We only show (a). Assume that $f(\mathbf{x})$ is concave and $F(u)$ is concave and increasing. Then a straightforward computation gives

$$\begin{aligned} G\big((1-t)\mathbf{x} + t\mathbf{y}\big) &= F\big(f((1-t)\mathbf{x} + t\mathbf{y})\big) \ge F\big((1-t)f(\mathbf{x}) + t f(\mathbf{y})\big) \\ &\ge (1-t)F(f(\mathbf{x})) + tF(f(\mathbf{y})) = (1-t)G(\mathbf{x}) + tG(\mathbf{y}) \end{aligned}$$

where the first inequality follows from the concavity of f and the monotonicity of F. The second inequality is implied by the concavity of F. □

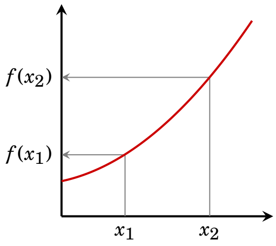
17.3 Monotone Univariate Functions

We now want to use derivatives to investigate the convexity or concavity of a given function. We start with univariate functions and look at the simpler case of monotonicity.

Definition 17.19

Monotone function. A function $f: D \subseteq \mathbb{R} \to \mathbb{R}$ is called **monotonically increasing [monotonically decreasing]** if

$$x_1 \le x_2 \quad \Rightarrow \quad f(x_1) \le f(x_2) \qquad \big[f(x_1) \ge f(x_2)\big]\,.$$



It is called **strictly increasing [strictly decreasing]** if

$$x_1 < x_2 \quad \Rightarrow \quad f(x_1) < f(x_2) \qquad \big[f(x_1) > f(x_2)\big]\,.$$

Notice that a function f is (strictly) monotonically decreasing if and only if $-f$ is (strictly) monotonically increasing. Moreover, the implication in Definition 17.19 can be replaced by an equivalence relation.

Lemma 17.20

A function $f: D \subseteq \mathbb{R} \to \mathbb{R}$ is [strictly] monotonically increasing if and only if

$$x_1 \le x_2 \quad \Leftrightarrow \quad f(x_1) \le f(x_2) \qquad \big[f(x_1) < f(x_2)\big]\,.$$

For a $\mathscr{C}^1$ function f we can use its derivative to verify monotonicity.

Theorem 17.21

Monotonicity and derivatives. Let $f: D \subseteq \mathbb{R} \to \mathbb{R}$ be a $\mathscr{C}^1$ function. Then the following holds.

(1) f is monotonically increasing on its domain D if and only if $f'(x) \ge 0$ for all $x \in D$.

(2) f is strictly increasing if $f'(x) > 0$ for all $x \in D$.

(3) If $f'(x_0) > 0$ for some $x_0 \in D$, then f is strictly increasing in an open neighborhood of x_0.

These statements holds analogously for decreasing functions.

Notice that (2) is a sufficient but not a necessary condition for strict monotonicity, see Problem 17.15.

Condition (2) can be replaced by a weaker condition that we state without proof:

(2') f is strictly increasing if $f'(x) > 0$ for almost all $x \in D$ (i.e., for all but a finite or countable number of points).

PROOF. (1) Assume that $f'(x) \ge 0$ for all $x \in D$. Let $x_1, x_2 \in D$ with $x_1 < x_2$. Then by the mean value theorem (Theorem 14.10) there exists a $\xi \in [x_1, x_2]$ such that

$$f(x_2) - f(x_1) = f'(\xi)(x_2 - x_1) \ge 0\,.$$

Hence $f(x_1) \le f(x_2)$ and thus f is monotonically increasing. Conversely, if $f(x_1) \le f(x_2)$ for all $x_1, x_2 \in D$ with $x_1 < x_2$, then

$$\frac{f(x_2) - f(x_1)}{x_2 - x_1} \ge 0 \quad \text{and thus} \quad f'(x_1) = \lim_{x_2 \to x_1} \frac{f(x_2) - f(x_1)}{x_2 - x_1} \ge 0$$

for all $x_1 \in D$. For the proof of (2) and (3) see Problem 17.15. □

17.4 Convexity of $\mathscr{C}^2$ Functions

For univariate $\mathscr{C}^2$ functions we can use the second derivative to verify convexity of the function, similar to Theorem 17.21.

Theorem 17.22

Convexity of univariate functions. Let $f: D \subseteq \mathbb{R} \to \mathbb{R}$ be a $\mathscr{C}^2$ function on an open interval $D \subseteq \mathbb{R}$. Then f is convex [concave] in D if and only if $f''(x) \geq 0$ $\left[f''(x) \leq 0 \right]$ for all $x \in D$.

PROOF IDEA. In order to verify the necessity of the condition we apply Theorem 17.11 to show that f' is increasing. Thus $f''(x) \geq 0$ by Theorem 17.21.

The sufficiency of the condition immediately follows from the Lagrange form of the reminder of the Taylor polynomial similar to the proof of Theorem 17.21.

PROOF. Assume that f is convex. Then Theorem 17.11 implies

$$f'(x_1) \leq \frac{f(x_2) - f(x_1)}{x_2 - x_1} \leq f'(x_2)$$

for all $x_1, x_2 \in D$ with $x_1 < x_2$. Hence f' is monotonically increasing and thus $f''(x) \geq 0$ for all $x \in D$ by Theorem 17.21, as claimed.

Conversely, assume that $f''(x) \geq 0$ for all $x \in D$. Then the Lagrange's form of the remainder of the first order Taylor series (Theorem 15.5) gives

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + \frac{f''(\xi)}{2}(x - x_0)^2 \geq f(x_0) + f'(x_0)(x - x_0)$$

and thus

$$f(x) - f(x_0) \geq f'(x_0)(x - x_0)\,.$$

Hence f is convex by Theorem 17.11. □

Similarly, we obtain a sufficient condition for strict convexity.

Theorem 17.23

Let $f: D \subseteq \mathbb{R} \to \mathbb{R}$ be a $\mathscr{C}^2$ function on an open interval $D \subseteq \mathbb{R}$. If $f''(x_0) > 0$ for some $x_0 \in D$, then f is strictly convex in an open neighborhood of x_0.

PROOF. Since f is a $\mathscr{C}^2$ function there exists an open ball $B_\varepsilon(x_0)$ such that $f''(x) > 0$ for all $x \in B_\varepsilon(x_0)$. Using the same argument as for Theorem 17.22 the statement follows. □

These results can be generalized for multivariate functions.

Theorem 17.24

Convexity of multivariate functions. A $\mathscr{C}^2$ function is convex (concave) on a convex, open set $D \subseteq \mathbb{R}^n$ if and only if the Hessian matrix $f''(\mathbf{x})$ is positive (negative) semidefinite for each $\mathbf{x} \in D$.

PROOF IDEA. We reduce the convexity of f to the convexity of all univariate reductions of f and apply Theorems 17.22 and 17.10.

PROOF. Let $\mathbf{x}, \mathbf{x}_0 \in D$ and $t \in [0,1]$. Define

$$g(t) = f\big((1-t)\mathbf{x}_0 + t\mathbf{x}\big) = f\big(\mathbf{x}_0 + t(\mathbf{x}-\mathbf{x}_0)\big)\,.$$

If g is convex for all $\mathbf{x}, \mathbf{x}_0 \in D$ and $t \in [0,1]$, then

$$\begin{aligned} f\big((1-t)\mathbf{x}_0 + t\mathbf{x}\big) &= g(t) = g\big((1-t)\cdot 0 + t\cdot 1\big) \\ &\le (1-t)g(0) + tg(1) = (1-t)f(\mathbf{x}_0) + tf(\mathbf{x}) \end{aligned}$$

i.e., f is convex. Similarly, if f is convex then g is convex. Applying the chain rule twice gives

$$\begin{aligned} g'(t) &= \nabla f\big(\mathbf{x}_0 + t(\mathbf{x}-\mathbf{x}_0)\big)\cdot(\mathbf{x}-\mathbf{x}_0)\,, \quad \text{and} \\ g''(t) &= (\mathbf{x}-\mathbf{x}_0)' f''\big(\mathbf{x}_0 + t(\mathbf{x}-\mathbf{x}_0)\big)\cdot(\mathbf{x}-\mathbf{x}_0)\,. \end{aligned}$$

By Theorem 17.22, g is convex if and only if $g''(t) \ge 0$ for all t. The latter is the case for all $\mathbf{x}, \mathbf{x}_0 \in D$ if and only if $f''(\mathbf{x})$ is positive semidefinite for each $\mathbf{x} \in D$ by Theorem 17.10. □

By a similar argument we find the multivariate extension of Theorem 17.23.

Theorem 17.25

Let f be a $\mathscr{C}^2$ function on a convex, open set $D \subseteq \mathbb{R}^n$ and $\mathbf{x}_0 \in D$. If $f''(\mathbf{x}_0)$ is positive (negative) definite, then f is strictly convex (strictly concave) in an open ball $B_\varepsilon(\mathbf{x}_0)$ centered at $\mathbf{x}_0$.

PROOF IDEA. Completely analogous to the proof of Theorem 17.24 except that we replace inequalities by strict inequalities. □

We can combine the results from Theorems 17.24 and 17.25 and our results from Linear Algebra as following. Let f be a $\mathscr{C}^2$ function on a convex, open set $D \subseteq \mathbb{R}^n$ and $\mathbf{x}_0 \in D$. Let $H_r(\mathbf{x})$ denotes the rth leading principle minor of $f''(\mathbf{x})$ then we find

- $H_r(\mathbf{x}_0) > 0$ for all $r \Longrightarrow f$ is strictly convex in some open ball $B_\varepsilon(\mathbf{x}_0)$.
- $(-1)^r H_r(\mathbf{x}_0) > 0$ for all $r \Longrightarrow f$ is strictly concave in $B_\varepsilon(\mathbf{x}_0)$.

The condition for semidefiniteness requires evaluations of all principle minors. Let $M_{i_1,\dots,i_r}$ denote a generic principle minor of order r of $f''(\mathbf{x})$. Then we have the following sufficient condition:

- $M_{i_1,\dots,i_r} \ge 0$ for all $\mathbf{x} \in D$ and all $i_1 < \dots i_r$ for $r = 1,\dots,n$
  $\Longleftrightarrow f$ is convex in D.
- $(-1)^r M_{i_1,\dots,i_r} \ge 0$ for all $\mathbf{x} \in D$ and all $i_1 < \dots i_r$ for $r = 1,\dots,n$
  $\Longleftrightarrow f$ is concave in D.

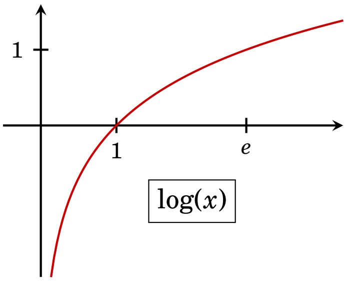
Example 17.26 **Logarithm and exponential function.** The logarithm function

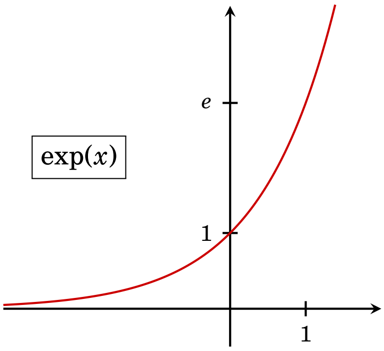
$$\log\colon D = (0,\infty) \to \mathbb{R},\, x \mapsto \log(x)$$

is strictly concave as its second derivative $(\log(x))'' = -\frac{1}{x} < 0$ is negative for all $x \in D$. The exponential function

$$\exp\colon D = \mathbb{R} \to (0,\infty),\, x \mapsto e^x$$

is a strictly convex as its second derivative $(\exp(x))'' = e^x > 0$ is positive for all $x \in D$. ♢





Example 17.27 Function $f(x,y) = x^4 + x^2 - 2xy + y^2$ is strictly convex in $D = \mathbb{R}^2$.

SOLUTION. Its Hessian matrix is

$$f''(x,y) = \begin{pmatrix} 12x^2+2 & -2 \\ -2 & 2 \end{pmatrix}$$

with leading principle minors $H_1 = 12x^2 + 2 > 0$ and $H_2 = |f''(x,y)| = 24x^2 \geq 0$. Observe that both are positive on $D_0 = \{(x,y)\colon x \neq 0\}$. Hence f is strictly convex on D_0. Since f is a $\mathscr{C}^2$ function and the closure of D_0 is $\overline{D}_0 = D$ we can conclude that f is convex on D. ♢

Example 17.28 **Cobb-Douglas function.** The Cobb-Douglas function

$$f\colon D = (0,\infty)^2 \to \mathbb{R}, (x,y) \mapsto f(x,y) = x^\alpha y^\beta$$

with $\alpha, \beta \geq 0$ and $\alpha + \beta \leq 1$ is concave.

SOLUTION. The Hessian matrix at (x,y) and its principle minors are

$$f''(x,y) = \begin{pmatrix} \alpha(\alpha-1)x^{\alpha-2}y^\beta & \alpha\beta x^{\alpha-1}y^{\beta-1} \\ \alpha\beta x^{\alpha-1}y^{\beta-1} & \beta(\beta-1)x^\alpha y^{\beta-2} \end{pmatrix},$$

$$\begin{aligned} M_1 &= \alpha(\alpha-1)x^{\alpha-2}y^\beta \leq 0\,, \\ M_2 &= \beta(\beta-1)x^\alpha y^{\beta-2} \leq 0\,, \\ M_{1,2} &= \alpha\beta(1-\alpha-\beta)x^{2\alpha-2}y^{2\beta-2} \geq 0\,. \end{aligned}$$

The Cobb-Douglas function is strict concave if $\alpha, \beta > 0$ and $\alpha + \beta < 1$. ♢

17.5 Quasi-Convex Functions

Convex and concave functions play a prominent rôle in static optimization. However, in many theorems *convexity* and *concavity* can be replaced by weaker conditions. In this section we introduce a notion that is based on level sets.

Definition 17.29

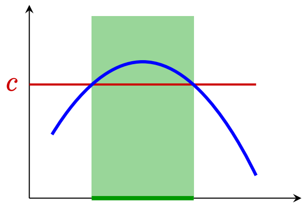
Level set. The set

$$U_c = \{\mathbf{x} \in D : f(\mathbf{x}) \geq c\} = f^{-1}\big([c,\infty)\big)$$

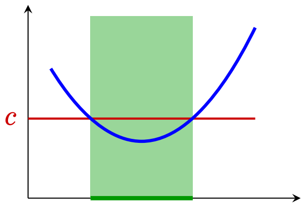
is called a **upper level set** of f. The set

$$L_c = \{\mathbf{x} \in D : f(\mathbf{x}) \leq c\} = f^{-1}\big((-\infty,c]\big)$$

is called a **lower level set** of f.



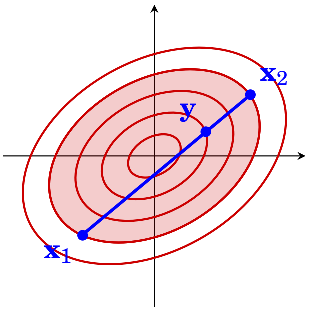
upper level set



lower level set

Lemma 17.30

Level sets of convex functions. Let $f : D \subseteq \mathbb{R}^n \to \mathbb{R}$ be a convex function and $c \in \mathbb{R}$. Then the lower level set $L_c = \{\mathbf{x} \in D : f(\mathbf{x}) \leq c\}$ is convex.



PROOF. Let $\mathbf{x}_1, \mathbf{x}_2 \in \{\mathbf{x} \in D : f(\mathbf{x}) \leq c\}$, i.e., $f(\mathbf{x}_1) \leq c$ and $f(\mathbf{x}_2) \leq c$. Then for every $\mathbf{y} = (1-t)\mathbf{x}_1 + t\mathbf{x}_2$ with $t \in [0,1]$ we find

$$f(\mathbf{y}) = f\big((1-t)\mathbf{x}_1 + t\mathbf{x}_2\big) \leq (1-t)f(\mathbf{x}_1) + t f(\mathbf{x}_2) \leq (1-t)c + tc = c$$

that is, $\mathbf{y} \in \{\mathbf{x} \in D : f(\mathbf{x}) \leq c\}$. Thus the lower level set $\{\mathbf{x} \in D : f(\mathbf{x}) \leq c\}$ is convex, as claimed. □

We will see in the next chapter that functions where all its lower level sets are convex behave in many situations similar to convex functions, that is, they are *quasi* convex. This motivates the following definition.

Definition 17.31

Quasi-convex. A function $f : D \subseteq \mathbb{R}^n \to \mathbb{R}$ is called **quasi-convex** if each of its lower level sets $L_c = \{\mathbf{x} \in D : f(\mathbf{x}) \leq c\}$ are convex.

Function f is called **quasi-concave** if each of its upper level sets $U_c = \{\mathbf{x} \in D : f(\mathbf{x}) \geq c\}$ are convex.

Analogously to Problem 17.9 we find that a function f is quasi-concave if and only if $-f$ is quasi-convex, see Problem 17.16.

Obviously every concave function is quasi-concave but not vice versa as the following examples shows.

Example 17.32

Function $f(x) = e^{-x^2}$ is quasi-concave but not concave. ♢

Theorem 17.33

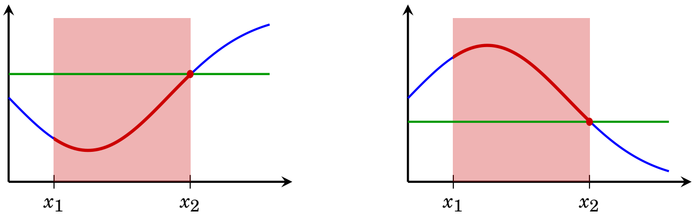
Characterization of quasi-convexity. A function f on a convex set $D \subseteq \mathbb{R}^n$ is quasi-convex if and only if

$$f\big((1-t)\mathbf{x}_1 + t\mathbf{x}_2\big) \le \max\big\{f(\mathbf{x}_1), f(\mathbf{x}_2)\big\}$$

for all $\mathbf{x}_1, \mathbf{x}_2 \in D$ and $t \in [0,1]$. The function is quasi-convex if and only if

$$f\big((1-t)\mathbf{x}_1 + t\mathbf{x}_2\big) \ge \min\big\{f(\mathbf{x}_1), f(\mathbf{x}_2)\big\}$$

for all $\mathbf{x}_1, \mathbf{x}_2 \in D$ and $t \in [0,1]$.



quasi-convex quasi-concave

PROOF IDEA. For $c = \max\big\{f(\mathbf{x}_1), f(\mathbf{x}_2)\big\}$ we find that

$$(1-t)\mathbf{x}_1 + t\mathbf{x}_2 \in L_c = \{\mathbf{x} \in D : f(\mathbf{x}) \le c\}$$

is equivalent to

$$f\big((1-t)\mathbf{x}_1 + t\mathbf{x}_2\big) \le c = \max\big\{f(\mathbf{x}_1), f(\mathbf{x}_2)\big\}\,.$$

PROOF. Let $\mathbf{x}_1, \mathbf{x}_2 \in D$ and $t \in [0,1]$. Let $c = \max\big\{f(\mathbf{x}_1), f(\mathbf{x}_2)\big\}$ and assume w.l.o.g. that $c = f(x_2) \ge f(x_1)$. If f is quasi-convex, then $(1-t)\mathbf{x}_1 + t\mathbf{x}_2 \in L_c = \{\mathbf{x} \in D : f(\mathbf{x}) \le c\}$ and thus $f((1-t)\mathbf{x}_1 + t\mathbf{x}_2) \le c = \max\big\{f(\mathbf{x}_1), f(\mathbf{x}_2)\big\}$. Conversely, if $f((1-t)\mathbf{x}_1 + t\mathbf{x}_2) \le c = \max\big\{f(\mathbf{x}_1), f(\mathbf{x}_2)\big\}$, then $(1-t)\mathbf{x}_1 + t\mathbf{x}_2 \in L_c$ and thus f is quasi-convex. The case for quasi-concavity is shown analogously. □

In Theorem 17.18 we have seen that some compositions of functions preserve convexity. Quasi-convexity is preserved under even milder condition.

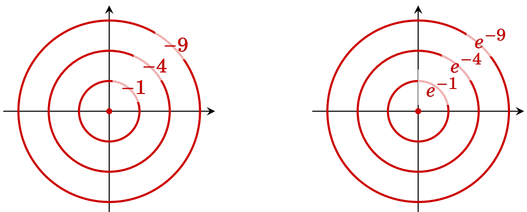
Theorem 17.34

Composite functions. Suppose that $f : D_f \subseteq \mathbb{R}^n \to \mathbb{R}$ and $F : D_F \subseteq \mathbb{R} \to \mathbb{R}$ are two functions such that $f(D_f) \subseteq D_F$. Then the following holds:

(a) If $f(\mathbf{x})$ is quasi-convex (quasi-concave) and $F(u)$ is increasing, then $G(\mathbf{x}) = F(f(\mathbf{x}))$ is quasi-convex (quasi-concave).

(b) If $f(\mathbf{x})$ is quasi-convex (quasi-concave) and $F(u)$ is decreasing, then $G(\mathbf{x}) = F(f(\mathbf{x}))$ is quasi-concave (quasi-convex).

PROOF IDEA. Monotone transformations preserve (in some sense) level sets of functions.



$f(x,y) = -x^2 - y^2$ $\exp(-x^2 - y^2)$

PROOF. Assume that f is quasi-convex and F is increasing. Thus by Theorem 17.33 $f\big((1-t)\mathbf{x}_1 + t\mathbf{x}_2\big) \le \max\big\{f(\mathbf{x}_1), f(\mathbf{x}_2)\big\}$ for all $\mathbf{x}_1, \mathbf{x}_2 \in D$ and $t \in [0,1]$. Moreover, $F(y_1) \le F(y_2)$ if and only if $y_1 \le y_2$. Hence we find for all $\mathbf{x}_1, \mathbf{x}_2 \in D$ and $t \in [0,1]$,

$$F\big(f((1-t)\mathbf{x}_1 + t\mathbf{x}_2)\big) \le F\big(\max\{f(\mathbf{x}_1), f(\mathbf{x}_2)\}\big) \le \max\big\{F(f(\mathbf{x}_1)), f(f(\mathbf{x}_2))\big\}$$

and thus $F \circ f$ is quasi-convex. The proof for the other cases is completely analogous. □

Theorem 17.34 allows to determine quasi-convexity or quasi-concavity of some functions. In Example 17.28 we have shown that the Cobb-Douglas function is concave for appropriate parameters. The computation was a bit tedious and it is not straightforward to extend the proof to functions of the form $\sum_{i=1}^n x_i^{\alpha_i}$. Quasi-concavity is much easier to show. Moreover, it holds for a larger range of parameters and our computation easily generalizes to many variables.

Cobb-Douglas function. The Cobb-Douglas function Example 17.35

$$f\colon D = (0,\infty)^2 \to \mathbb{R}, (x,y) \mapsto f(x,y) = x^\alpha y^\beta$$

with $\alpha, \beta \ge 0$ is quasi-concave.

SOLUTION. Observe that $f(x,y) = \exp\big(\alpha\log(x) + \beta\log(y)\big)$. Notice that $\log(x)$ is concave by Example 17.26. Thus $\alpha\log(x) + \beta\log(y)$ is concave by Theorem 17.9 and hence quasi-concave. Since the exponential function exp is monotonically increasing, it follows that the Cobb-Douglas function is quasi-concave if $\alpha, \beta > 0$. ♢

Notice that it is not possible to apply Theorem 17.18 to show concavity of the Cobb-Douglas function when $\alpha + \beta \le 1$.

CES function. Let $a_1, \ldots, a_n \ge 0$. Then function Example 17.36

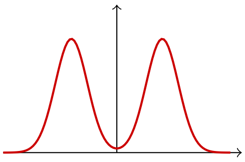
$$f(\mathbf{x}) = \left(\sum_{i=1}^n a_i x_i^r\right)^{1/r}$$

is quasi-concave for all $r \le 1$ and quasi-convex for all $r \ge 1$.

SOLUTION. Since $\big(x_i^r\big)'' = r(r-1)x_i^{r-2}$, we find that x_i^r is concave for $r \in [0,1]$ and convex otherwise. Hence the same holds for $\sum_{i=1}^n a_i x_i^r$ by Theorem 17.9. Since $F(y) = y^{1/r}$ is monotonically increasing if $r > 0$ and decreasing if $r < 0$, Theorem 17.34 implies that $f(\mathbf{x})$ is quasi-concave for all $r \le 1$ and quasi-convex for all $r \ge 1$. ♢

In opposition to Theorem 17.9 the sum of quasi-convex functions need not be quasi-convex.

Example 17.37



The two functions $f_1(x) = \exp\big(-(x-2)^2\big)$ and $f_2(x) = \exp\big(-(x+2)^2\big)$ are quasi-concave as each of their upper level sets are intervals (or empty). However, $f_1(x)+f_2(x)$ has two local maxima and thus cannot be quasi-concave.

There is also an analog to strict convexity. However, a definition using lower level set were not useful. So we start with the characterization of quasi-convexity in Theorem 17.33.

Definition 17.38

Strictly quasi-convex. A function f on a convex set $D \subseteq \mathbb{R}^n$ is called **strictly quasi-convex** if

$$f\big((1-t)\mathbf{x}_1 + t\mathbf{x}_2\big) < \max\{f(\mathbf{x}_1), f(\mathbf{x}_2)\}$$

for all $\mathbf{x}_1, \mathbf{x}_2$ with $\mathbf{x}_1 \neq \mathbf{x}_2$, and $t \in (0,1)$. It is called **strictly quasi-concave** if

$$f\big((1-t)\mathbf{x}_1 + t\mathbf{x}_2\big) > \min\{f(\mathbf{x}_1), f(\mathbf{x}_2)\}$$

for all $\mathbf{x}_1, \mathbf{x}_2$ with $\mathbf{x}_1 \neq \mathbf{x}_2$, and $t \in (0,1)$.

Our last result shows, that we also can use tangents to characterize quasi-convex function. Again, the condition is weaker than the corresponding condition in Theorem 17.11.

Theorem 17.39

Tangents of quasi-convex functions. A $\mathscr{C}^1$ function f is quasi-convex in an open, convex set D if and only if

$$f(\mathbf{x}) \leq f(\mathbf{x}_0) \quad \text{implies} \quad \nabla f(\mathbf{x}_0)\cdot(\mathbf{x}-\mathbf{x}_0) \leq 0$$

for all $\mathbf{x}, \mathbf{x}_0 \in D$. It is quasi-concave if and only if

$$f(\mathbf{x}) \geq f(\mathbf{x}_0) \quad \text{implies} \quad \nabla f(\mathbf{x}_0)\cdot(\mathbf{x}-\mathbf{x}_0) \geq 0$$

for all $\mathbf{x}, \mathbf{x}_0 \in D$.

PROOF. Assume that f is quasi-convex and $f(\mathbf{x}) \leq f(\mathbf{x}_0)$. Define $g(t) = f\big((1-t)\mathbf{x}_0 + t\mathbf{x}\big) = f\big(\mathbf{x}_0 + t(\mathbf{x}-\mathbf{x}_0)\big)$. Then Theorem 17.33 implies that $g(0) = f(\mathbf{x}_0) \geq g(t)$ for all $t \in [0,1]$ and hence $g'(0) \leq 0$. By the chain rule we find $g'(t) = \nabla f\big(\mathbf{x}_0 + t(\mathbf{x}-\mathbf{x}_0)\big)(\mathbf{x}-\mathbf{x}_0)$ and consequently $g'(0) = \nabla f(\mathbf{x}_0)\cdot(\mathbf{x}-\mathbf{x}_0) \leq 0$ as claimed.

For the converse assume that f is not quasi-convex. Then there exist $\mathbf{x}, \mathbf{x}_0 \in D$ with $f(\mathbf{x}) \leq f(\mathbf{x}_0)$ and a $\mathbf{z} = \mathbf{x}_0 + t(\mathbf{x}-\mathbf{x}_0) \in D$ for some $t \in (0,1)$ such that $f(\mathbf{z}) > f(\mathbf{x}_0)$. Define $g(t)$ as above. Then $g(t) > g(0)$ and there exists a $\tau \in (0,1)$ such that $g'(\tau) > 0$ by the Mean Value Theorem, and thus $\nabla f(\mathbf{z}_0)\cdot(\mathbf{x}-\mathbf{z}_0) > 0$, where $\mathbf{z}_0 = g(\tau)$. We state without proof that we can find such a point $\mathbf{z}_0$ where $f(\mathbf{z}_0) \geq f(\mathbf{x})$. Thus the given condition is violated for some points.

The proof for the second statement is completely analogous. □

— Exercises

17.1 Let

$$f(x) = x^4 + \frac{4}{3}x^3 - 24x^2 + 8\,.$$

(a) Determinte the regions where f is monotonically increasing and monotonically decreasing, respectively.

(b) Determinte the regions where f is concave and convex, respectively.

17.2 A function $f\colon \mathbb{R} \to (0,\infty)$ is called *log-concave* if $\ln\circ f$ is a concave function.

Which of the following functions is log-concave?

(a) $f(x) = 3\,\exp(-x^4)$

(b) $g(x) = 4\,\exp(-x^7)$

(c) $h(x) = 2\,\exp(x^2)$

(d) $s\colon (-1,1) \to (0,\infty)$, $x \mapsto s(x) = 1 - x^4$

17.3 Determine whether the following functions are convex, concave or neither.

(a) $f(x) = \exp\left(-\sqrt{x}\right)$ on $D = [0,\infty)$.

(b) $f(\mathbf{x}) = \exp\left(-\sum_{i=1}^{n}\sqrt{x_i}\right)$ on $D = [0,\infty)^n$.

17.4 Determine whether the following functions on $\mathbb{R}^2$ are (strictly) convex or (strictly) concave or neither.

(a) $f(x,y) = x^2 - 2xy + 2y^2 + 4x - 8$

(b) $g(x,y) = 2x^2 - 3xy + y^2 + 2x - 4y - 2$

(c) $h(x,y) = -x^2 + 4xy - 4y^2 + 1$

17.5 Show that function

$$f(x,y) = ax^2 + 2bxy + cy^2 + px + qy + r$$

is strictly concave if $ac - b^2 > 0$ and $a < 0$, and strictly convex if $ac - b^2 > 0$ and $a > 0$.

Find necessary and sufficient conditions for (strict) convexity/concavity of f.

17.6 Show that $f(x,y) = \exp(-x^2 - y^2)$ is quasi-concave in $D = \mathbb{R}^2$ but not concave.

Apply Theorem 17.34.

Is there a domain where f is (strictly) concave? Compute the largest of such domains.

— Problems

17.7 Let $S_1,\dots,S_k$ be convex sets in $\mathbb{R}^n$. Show that their intersection $\bigcap_{i=1}^k S_i$ is convex (Theorem 17.3).

Give an example where the union of convex sets is not convex.

17.8 Show that the sets H, H^+, and H^- in Example 17.4 are convex.

17.9 Show that a function $f\colon D \subseteq \mathbb{R}^n \to \mathbb{R}$ is (strictly) concave if and only if function $g\colon D \subseteq \mathbb{R}^n \to \mathbb{R}$ with $g(\mathbf{x}) = -f(\mathbf{x})$ is (strictly) convex.

17.10 A function $f\colon \mathbb{R} \to (0,\infty)$ is called *log-concave* if $\ln\circ f$ is a concave function.

Show that every concave function $f\colon \mathbb{R} \to (0,1)$ is log-concave.

17.11 Let $T\colon (0,\infty) \to \mathbb{R}$ be a strictly monotonically increasing twice differentiable transformation. A function $f\colon \mathbb{R} \to (0,\infty)$ is called T-*concave* if $T\circ f$ is a concave function.

Consider the family $T_c(x)$, $c \le 0$, of transformations with $T_0(x) = \ln(x)$ and $T_c(x) = -x^c/c$ for $c < 0$.

(a) Show that all transformations T_c satisfy the above conditions for all $c \le 0$.

(b) Show that $f(x) = \exp(-x^2)$ is $T_{-1/2}$-concave.

(c) Show that $f(x) = \exp(-x^2)$ is T_c-concave for alle $c \le 0$.

(d) Show that every T_{c_0}-concave function $f\colon \mathbb{R} \to (0,\infty)$ with $c_0 < 0$ is also T_c-concave for all $c \le c_0$.
HINT: f is T_c-concave if and only if $(T(f(x)))''/(cf(x)^{(}c-2)) \le 0$ for all x. Compute this term and derive a condition on c.

17.12 Prove Theorem 17.9.

17.13 Prove Jensen's inequality (Theorem 17.13).

HINT: For $k = 2$ the theorem is equivalent to the definition of concavity. For $k \ge 3$ use induction.

HINT: Use Lemma 17.16.

17.14 Prove Lemma 17.17.

17.15 Prove (2) and (3) of Theorem 17.21.

HINT: Give a strictly increasing function f where $f'(0) = 0$.

Condition (2) (i.e., $f'(x) > 0$ for all $x \in D$) is sufficient for f being strictly monotonically increasing. Give a counterexample that shows that this condition is not necessary.

Suppose one wants to prove the (false!) statement that $f'(x) > 0$ for each $x \in D_f$ for every strictly increasing function f. Thus he or she uses the same argument as in the proof of Theorem 17.21(1). Where does this argument fail?

17.16 Show that a function $f: D \subseteq \mathbb{R}^n \to \mathbb{R}$ is (strictly) quasi-concave if and only if function $g: D \subseteq \mathbb{R}^n \to \mathbb{R}$ with $g(\mathbf{x}) = -f(\mathbf{x})$ is (strictly) quasi-convex.

18

Static Optimization

We want to find the highest peak in our world.

18.1 Extremal Points

We start with so called global extrema.

Extremal points. Let $f \colon D \subseteq \mathbb{R}^n \to \mathbb{R}$. Then $\mathbf{x}^* \in D$ is called a (global) **maximum** of f if Definition 18.1

$$f(\mathbf{x}^*) \geq f(\mathbf{x}) \quad \text{for all } \mathbf{x} \in D.$$

It is called a **strict maximum** if the inequality is strict for $\mathbf{x} \neq \mathbf{x}^*$.

Similarly, $\mathbf{x}^* \in D$ is called a (global) **minimum** of f if $f(\mathbf{x}^*) \leq f(\mathbf{x})$ for all $\mathbf{x} \in D$.

A **stationary point** $\mathbf{x}_0$ of a function f is a point where the gradient vanishes, i.e, Definition 18.2

$$\nabla f(\mathbf{x}_0) = 0\,.$$

Necessary first-order conditions. Let f be a $\mathscr{C}^1$ function on an open set $D \subseteq \mathbb{R}^n$ and let $\mathbf{x}^* \in D$ be an extremal point. Then $\mathbf{x}^*$ is a stationary point of f, i.e., Theorem 18.3

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

PROOF. If $\mathbf{x}^*$ is an extremal point then all directional derivatives are 0 and thus the result follows. □

Sufficient conditions. Let f be a $\mathscr{C}^1$ function on an open set $D \subseteq \mathbb{R}^n$ and let $\mathbf{x}^* \in D$ be a stationary point of f. Theorem 18.4

If f is (strictly) convex in D, then $\mathbf{x}^*$ is a (strict) minimum of f.

If f is (strictly) concave in D, then $\mathbf{x}^*$ is a (strict) maximum of f.

PROOF. Assume that f is strictly convex. Then by Theorem 17.11

$$f(\mathbf{x})-f(\mathbf{x}^*) > \nabla f(\mathbf{x}^*)\cdot(\mathbf{x}-\mathbf{x}^*) = 0\cdot(\mathbf{x}-\mathbf{x}^*) = 0$$

and hence $f(\mathbf{x}) > f(\mathbf{x}^*)$ for all $\mathbf{x} \neq \mathbf{x}^*$, as claimed. The other statements follow analogously. □

Example 18.5

Cobb-Douglas function. We want to find the (global) maxima of

$$f\colon D=[0,\infty)^2 \to \mathbb{R},\ f(x,y) = 4x^{\frac{1}{4}}y^{\frac{1}{4}} - x - y\,.$$

SOLUTION. A straightforward computation yields

$$f_x = x^{-\frac{3}{4}}y^{\frac{1}{4}} - 1$$
$$f_y = x^{\frac{1}{4}}y^{-\frac{3}{4}} - 1$$

and thus $\mathbf{x}_0 = (1,1)$ is the only stationary point of this function. As f is strictly concave (see Example 17.28) $\mathbf{x}_0$ is the global maximum of f. ♢

Definition 18.6

Local extremal points. Let $f\colon D \subseteq \mathbb{R}^n \to \mathbb{R}$. Then $\mathbf{x}^* \in D$ is called a **local maximum** of f if there exists an $\varepsilon > 0$ such that

$$f(\mathbf{x}^*) \geq f(\mathbf{x}) \quad \text{for all } \mathbf{x} \in B_\varepsilon(\mathbf{x}^*).$$

It is called a **strict local maximum** if the inequality is strict for $\mathbf{x} \neq \mathbf{x}^*$.

Similarly, $\mathbf{x}^* \in D$ is called a **local minimum** of f if there exists an $\varepsilon > 0$ such that $f(\mathbf{x}^*) \leq f(\mathbf{x})$ for all $\mathbf{x} \in B_\varepsilon(\mathbf{x}^*)$.

Local extrema necessarily are stationary points.

Theorem 18.7

Sufficient conditions for local extremal points. Let f be a $\mathscr{C}^2$ function on an open set $D \subseteq \mathbb{R}^n$ and let $\mathbf{x}^* \in D$ be a stationary point of f.

If $f''(\mathbf{x}^*)$ is positive definite, then $\mathbf{x}^*$ is a strict local minimum of f.

If $f''(\mathbf{x}^*)$ is negative definite, then $\mathbf{x}^*$ is a strict local maximum of f.

PROOF. Assume that $f''(\mathbf{x}^*)$ is positive definite. Since f'' is continuous, there exists an ε such that $f''(\mathbf{x})$ is positive definite for all $\mathbf{x} \in B_\varepsilon(\mathbf{x}^*)$ and hence f is strictly convex in $B_\varepsilon(\mathbf{x}^*)$. Consequently, $\mathbf{x}^*$ is a strict minimum in $B_\varepsilon(\mathbf{x}^*)$ by Theorem 18.4, i.e., a strict local minimum of f. □

Example 18.8

We want to find all local maxima of

$$f(x,y) = \frac{1}{6}x^3 - x + \frac{1}{4}xy^2\,.$$

SOLUTION. The partial derivative of f are given as

$$\begin{aligned} f_x &= \frac{1}{2}x^2 - 1 + \frac{1}{4}y^2, \\ f_y &= \frac{1}{2}xy, \end{aligned}$$

and hence we find the stationary points $\mathbf{x}_1 = (0,2)$, $\mathbf{x}_2 = (0,-2)$, $\mathbf{x}_3 = (\sqrt{2},0)$, and $\mathbf{x}_4 = (-\sqrt{2},0)$. In order to apply Theorem 17.25 we need the Hessian of f,

$$f''(x,y) = \begin{pmatrix} f_{xx}(\mathbf{x}) & f_{xy}(\mathbf{x}) \\ f_{yx}(\mathbf{x}) & f_{yy}(\mathbf{x}) \end{pmatrix} = \begin{pmatrix} x & \frac{1}{2}y \\ \frac{1}{2}y & \frac{1}{2}x \end{pmatrix}.$$

We then find $f''(\mathbf{x}_3) = \begin{pmatrix} \sqrt{2} & 0 \\ 0 & \frac{\sqrt{2}}{2} \end{pmatrix}$. Its leading principle minors are both positive, $H_1 = \sqrt{2} > 0$ and $H_2 = 1 > 0$, and hence $\mathbf{x}_3$ is a local minimum. Similarly we find that $\mathbf{x}_4$ is a local maximum. ♢

Besides (local) extrema there are also other types of stationary points.

Definition 18.9

Saddle point. Let f be a $\mathscr{C}^2$ function on an open set $D \subseteq \mathbb{R}^n$. A stationary point $\mathbf{x}_0 \in D$ is called a **saddle point** if $f''(\mathbf{x}_0)$ is indefinite, that is, if f is neither convex nor concave in any open ball around $\mathbf{x}^*$.

Example 18.10

In Example 18.8 we have found two additional stationary points: $\mathbf{x}_1 = (0,2)$ and $\mathbf{x}_2 = (0,-2)$. However, the Hessian of f at $\mathbf{x}_1$,

$$f''(\mathbf{x}_1) = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

is indefinite as it has leading principle minors $H_1 = 0$ and $H_2 = -1 < 0$. Consequently $\mathbf{x}_1$ is a saddle point. ♢

18.2 The Envelope Theorem

Let $f(\mathbf{x},\mathbf{r})$ be a $\mathscr{C}^1$ function with (endogenous) variable $\mathbf{x} \in D \subseteq \mathbb{R}^n$ and parameter (exogenous variable) $\mathbf{r} \in \mathbb{R}^k$. An extremal point of f may depend on $\mathbf{r}$. So let $\mathbf{x}^*(\mathbf{r})$ denote an extremal point for a given parameter $\mathbf{r}$ and let

$$f^*(\mathbf{r}) = \max_{\mathbf{x}\in D} f(\mathbf{x},\mathbf{r}) = f(\mathbf{x}^*(\mathbf{r}),\mathbf{r})$$

be the **value function**.

Theorem 18.11

Envelope theorem. Let $f(\mathbf{x},\mathbf{r})$ be a $\mathscr{C}^1$ function on $D \times \mathbb{R}^k$ where $D \subseteq \mathbb{R}^n$. Let $\mathbf{x}^*(\mathbf{r})$ denote an extremal point for a given parameter $\mathbf{r}$ and assume that $\mathbf{r} \mapsto \mathbf{x}^*(\mathbf{r})$ is differentiable. Then

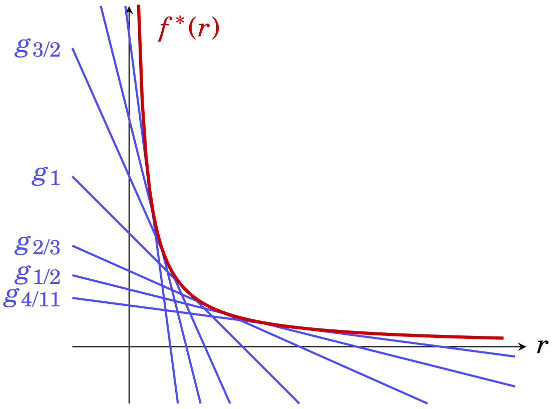
$$\frac{\partial f^*(\mathbf{r})}{\partial r_j} = \left.\frac{\partial f(\mathbf{x},\mathbf{r})}{\partial r_j}\right|_{\mathbf{x}=\mathbf{x}^*(\mathbf{r})}$$

PROOF IDEA. The chain rule implies

$$\begin{aligned}\frac{\partial f^*(\mathbf{r})}{\partial r_j} &= \frac{\partial f(\mathbf{x}^*(\mathbf{r}),\mathbf{r})}{\partial r_j} \\ &= \sum_{i=1}^{n} \underbrace{f_{x_i}(\mathbf{x}^*(\mathbf{r}),\mathbf{r})}_{=0} \cdot \frac{\partial x_i^*(\mathbf{r})}{\partial r_j} + \left.\frac{\partial f(\mathbf{x},\mathbf{r})}{\partial r_j}\right|_{\mathbf{x}=\mathbf{x}^*(\mathbf{r})} = \left.\frac{\partial f(\mathbf{x},\mathbf{r})}{\partial r_j}\right|_{\mathbf{x}=\mathbf{x}^*(\mathbf{r})}\end{aligned}$$

as claimed. □

The following figure illustrates this theorem. Let $f(x,r) = \sqrt{x} - rx$ and $f^*(r) = \max_x f(x,r)$. $g_x(r) = f(r,x)$ denotes function f with argument x fixed. Observe that $f^*(r) = \max_x g_x(r)$.



See Lecture 11 in *Mathematische Methoden* for further examples.

18.3 Constraint Optimization – The Lagrange Function

In this section we consider the optimization problem

$$\begin{aligned}&\max\ (\min) \quad f(x_1,\dots,x_n) \\ &\text{subject to} \quad g_j(x_1,\dots,x_n) = c_j, \quad j = 1,\dots,m \quad (m < n)\end{aligned}$$

or in vector notation

$$\max\ (\min) \quad f(\mathbf{x}) \quad \text{subject to} \quad \mathbf{g}(\mathbf{x}) = \mathbf{c}\,.$$

Definition 18.12

Lagrange function. Function

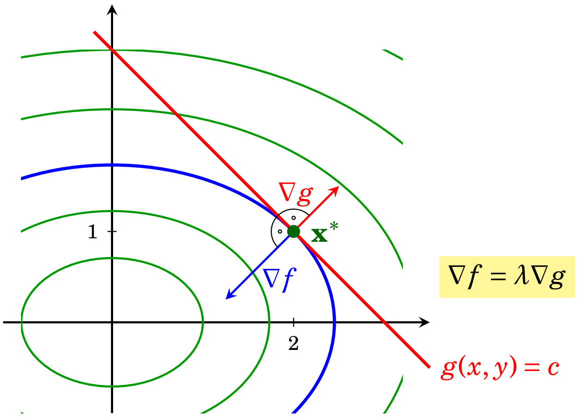
$$\mathscr{L}(\mathbf{x};\boldsymbol{\lambda}) = f(\mathbf{x}) - \boldsymbol{\lambda}'(\mathbf{g}(\mathbf{x}) - \mathbf{c}) = f(\mathbf{x}) - \sum_{j=1}^{m} \lambda_j (g_j(\mathbf{x}) - c_j)$$

is called the **Lagrange function** (or **Lagrangian**) of the above constraint optimization problem. The numbers λ_j are called **Lagrange multipliers**.

In order to find candidates for solutions of the constraint optimization problem we have to find stationary points of the Lagrange function. We state this condition without a proof.

Necessary condition. Suppose that f and $\mathbf{g}$ are $\mathscr{C}^1$ functions and $\mathbf{x}^*$ (locally) solves the constraint optimization problem and $\mathbf{g}'(\mathbf{x}^*)$ has maximal rank m, then there exist a unique vector $\boldsymbol{\lambda}^* = (\lambda_1^*, \dots, \lambda_m^*)$ such that $\nabla \mathscr{L}(\mathbf{x}^*, \boldsymbol{\lambda}^*) = 0$. Theorem 18.13

This necessary condition implies that $\partial f'(\mathbf{x}^*) = \boldsymbol{\lambda}^* \mathbf{g}'(\mathbf{x}^*)$. The following figure illustrates the situation for the case of two variables x and y and one constraint $g(x,y) = c$. Then we have find $\nabla f = \lambda \nabla g$, that is, in an optimal point ∇f is some multiple of ∇g.



Also observe that a point $\mathbf{x}$ is admissible (i.e., satisfies constraint $\mathbf{g}(\mathbf{x}) = \mathbf{c}$) if and only if $\frac{\partial \mathscr{L}}{\partial \boldsymbol{\lambda}}(\mathbf{x}, \boldsymbol{\lambda}) = 0$ for some vector $\boldsymbol{\lambda} = 0$.

Sufficient condition. Let f and $\mathbf{g}$ be $\mathscr{C}^1$. Suppose there exists a $\boldsymbol{\lambda}^* = (\lambda_1^*, \dots, \lambda_m^*)$ and an admissible $\mathbf{x}^*$ such that $(\mathbf{x}^*, \boldsymbol{\lambda}^*)$ is a stationary point of $\mathscr{L}$, i.e., $\nabla \mathscr{L}(\mathbf{x}^*, \boldsymbol{\lambda}^*) = 0$. If $\mathscr{L}(\mathbf{x}, \boldsymbol{\lambda}^*)$ is concave (convex) in $\mathbf{x}$, then $\mathbf{x}^*$ solves the constraint maximization (minimization) problem. Theorem 18.14

Proof. By Theorem 18.4 these conditions imply that $\mathbf{x}^*$ is a maximum of $\mathscr{L}(\mathbf{x}, \boldsymbol{\lambda}^*)$ w.r.t. $\mathbf{x}$, i.e.,

$$\begin{aligned}\mathscr{L}(\mathbf{x}^*; \boldsymbol{\lambda}^*) &= f(\mathbf{x})^* - \sum_{j=1}^m \lambda_j^*(g_j(\mathbf{x}^*) - c_j) \\ &\geq f(\mathbf{x}) - \sum_{j=1}^m \lambda_j^*(g_j(\mathbf{x}) - c_j) = \mathscr{L}(\mathbf{x}; \boldsymbol{\lambda}^*).\end{aligned}$$

However, all admissible $\mathbf{x}$ satisfy $g_j(\mathbf{x}) = c_j$ for all j and thus $f(\mathbf{x}^*) \geq f(\mathbf{x})$ for all admissible $\mathbf{x}$. Hence $\mathbf{x}^*$ solves the constraint maximization problem. □

Similar to Theorem 18.7 we can find sufficient conditions for local solutions of the constraint optimization problem. That is, $(\mathbf{x}^*, \boldsymbol{\lambda}^*)$ is a stationary point of $\mathscr{L}$ and $\mathscr{L}$ w.r.t. $\mathbf{x}$ is strictly concave (strictly convex) in some open ball around $(\mathbf{x}^*, \boldsymbol{\lambda}^*)$, then $\mathbf{x}^*$ solves the local constraint maximization (minimization) problem. Such an open ball exists if the Hessian of $\mathscr{L}$ w.r.t. $\mathbf{x}$ is negative (positive) definite in $(\mathbf{x}^*, \boldsymbol{\lambda}^*)$.

However, such a condition is too strong. There is no need to investigate the behavior of $\mathscr{L}$ for points $\mathbf{x}$ that do not satisfy constraint $\mathbf{g}(\mathbf{x}) = \mathbf{c}$. Hence (roughly spoken) it is sufficient that the Lagrange function $\mathscr{L}$ is strictly concave on the affine subspace spanned by the gradients $\nabla g_1(\mathbf{x}^*), \dots, \nabla g_m(\mathbf{x}^*)$. Again it is sufficient to look at the definiteness of the Hessian $\mathscr{L}''$ at $\mathbf{x}^*$. ($\mathscr{L}''$ denotes the Hessian w.r.t. $\mathbf{x}$.)

Lemma 18.15

Let f and $\mathbf{g}$ be $\mathscr{C}^1$. Suppose there exists a $\boldsymbol{\lambda}^* = (\lambda_1^*, \dots, \lambda_m^*)$ and an admissible $\mathbf{x}^*$ such that $\nabla\mathscr{L}(\mathbf{x}^*, \boldsymbol{\lambda}^*) = 0$. If there exists an open ball around $\mathbf{x}^*$ such that the quadratic form

$$\mathbf{h}'\mathscr{L}''(\mathbf{x}^*; \boldsymbol{\lambda}^*)\mathbf{h}$$

is negative (positive) definite for all $\mathbf{h} \in \operatorname{span}\big(\nabla g_1(\mathbf{x}^*), \dots, \nabla g_m(\mathbf{x}^*)\big)$, then $\mathbf{x}^*$ solves the local constraint maximization (minimization) problem.

This condition can be verified by means of a theorem from Linear Algebra which requires the concept of the border Hessian.

Definition 18.16

Border Hessian. The matrix

$$\bar{\mathbf{H}}(\mathbf{x}; \boldsymbol{\lambda}) = \begin{pmatrix} 0 & \mathbf{g}'(\mathbf{x}) \\ (\mathbf{g}'(\mathbf{x}))' & \mathscr{L}''(\mathbf{x}; \boldsymbol{\lambda}) \end{pmatrix}$$

$$= \begin{pmatrix} 0 & \dots & 0 & \frac{\partial g_1}{\partial x_1} & \dots & \frac{\partial g_1}{\partial x_n} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & \frac{\partial g_m}{\partial x_1} & \dots & \frac{\partial g_m}{\partial x_n} \\ \frac{\partial g_1}{\partial x_1} & \dots & \frac{\partial g_m}{\partial x_1} & \mathscr{L}_{x_1x_1} & \dots & \mathscr{L}_{x_1x_n} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \frac{\partial g_1}{\partial x_n} & \dots & \frac{\partial g_m}{\partial x_n} & \mathscr{L}_{x_nx_1} & \dots & \mathscr{L}_{x_nx_n} \end{pmatrix}$$

is called the **border Hessian** of $\mathscr{L}(\mathbf{x}; \boldsymbol{\lambda}) = f(\mathbf{x}) - \boldsymbol{\lambda}'(\mathbf{g}(\mathbf{x}) - \mathbf{c})$.
We denote its leading principal minors by

$$B_r(\mathbf{x}) = \begin{vmatrix} 0 & \dots & 0 & \frac{\partial g_1}{\partial x_1}(\mathbf{x}; \boldsymbol{\lambda}) & \dots & \frac{\partial g_1}{\partial x_r}(\mathbf{x}; \boldsymbol{\lambda}) \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & \frac{\partial g_m}{\partial x_1}(\mathbf{x}; \boldsymbol{\lambda}) & \dots & \frac{\partial g_m}{\partial x_r}(\mathbf{x}; \boldsymbol{\lambda}) \\ \frac{\partial g_1}{\partial x_1}(\mathbf{x}; \boldsymbol{\lambda}) & \dots & \frac{\partial g_m}{\partial x_1}(\mathbf{x}; \boldsymbol{\lambda}) & \mathscr{L}_{x_1x_1}(\mathbf{x}; \boldsymbol{\lambda}) & \dots & \mathscr{L}_{x_1x_r}(\mathbf{x}; \boldsymbol{\lambda}) \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \frac{\partial g_1}{\partial x_r}(\mathbf{x}; \boldsymbol{\lambda}) & \dots & \frac{\partial g_m}{\partial x_r}(\mathbf{x}; \boldsymbol{\lambda}) & \mathscr{L}_{x_rx_1}(\mathbf{x}; \boldsymbol{\lambda}) & \dots & \mathscr{L}_{x_rx_r}(\mathbf{x}; \boldsymbol{\lambda}) \end{vmatrix}.$$

Sufficient condition for local optimum. Let f and $\mathbf{g}$ be $\mathscr{C}^1$. Suppose there exists a $\boldsymbol{\lambda}^* = (\lambda_1^*, \dots, \lambda_m^*)$ and an admissible $\mathbf{x}^*$ such that $\nabla\mathscr{L}(\mathbf{x}^*, \boldsymbol{\lambda}^*) = 0$. Theorem 18.17

(a) If $(-1)^r B_r(\mathbf{x}) > 0$ for all $r = m+1, \dots, n$, then $\mathbf{x}^*$ solves the local constraint maximization problem.

(b) If $(-1)^m B_r(\mathbf{x}) > 0$ for all $r = m+1, \dots, n$, then $\mathbf{x}^*$ solves the local constraint minimization problem.

See Lecture 12 in *Mathematische Methoden* for examples.

18.4 Kuhn-Tucker Conditions

In this section we consider the optimization problem

$$\begin{aligned}
&\max \quad f(x_1, \dots, x_n) \\
&\text{subject to} \quad g_j(x_1, \dots, x_n) \le c_j, \quad j = 1, \dots, m \quad (m < n), \\
&\quad \text{and} \quad x_i \ge 0, \quad i = 1, \dots n \quad \text{(non-negativity constraint)}
\end{aligned}$$

or in vector notation

$$\max \quad f(\mathbf{x}) \quad \text{subject to} \quad \mathbf{g}(\mathbf{x}) \ge \mathbf{c} \quad \text{and} \quad \mathbf{x} \ge 0.$$

Again let

$$\mathscr{L}(\mathbf{x}; \boldsymbol{\lambda}) = f(\mathbf{x}) - \boldsymbol{\lambda}'(\mathbf{g}(\mathbf{x}) - \mathbf{c}) = f(\mathbf{x}) - \sum_{j=1}^{m} \lambda_j (g_j(\mathbf{x}) - c_j)$$

denote the Lagrange function of this problem.

Kuhn-Tucker condition. The conditions Definition 18.18

$$\begin{aligned}
&\frac{\partial\mathscr{L}}{\partial x_j} \le 0, \quad x_j \ge 0 \quad \text{and} \quad x_j \frac{\partial\mathscr{L}}{\partial x_j} = 0 \\
&\frac{\partial\mathscr{L}}{\partial \lambda_i} \ge 0, \quad \lambda_i \ge 0 \quad \text{and} \quad \lambda_i \frac{\partial\mathscr{L}}{\partial \lambda_i} = 0
\end{aligned}$$

are called the **Kuhn-Tucker conditions** of the problem.

Kuhn-Tucker sufficient condition. Suppose that f and $\mathbf{g}$ are $\mathscr{C}^1$ functions and there exists a $\boldsymbol{\lambda}^* = (\lambda_1^*, \dots, \lambda_m^*)$ and an admissible $\mathbf{x}^*$ such that Theorem 18.19

(1) The objective function f is concave.

(2) The functions g_j are convex for $j = 1, \dots, m$.

(3) The point $\mathbf{x}^*$ satisfies the Kuhn-Tucker conditions.

Then $\mathbf{x}^*$ solves the constraint maximization problem.

See Lecture 13 in *Mathematische Methoden* for examples.

— Exercises

18.1 Compute all local and global extremal points of the functions

(a) $f(x) = (x-3)^6$

(b) $g(x) = \frac{x^2+1}{x}$

18.2 Compute the local and global extremal points of the functions

(a) $f\colon [0,\infty] \to \mathbb{R}, x \mapsto \frac{1}{x} + x$

(b) $f\colon [0,\infty] \to \mathbb{R}, x \mapsto \sqrt{x} - x$

(c) $g\colon \mathbb{R} \to \mathbb{R}, x \mapsto e^{-2x} + 2x$

18.3 Compute all local extremal points and saddle points of the following functions. Are the local extremal points also globally extremal.

(a) $f(x,y) = -x^2 + xy + y^2$

(b) $f(x,y) = \frac{1}{x}\ln(x) - y^2 + 1$

(c) $f(x,y) = 100(y-x^2)^2 + (1-x)^2$

(d) $f(x,y) = 3x + 4y - e^x - e^y$

18.4 Compute all local extremal points and saddle points of the following functions. Are the local extremal points also globally extremal.

$$f(x_1,x_2,x_3) = (x_1^3 - x_1)x_2 + x_3^2 .$$

18.5 We are given the following constraint optimization problem

$$\max(\min) \quad f(x,y) = x^2 y \quad \text{subject to} \quad x + y = 3.$$

(a) Solve the problem graphically.

(b) Compute all stationary points.

(c) Use the bordered Hessian to determine whether these stationary points are (local) maxima or minima.

18.6 Compute all stationary points of the constraint optimization problem

$$\max\ (\min) \quad f(x_1,x_2,x_3) = \frac{1}{3}(x_1-3)^3 + x_2 x_3$$

$$\text{subject to} \quad x_1 + x_2 = 4 \text{ and } x_1 + x_3 = 5.$$

18.7 A household has an income m and can buy two commodities with prices p_1 and p_2. We have

$$p_1 x_1 + p_2 x_2 = m$$

where x_1 and x_2 denote the quantities. Assume that the household has a utility function

$$u(x_1,x_2) = \alpha \ln(x_1) + (1-\alpha)\ln(x_2)$$

where $\alpha \in (0,1)$.

(a) Solve this constraint optimization problem.

(b) Compute the change of the optimal utility function when the price of commodity 1 changes.

(c) Compute the change of the optimal utility function when the income m changes.

18.8 We are given the following constraint optimization problem

$$\max \quad f(x,y) = -(x-2)^2 - y \quad \text{subject to} \quad x+y \le 1, \quad x,y \ge 0\,.$$

(a) Solve the problem graphically.

(b) Solve the problem by means of the Kuhn-Tucker conditions.

— Problems

18.9 Our definition of a local maximum (Definition 18.6) is quite simple but has unexpected consequences: There exist functions where a global minimum is a local maximum. Give an example for such a function. How could Definition 18.6 be "repaired"?

18.10 Let $f\colon \mathbb{R}^n \to \mathbb{R}$ and $T\colon \mathbb{R} \to \mathbb{R}$ be a strictly monotonically increasing transformation. Show that $\mathbf{x}^*$ is a maximum of f if and only if $\mathbf{x}^*$ is a maximum of the transformed function $T \circ f$.

19

Integration

We know the boundary of some domain. What is its area?

In this chapter we deal with two topics that seem to be quite distinct: We want to invert the result of differentiation and we want to compute the area of a region that is enclosed by curves. These two tasks are linked by the Fundamental Theorem of Calculus.

19.1 The Antiderivative of a Function

Definition 19.1

A univariate function F is called an **antiderivative** of some function f if $F'(x) = f(x)$.

Motivated by the Fundamental Theorem of Calculus (p. 206) the antiderivative is usually called the **indefinite integral** (or **primitive integral**) of f and denoted by

$$F(x) = \int f(x)\,dx\,.$$

Finding antiderivatives is quite a hard issue. In opposition to differentiation often no straightforward methods exist. Roughly spoken we have to do the following:

Make an educated guess and verify by differentiation.

Example 19.2

Find the antiderivative of $f(x) = \ln(x)$.

SOLUTION. Guess: $F(x) = x(\ln(x) - 1)$.
Verify: $F'(x) = (x(\ln(x)-1))' = 1 \cdot (\ln(x)-1) + x \cdot \frac{1}{x} = \ln(x)$. ◊

It is quite obvious that $F(x) = x(\ln(x) - 1) + 123$ is also an antiderivative of $\ln(x)$ as is $F(x) = x(\ln(x) - 1) + c$ for every $c \in \mathbb{R}$.

Lemma 19.3

If $F(x)$ is an antiderivative of some function $f(x)$, then $F(x) + c$ is also an antiderivative of $f(x)$ for every $c \in \mathbb{R}$. The constant c is called the **integration constant**.

$f(x)$	$\int f(x)\,dx$	
0	c	
x^α	$\frac{1}{\alpha+1}\cdot x^{\alpha+1}+c$	for $\alpha\neq -1$
e^x	e^x+c	
$\frac{1}{x}$	$\ln\lvert x\rvert+c$	
$\cos(x)$	$\sin(x)+c$	
$\sin(x)$	$-\cos(x)+c$	

Table 19.4

Table of antide
of some elemen
functions.

Summation rule:	$\int \alpha f(x)+\beta g(x)\,dx = \alpha\int f(x)\,dx+\beta\int g(x)\,dx$
By parts:	$\int f(x)\cdot g'(x)\,dx = f(x)\cdot g(x)-\int f'(x)\cdot g(x)\,dx$
By substitution:	$\int f\big(g(x)\big)\cdot g'(x)\,dx = \int f(z)\,dz$ where $z=g(x)$ and $dz=g'(x)\,dx$

Table 19.5

Rules for indef
integrals.

Fortunately there exist some tools that ease the task of “guessing” the antiderivative. Table 19.4 lists basic integrals. Observe that we get these antiderivatives simply by exchanging the columns in our table of derivatives (Table 14.8).

Table 19.5 lists integration rules that allow to reduce the issue of finding indefinite integrals of complicated expressions to simpler ones. Again these rules can be directly derived from the corresponding rules in Table 14.9 for computing derivatives. There exist many other such rules which are, however, often only applicable to special functions. Computer algebra systems like *Maxima* thus use much larger tables for basic integrals and integration rules for finding indefinite integrals.

DERIVATION OF THE INTEGRATION RULES. The *summation rule* is just a consequence of the linearity of the differential operator.

For *integration by parts* we have to assume that both f and g are differentiable. Thus we find by means of the product rule

$$\begin{aligned} f(x)\cdot g(x) &= \int \big(f(x)\cdot g(x)\big)'\,dx = \int \big(f'(x)g(x)+f(x)g'(x)\big)\,dx \\ &= \int f'(x)g(x)\,dx+\int f(x)g'(x)\,dx \end{aligned}$$

and hence the rule follows.

For *integration by substitution* let F denote an antiderivative of f.

Then we find

$$\begin{aligned}\int f(z)\,dz &= F(z)\\ &= F\big(g(x)\big) = \int \Big(F\big(g(x)\big)\Big)'\,dx = \int F'\big(g(x)\big)\,g'(x)\,dx\\ &= \int f\big(g(x)\big)\,g'(x)\,dx\end{aligned}$$

that is, if the integrand is of the form $f\big(g(x)\big)\,g'(x)$ we first compute the indefinite integral $\int f(z)\,dz$ and then substitute $z = g(x)$. □

Compute the indefinite integral of $f(x) = 4x^3 - x^2 + 3x - 5$. Example 19.6

SOLUTION. By the summation rule we find

$$\begin{aligned}\int f(x)\,dx &= \int 4x^3 - x^2 + 3x - 5\,dx\\ &= 4\int x^3\,dx - \int x^2\,dx + 3\int x\,dx - 5\int dx\\ &= 4\frac{1}{4}x^4 - \frac{1}{3}x^3 + 3\frac{1}{2}x^2 - 5x + c\\ &= x^4 - \frac{1}{3}x^3 + \frac{3}{2}x^2 - 5x + c\,. \qquad \diamondsuit\end{aligned}$$

Compute the indefinite integral of $f(x) = x\,e^x$. Example 19.7

SOLUTION. Integration by parts yields

$$\int f(x)\,dx = \int \underbrace{x}_{f}\cdot\underbrace{e^x}_{g'}\,dx = \underbrace{x}_{f}\cdot\underbrace{e^x}_{g}\,dx - \int \underbrace{1}_{f'}\cdot\underbrace{e^x}_{g}\,dx = x\cdot e^x - e^x + c\,.$$

$$\begin{aligned}f(x) &= x &&\Rightarrow\quad f'(x) = 1\\ g'(x) &= e^x &&\Rightarrow\quad g(x) = e^x\end{aligned} \qquad \diamondsuit$$

Compute the indefinite integral of $f(x) = 2x\,e^{x^2}$. Example 19.8

SOLUTION. By substitution we find

$$\int f(x)\,dx \int \exp(\underbrace{x^2}_{g(x)})\cdot\underbrace{2x}_{g'(x)}\,dx = \int \exp(z)\,dz = e^z + c = e^{x^2} + c\,.$$

$$z = g(x) = x^2 \quad\Rightarrow\quad dz = g'(x)\,dx = 2x\,dx \qquad \diamondsuit$$

Compute the indefinite integral of $f(x) = x^2\cos(x)$. Example 19.9

SOLUTION. Integration by parts yields

$$\int f(x)\,dx \int \underbrace{x^2}_{f}\cdot\underbrace{\cos(x)}_{g'}\,dx = \underbrace{x^2}_{f}\cdot\underbrace{\sin(x)}_{g} - \int \underbrace{2x}_{f'}\cdot\underbrace{\sin(x)}_{g}\,dx\,.$$

For the last term we have to apply integration by parts again:

$$\begin{aligned}\int \underbrace{2x}_{f} \cdot \underbrace{\sin(x)}_{g'}\,dx &= \underbrace{2x}_{f} \cdot \underbrace{(-\cos(x))}_{g} - \int \underbrace{2}_{f'} \cdot \underbrace{(-\cos(x))}_{g}\,dx \\ &= -2x\cdot\cos(x) - 2\cdot(-\sin(x)) + c\,.\end{aligned}$$

Therefore we have

$$\begin{aligned}\int x^2\cos(x)\,dx &= x^2\sin(x) - \big(-2x\cos(x) + 2\sin(x) + c\big) \\ &= x^2\sin(x) + 2x\cos(x) - 2\sin(x) + c\,.\end{aligned}$$

◊

Sometimes the application of these integration rules might not be obvious as the following examples shows.

Example 19.10

Compute the indefinite integral of $f(x) = \ln(x)$.

SOLUTION. We write $f(x) = 1\cdot\ln(x)$. Integration by parts yields

$$\int \underbrace{\ln(x)}_{f}\cdot\underbrace{1}_{g'}\,dx = \underbrace{\ln(x)}_{f}\cdot\underbrace{x}_{g}\,dx - \int \underbrace{\frac{1}{x}}_{f'}\cdot\underbrace{x}_{g}\,dx = \ln(x)\cdot x - x + c$$

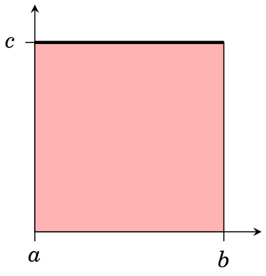
$$\begin{array}{lll} f(x) = \ln(x) & \Rightarrow & f'(x) = \frac{1}{x} \\ g'(x) = 1 & \Rightarrow & g(x) = x \end{array}$$

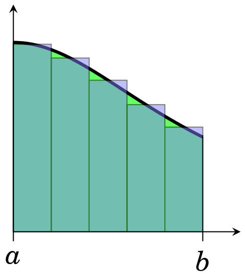
◊

We again want to note that there are no simple recipes for finding indefinite integrals. Even with integration rules like those in Table 19.5 there remains still trial and error. (Of course experience increases the change of successful guesses significantly.)

There are even two further obstacles: (1) not all functions have an antiderivative; (2) the indefinite integral may exist but it is not possible to express it in terms of elementary functions. The density of the normal distribution, $\varphi(x) = \exp(-x^2)$, is the most prominent example.

19.2 The Riemann Integral



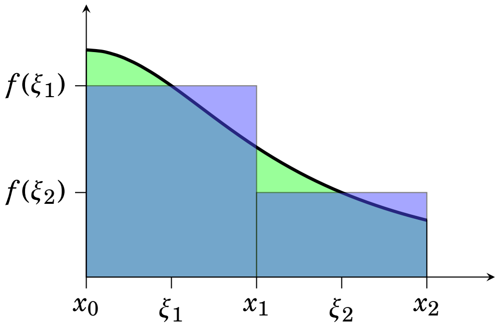


Suppose we are given some nonnegative function f over some interval $[a,b]$ and we have to compute the area A below the graph of f. If $f(x) = c$ is a constant function, then this task is quite simple: The region in question is a rectangle and we find by basic geometry (length of base × height)

$$A = c\cdot(b-a)\,.$$

For general functions with "irregular"-shaped graphs we may approximate the function by a **step function** (or *staircase function*), i.e. a piecewise constant function. The area for the step function can then be computed for each of the rectangles and added up for the total area.

Thus we select points $x_0 = 0 < x_1 < \ldots < x_n = b$ and compute f at intermediate points $\xi \in (x_{i-1}, x_i)$, for $i = 1,\ldots,n$.

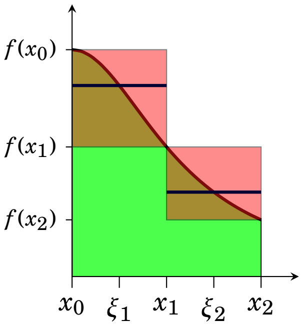


Hence we find for the area

$$A \approx \sum_{i=1}^{n} f(\xi_i) \cdot (x_i - x_{i-1}) \,.$$

If f is a monotonically decreasing function and the points $x_0, x_1, \dots, x_n$ are selected equidistant, i.e., $(x_i - x_{i-1}) = \frac{1}{n}(b-a)$, then we find for the approximation error

$$\left| A - \sum_{i=1}^{n} f(\xi_i) \cdot (x_i - x_{i-1}) \right| \le (f_{\max} - f_{\min})(b-a)\frac{1}{n} \to 0 \quad \text{as } n \to \infty.$$



Thus when we increase the number of points n, then this so called *Riemann sum* converges to area A. For a nonmonotone function the limit may not exist. If it exists we get the area under the graph.

Definition 19.11

Riemann integral. Let f be some function defined on $[a,b]$. Let $(Z_k) = \left(\left\{x_0^{(k)}, x_1^{(k)}, \dots, x_n^{(k)}\right\}\right)$ be a sequence of point sets such that $x_0^{(k)} = a < x_1^{(k)} < \dots x_{k-1}^{(k)} < x_k^{(k)} = b$ for all $k = 1, 2, \dots$ and $\max_{i=1,\dots,k}\left(x_i^{(k)} - x_{i-1}^{(k)}\right) \to 0$ as $k \to \infty$. Let $\xi_i^{(k)} \in \left(x_{i-1}^{(k)}, x_i^{(k)}\right)$. If the **Riemann sum**

$$I_k = \sum_{i=1}^{k} f(\xi_i^{(k)}) \cdot \left(x_i^{(k)} - x_{i-1}^{(k)}\right)$$

converges for all such sequences (Z_k) then the function $f\colon [a,b] \to \mathbb{R}$ is **Riemann integrable**. The limit is called the **Riemann integral** of f and denoted by

$$\int_a^b f(x)\,dx = \lim_{k\to\infty} \sum_{i=1}^{k} f(\xi_i^{(k)}) \cdot \left(x_i^{(k)} - x_{i-1}^{(k)}\right) .$$

This definition requires some remarks.

- This limit (if it exists) is uniquely determined.
- Not all functions are Riemann integrable, that is, there exist functions where this limit does not exist for some choices of sequence (Z_k). However, bounded “nice” (in particular continuous) functions are always Riemann integrable.

Let f and g be integrable functions and $\alpha, \beta \in \mathbb{R}$. Then we find

$$\int_a^b (\alpha f(x) + \beta g(x))\,dx = \alpha \int_a^b f(x)\,dx + \beta \int_a^b g(x)\,dx$$

$$\int_a^b f(x)\,dx = -\int_b^a f(x)\,dx$$

$$\int_a^a f(x)\,dx = 0$$

$$\int_a^c f(x)\,dx = \int_a^b f(x)\,dx + \int_b^c f(x)\,dx$$

$$\int_a^b f(x)\,dx \le \int_a^b g(x)\,dx \qquad \text{if } f(x) \le g(x) \text{ for all } x \in [a,b]$$

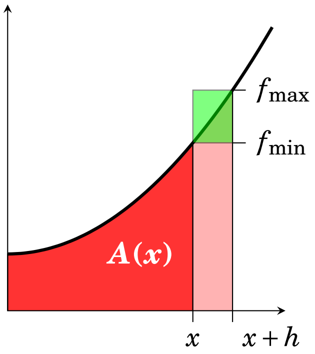
Table 19.12

Properties of d integrals.

- There also exist other concepts of integration. However, for continuous functions these coincide. Thus we will say **integrable** and **integral** for short.
- As we will see in the next section integrals are usually called *definite integrals*.
- From the definition of the integral we immediately see that for regions where function f is *negative* the integral also is *negative*.
- Similarly, as the definition of *Riemann sum* contains the term $\left(x_i^{(k)} - x_{i-1}^{(k)}\right)$ instead of its absolute value $\left|x_i^{(k)} - x_{i-1}^{(k)}\right|$, the integral of a *positive* function becomes *negative* if the interval (a,b) is traversed from right to left.

Table 19.12 lists important properties of the definite integral. These can be derived from the definition of integrals and the rules for limits (Theorem 14.3 on p. 136).

19.3 The Fundamental Theorem of Calculus



We have defined the integral as the limit of Riemann sums. However, we still need a efficient method to compute the integral. On the other hand we did not establish any condition that ensure the existence of the antiderivative of a given function. Astonishingly these two apparently distinct problems are closely connected.

Let f be some *continuous* function and suppose that the area of f under the graph in the interval $[0,x]$ is given by $A(x)$. We then get the area under the curve of f in the interval $[x,x+h]$ for some h by subtraction, $A(x+h) - A(x)$. As f is continuous it has a maximum $f_{\max}(h)$ and a minimum $f_{\min}(h)$ on $[x, x+h]$. Then we find

$$\begin{aligned} f_{\min}(h) \cdot h &\le A(x+h) - A(x) \le f_{\max}(h) \cdot h \\ f_{\min}(h) &\le \frac{A(x+h)-A(x)}{h} \le f_{\max}(h) \end{aligned}$$

If $h \to 0$ we then find by continuity of f,

$$\lim_{h\to 0} f_{\min}(h) = \lim_{h\to 0} f_{\max}(h) = f(x)$$

and hence

$$f(x) \le \underbrace{\lim_{h\to 0} \frac{A(x+h)-A(x)}{h}}_{=A'(x)} \le f(x) .$$

Consequently, $A(x)$ is differentiable and we arrive at

$$A'(x) = f(x)$$

that is, $A(x)$ is an antiderivative of f.

This observation is formally stated in the two parts of the Fundamental Theorem of Calculus which we state without a stringent proof.

Theorem 19.13

First fundamental theorem of calculus. Let $f\colon [a,b] \to \mathbb{R}$ be a continuous function that admits an antiderivative F on $[a,b]$, then

$$\int_a^b f(x)\,dx = F(b) - F(a) .$$

Theorem 19.14

Second fundamental theorem of calculus. Let $f\colon [a,b] \to \mathbb{R}$ be a continuous function and F defined for all $x \in [a,b]$ as the integral

$$F(x) = \int_a^x f(t)\,dt .$$

Then F is differentiable on (a,b) and

$$F'(x) = f(x) \quad \text{for all } x \in (a,b).$$

An immediate corollary is that every continuous function has an antiderivative, namely the integral function F.

Notice that the first part states that we simply can use the indefinite integral to compute the integral of continuous functions, $\int_a^b f(x)\,dx$. In contrast, the second part gives us a sufficient condition for the existence of the antiderivative of a function.

By parts: $$\int_a^b f(x)g'(x)\,dx = f(x)g(x)\Big|_a^b - \int_a^b f'(x)g(x)\,dx$$

By substitution: $$\int_a^b f(g(x))\cdot g'(x)\,dx = \int_{g(a)}^{g(b)} f(z)\,dz$$

where $z = g(x)$ and $dz = g'(x)\,dx$

Table 19.17

Rules for defin
integrals.

19.4 The Definite Integral

Theorem 19.13 provides a method to compute the integral of a function without dealing with limits of Riemann sums. This motivates the term *definite integral*.

Definition 19.15

Let $f\colon [a,b] \to \mathbb{R}$ be a continuous function and F an antiderivative of f. Then

$$\int_a^b f(x)\,dx = F(x)\Big|_a^b = F(b) - F(a)$$

is called the **definite integral** of f.

Example 19.16

Compute the definite integral of $f(x) = x^2$ in the interval $[0,1]$.

SOLUTION. $\int_0^1 x^2\,dx = \frac{1}{3}x^3\Big|_0^1 = \frac{1}{3}\cdot 1^3 - \frac{1}{3}\cdot 0^3 = \frac{1}{3}$. ♢

The rules for indefinite integrals in Table 19.5 can be easily translated into rules for the definite integral, see Table 19.17

Example 19.18

Compute $\int_e^{10} \frac{1}{\log(x)}\cdot\frac{1}{x}\,dx$.

SOLUTION.

$$\int_e^{10} \frac{1}{\log(x)}\cdot\frac{1}{x}\,dx = \int_1^{\log(10)} \frac{1}{z}\,dz$$

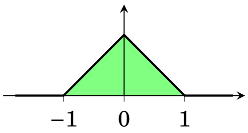
$$z = \log(x) \quad\Rightarrow\quad dz = \frac{1}{x}\,dx$$

$$= \log(z)\Big|_1^{\log(10)} = \log(\log(10)) - \log(1) \approx 0.834\,.$$ ♢

Example 19.19

Compute $\int_{-2}^{2} f(x)\,dx$ where

$$f(x) = \begin{cases} 1+x & \text{for } -1 \le x < 0 \\ 1-x & \text{for } 0 \le x < 1 \\ 0 & \text{for } x < -1 \text{ and } x \ge 1 \end{cases}$$



SOLUTION.

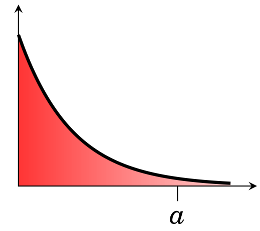
$$\begin{aligned}\int_{-2}^{2} f(x)\,dx &= \int_{-2}^{-1} f(x)\,dx + \int_{-1}^{0} f(x)\,dx + \int_{0}^{1} f(x)\,dx + \int_{1}^{2} f(x)\,dx \\ &= \int_{-2}^{-1} 0\,dx + \int_{-1}^{0} (1+x)\,dx + \int_{0}^{1} (1-x)\,dx + \int_{1}^{2} 0\,dx \\ &= \left(x + \frac{1}{2}x^2\right)\Big|_{-1}^{0} + \left(x - \frac{1}{2}x^2\right)\Big|_{0}^{1} \\ &= \frac{1}{2} + \frac{1}{2} = 1\,. \qquad \diamondsuit\end{aligned}$$

19.5 Improper Integrals

Example 19.20

Suppose we want to compute $\int_0^b e^{-\lambda x}\,dx$. We then get

$$\int_0^b e^{-\lambda x}\,dx = \int_0^{-\lambda b} e^z \left(-\tfrac{1}{\lambda}\right) dz = -\tfrac{1}{\lambda} e^z \Big|_0^{-\lambda b} = \frac{1}{\lambda}\left(1 - e^{-\lambda b}\right) . \qquad \diamondsuit$$



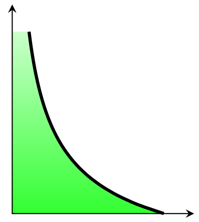
So what happens if b tends to ∞, i.e., when the domain of integration is unbounded. Obviously

$$\lim_{b\to\infty} \int_0^b e^{-\lambda x}\,dx = \lim_{b\to\infty} \frac{1}{\lambda}\left(1 - e^{-\lambda b}\right) = \frac{1}{\lambda}\,.$$

Thus we may use the symbol

$$\int_0^\infty f(x)\,dx$$

for this limit. Similarly we may want to compute the integral $\int\limits_0^1 \frac{1}{\sqrt{x}}\,dx$. But then 0 does not belong to the domain of f as $f(0)$ is not defined. We then replace the lower bound 0 by some $a > 0$, compute the integral and find the limit for $a \to 0$. We again write



$$\int_0^1 \frac{1}{\sqrt{x}}\,dx = \lim_{a\to 0^+} \int_a^1 \frac{1}{\sqrt{x}}\,dx$$

where 0^+ indicates that we are looking at the limit from the right hand side.

Definition 19.21

Integrals of functions that are unbounded at a or b or have unbounded domain (i.e., $a = -\infty$ or $b = \infty$) are called **improper integrals**. They are defined as limits of proper integrals. If the limit

$$\int_0^b f(x)\,dx = \lim_{t\to b} \int_0^t f(x)\,dx$$

exists we say that the improper integral *converges*. Otherwise we say that it *diverges*.

For practical reasons we demand that this limit exists if and only if $\lim_{t\to\infty}\int_0^t |f(x)|\,dx$ exists.

Example 19.22 Compute the improper integral $\int_0^1 \frac{1}{\sqrt{x}}\,dx$.

SOLUTION.

$$\int_0^1 \frac{1}{\sqrt{x}}\,dx = \lim_{t\to 0}\int_t^1 x^{-\frac{1}{2}}\,dx = \lim_{t\to 0} 2\sqrt{x}\Big|_t^1 = \lim_{t\to 0}(2-2\sqrt{t}) = 2\,. \qquad \diamondsuit$$

Example 19.23 Compute the improper integral $\int_1^\infty \frac{1}{x^2}\,dx$.

SOLUTION.

$$\int_1^\infty \frac{1}{x^2}\,dx = \lim_{t\to\infty}\int_1^t x^{-2}\,dx = \lim_{t\to\infty} -\frac{1}{x}\Big|_1^t = \lim_{t\to\infty} -\frac{1}{t} - (-1) = 1\,. \qquad \diamondsuit$$

Example 19.24 Compute the improper integral $\int_1^\infty \frac{1}{x}\,dx$.

SOLUTION.

$$\int_1^\infty \frac{1}{x}\,dx = \lim_{t\to\infty}\int_1^t \frac{1}{x}\,dx = \lim_{t\to\infty} \log(x)\Big|_1^t = \lim_{t\to\infty} \log(t) - \log(1) = \infty\,.$$

The improper integral diverges. ♢

19.6 Differentiation under the Integral Sign

We are given some continuous function f with antiderivative F, i.e., $F'(x) = f(x)$. If we differentiate the definite integral $\int_a^x f(t)\,dt = F(x) - F(a)$ w.r.t. its upper bound we obtain

$$\frac{d}{dx}\int_a^x f(t)\,dt = (F(x)-F(a))' = F'(x) = f(x)\,.$$

That is, the derivative of the definite integral w.r.t. the upper limit of integration is equal to the integrand evaluated at that point.

We can generalize this result. Suppose that both lower and upper limit of the definite integral are differentiable functions $a(x)$ and $b(x)$, respectively. Then we find by the chain rule

$$\frac{d}{dx}\int_{a(x)}^{b(x)} f(t)\,dt = (F(b(x)) - F(a(x)))' = f(b(x))\,b'(x) - f(a(x))\,a'(x)\,.$$

Now suppose that $f(x,t)$ is a continuous function of two variables and consider the function $F(x)$ defined by

$$F(x) = \int_a^b f(x,t)\,dt\,.$$

Its derivative $F'(x)$ can be computed as

$$\begin{aligned}F'(x) &= \lim_{h\to 0} \frac{F(x+h)-F(x)}{h}\\ &= \lim_{h\to 0}\int_a^b \frac{f(x+h,t)-f(x,t)}{h}\,dt\\ &= \int_a^b \lim_{h\to 0} \frac{f(x+h,t)-f(x,t)}{h}\,dt\\ &= \int_a^b \frac{\partial f(x,t)}{\partial x}\,dt\,.\end{aligned}$$

That is, in order to get the derivative of the integral with respect to parameter x we differentiate under the integral sign.

Of course the partial derivative $f_x(x,t)$ must be an integrable function which is satisfied whenever it is continuous by the Fundamental Theorem.

It is important to note that both the (Riemann-) integral and the partial derivative are limits. Thus we have to exchange these two limits in our calculation. Notice, however, that this is a problematic step and its validation requires tools from advanced calculus.

In general exchanging limits can change the result!

We now can combine our observations into a single formula.

Theorem 19.25

Leibniz's formula. Let

$$F(x) = \int_{a(x)}^{b(x)} f(x,t)\,dt$$

where the function $f(x,t)$ and its partial derivative $f_x(x,t)$ are continuous in both x and t in the region $\{(x,t)\colon x_0 \le x \le x_1, a(x) \le t \le b(x)\}$ and the functions $a(x)$ and $b(x)$ are $\mathscr{C}^1$ functions over $[x_0, x_1]$. Then

$$F'(x) = f(x,b(x))\,b'(x) - f(x,a(x))\,a'(x) + \int_{a(x)}^{b(x)} \frac{\partial f(x,t)}{\partial x}\,dt\,.$$

PROOF. Let $H(x,a,b) = \int\limits_a^b f(x,t)\,dt$. Then $F(x) = H(x,a(x),b(x))$ and we find by the chain rule

$$F'(x) = H_x + H_a a'(x) + H_b b'(x)\,.$$

Since $H_x = \int_a^b f_x(x,t)\,dt$, $H_a = -f(x,a)$ and $H_b = f(x,b)$, the result follows. □

Example 19.26

Compute $F'(x)$, $x \ge 0$, when $F(x) = \int_x^{2x} tx^2\,dt$.

SOLUTION. Let $f(x,t) = tx^2$, $a(x) = x$ and $b(x) = 2x$. By Leibniz's formula we find

$$\begin{aligned}F'(x) &= (2x)\cdot x^2\cdot 2 - (x)\cdot x^2 \cdot 1 + \int_x^{2x} 2xt\,dt\\ &= 4x^3 - x^3 + (2x\frac{1}{2}t^2)\Big|_x^{2x} = 4x^3 - x^3 + (4x^3 - x^3) = 6x^3\,.\end{aligned}$$

♢

Leibniz formula also works for improper integrals provided that the integral $\int_{a(x)}^{b(x)} f'_x(x,t)\,dt$ converges:

$$\frac{d}{dx}\int_a^{\infty} f(x,t)\,dt = \int_a^{\infty} \frac{\partial f(x,t)}{\partial x}\,dt$$

Example 19.27

Let $K(t)$ denote the capital stock of some firm at time t, and let $p(t)$ be the price per unit of capital. Let $R(t)$ denote the rental price per unit of capital and let r be some constant interest rate. In capital theory, one principle for the determining of the correct price of the firm's capital is given by the equation

$$p(t)K(t) = \int_t^{\infty} R(\tau)K(\tau)e^{-r(\tau-t)}\,d\tau \qquad \text{for all } t.$$

That is, the current cost of capital should equal the discounted present value of the returns from lending it. Find an expression for $R(t)$ by differentiating the equation w.r.t. t.

SOLUTION. By differentiation the left hand side using the product rule and the right hand side using Leibniz's formula we arrive at

$$\begin{aligned} p'(t)K(t) + p(t)K'(t) &= -R(t)K(t) + \int_t^{\infty} R(\tau)K(\tau)\,r\,e^{-r(\tau-t)}\,d\tau \\ &= -R(t)K(t) + r\,p(t)K(t) \end{aligned}$$

and consequently

$$R(t) = \left(r - \frac{K'(t)}{K(t)}\right)p(t) - p'(t)\,. \qquad \diamondsuit$$

— Exercises

19.1 Compute the following indefinite integrals:

(a) $\int x\ln(x)\,dx$ (b) $\int x^2\sin(x)\,dx$ (c) $\int 2x\sqrt{x^2+6}\,dx$

(d) $\int e^{x^2}x\,dx$ (e) $\int \frac{x}{3x^2+4}\,dx$ (f) $\int x\sqrt{x+1}\,dx$

(g) $\int \frac{3x^2+4}{x}\,dx$ (h) $\int \frac{\ln(x)}{x}\,dx$

19.2 Compute the following definite integrals:

(a) $\int_1^4 2x^2-1\,dx$ (b) $\int_0^2 3e^x\,dx$

(c) $\int_1^4 3x^2+4x\,dx$ (d) $\int_0^{\frac{\pi}{3}} \frac{-\sin(x)}{3}\,dx$

(e) $\int_0^1 \frac{3x+2}{3x^2+4x+1}\,dx$

19.3 Compute the following improper integrals:

(a) $\int_0^\infty -e^{-3x}\,dx$ (b) $\int_0^1 \frac{2}{\sqrt[4]{x^3}}\,dx$ (c) $\int_0^\infty \frac{x}{x^2+1}\,dx$

19.4 The marginal costs for a cost function $C(x)$ are given by $30-0.05x$. Reconstruct $C(x)$ when the fixed costs are €2000.

19.5 Compute the expectation of a so called *half-normal* distributed random variate which has domain $[0,\infty)$ and probability density function

$$f(x)=\sqrt{\frac{2}{\pi}}\exp\left(-\frac{x^2}{2}\right).$$

HINT: The **expectation** of a random variate X with density f is defined as

$$\mathbb{E}(X)=\int_{-\infty}^{\infty} x\,f(x)\,dx\,.$$

19.6 Compute the expectation of a normal distributed random variate with probability density function

$$f(x)=\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right).$$

HINT: $\int_{-\infty}^{\infty} f(x)\,dx = \int_{-\infty}^{0} f(x)\,dx+\int_{0}^{\infty} f(x)\,dx$

— Problems

19.7 For which value of $\alpha \in \mathbb{R}$ do the following improper integrals converge? What are their values?

(a) $\displaystyle\int_0^1 x^\alpha\,dx$ (b) $\displaystyle\int_1^\infty x^\alpha\,dx$ (c) $\displaystyle\int_0^\infty x^\alpha\,dx$

HINT: Show that the improper integral $\int_{-\infty}^{\infty} f(x)\,dx$ diverges.

19.8 Let X be a so called *Cauchy* distributed random variate with probability density function

$$f(x) = \frac{1}{\pi(1+x^2)}\,.$$

Show that X does not have an expectation.

Why is the following approach incorrect?

$$\mathbb{E}(X) = \lim_{t\to\infty}\int_{-t}^{t}\frac{x}{\pi(1+x^2)}\,dx = \lim_{t\to\infty} 0 = 0\,.$$

19.9 Compute for $T \geq 0$

$$\frac{d}{dx}\int_0^{g(x)} U(x)\,e^{-(t-T)}\,dt\,.$$

Which conditions on $g(x)$ and $U(x)$ must be satisfied?

19.10 Let f be the probability density function of some absolutely continuous distributed random variate X. The *moment generating function* of f is defined as

$$M(t) = \mathbb{E}(e^{tX}) = \int_{-\infty}^{\infty} e^{tx} f(x)\,dx\,.$$

Show that $M'(0) = \mathbb{E}(X)$, i.e., the expectation of X.

19.11 The gamma function $\Gamma(z)$ is an extension of the factorial function. That is, if n is a positive integer, then

$$\Gamma(n) = (n-1)!$$

For positive real numbers z it is defined as

$$\Gamma(z) = \int_0^\infty t^{z-1} e^{-t}\,dt\,.$$

(a) Use integration by parts and show that

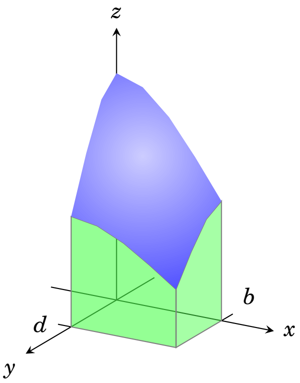
$$\Gamma(z+1) = z\,\Gamma(z)\,.$$

(b) Compute $\Gamma'(z)$ by means of Leibniz's formula.

20

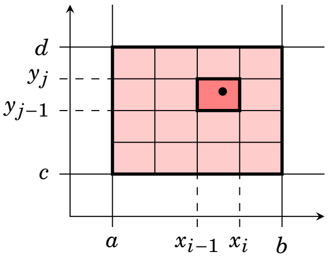
Multiple Integrals

What is the volume *of a smooth mountain?*



The idea of Riemann integration can be extended to the computation of volumes under the graph of bivariate and multivariate functions. However, difficulties arise as the domain of such functions are not simple intervals in general but can be irregular shaped regions.

20.1 The Riemann Integral



Let us start with the simple case where the domain of some bivariate function is the Cartesian product of two closed intervals, i.e., a rectangle

$$R = [a,b] \times [c,d] = \{(x,y) \in \mathbb{R}^2 : a \le x \le b,\ c \le y \le d\}\,.$$

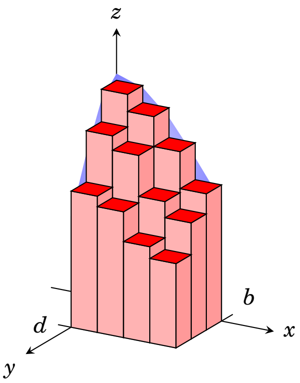
Analogously to Section 19.2 we partition R into rectangular subregions

$$R_{ij} = [x_{i-1}, x_i] \times [y_{j-1}, y_j] \qquad \text{for } 1 \le i \le n \text{ and } 1 \le j \le k$$

where $a = x_0 < x_1 < \ldots < x_n = b$ and $c = y_0 < y_1 < \ldots < y_k = d$.

For $f\colon R \subset \mathbb{R}^2 \to \mathbb{R}$ we compute $f(\xi_i, \zeta_i)$ for points $(\xi_i, \zeta_j) \in R_{ij}$ and approximate the volume V under the graph of f by the Riemann sum

$$V \approx \sum_{i=1}^{n} \sum_{j=1}^{k} f(\xi_i, \zeta_j)(x_i - x_{i-1})(y_j - y_{j-1})\,,$$



Observe that $(x_i - x_{i-1})(y_j - y_{j-1})$ simply is the area of rectangle R_{ij}. Each term in this sum is just the volume of the bar $[x_{i-1}, x_i] \times [y_{j-1}, y_j] \times [0, f(\xi_i, \zeta_j)]$.

Now suppose that we refine this partition such that the diameter of the largest rectangle tends to 0. If the Riemann sum converges for every such sequence of partitions for arbitrarily chosen points (ξ_i, ζ_i) then this limit is called the **double integral** of f over R and denoted by

$$\iint_R f(x,y)\,dx\,dy = \lim_{\operatorname{diam}(R_{ij}) \to 0} \sum_{i=1}^{n} \sum_{j=1}^{k} f(\xi_i, \zeta_j)(x_i - x_{i-1})(y_j - y_{j-1})\,.$$

Table 20.1

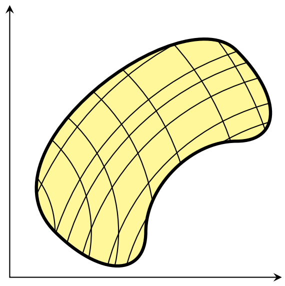
Properties of d integrals.

Let f and g be integrable functions over some domain D. Let D_1, D_2 be a partition of D, i.e., $D_1 \cup D_2 = D$ and $D_1 \cap D_2 = \emptyset$. Then we find

$$\iint_D (\alpha f(x,y) + \beta g(x,y))\,dx\,dy$$
$$= \alpha \iint_D f(x,y)\,dx\,dy + \beta \iint_D g(x,y)\,dx\,dy$$

$$\iint_D f(x,y)\,dx = \iint_{D_1} f(x,y)\,dx\,dy + \iint_{D_2} f(x,y)\,dx\,dy$$

$$\iint_D f(x,y)\,dx\,dy \le \iint_D g(x,y)\,dx\,dy$$
$$\text{if } f(x,y) \le g(x,y) \text{ for all } (x,y) \in D$$



It must be noted here that for the definition of the Riemann integral the partition of R need not consist of rectangles. Thus the same idea also works for non-rectangular domains D which may have a quite irregular shape. However, the process of convergence requires more technical details than for the case of univariate functions. For example, the partition has to consist of subdomains D_i of D for which we can determine their areas. Then we have

$$\iint_D f(x,y)\,dx\,dy = \lim_{\operatorname{diam}(D_i)\to 0} \sum_{i=1}^{n} f(\xi_i,\zeta_i)A(D_i)$$

where $A(D_i)$ denotes the area of subdomain D_i. Of course this only works if this limit exists and if it is independent from the particular partition D_i and the choice of the points $(\xi_i,\zeta_i) \in D_i$.

By this definition we immediately get properties that are similar to those of definite integrals, see Table 20.1.

20.2 Double Integrals over Rectangles

As far we only have a concept for the volume below the graph of a bivariate function. However, we also need a convenient method to compute it. So let us again assume that f is a continuous positive function defined on a rectangular domain $R = [a,b] \times [c,d]$. We then write

$$\iint_R f(x,y)\,dx\,dy = \int_a^b \int_c^d f(x,y)\,dy\,dx$$

in analogy to univariate definite integrals.

Let t be an arbitrary point in $[a,b]$. Then let $V(t)$ denote the volume

$$V(t) = \int_a^t \int_c^d f(x,y)\,dy\,dx\,.$$

We also obtain a univariate function $g(y) = f(t,y)$ defined on the interval $[c,d]$. Thus

$$A(t) = \int_c^d g(y)\,dy = \int_c^d f(t,y)\,dy$$

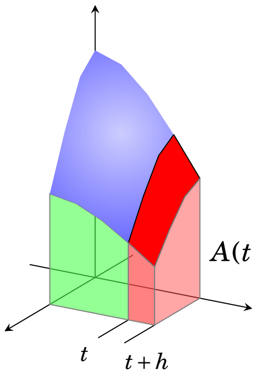
is the area of the (2-dimensional) set $\{(t,y,z)\colon 0 \le z \le f(t,y),\ y \in [c,d]\}$. Hence we find

$$V(t+h) - V(t) \approx A(t)\cdot h$$

and consequently

$$V'(t) = \lim_{h\to 0} \frac{V(t+h)-V(t)}{h} = A(t)$$

that is, $V(t)$ is an antiderivative of $A(t)$. Here we have used (but did not formally proof) that $A(t)$ is also a continuous function of t.



By this observation we only need to compute the definite integral $\int_c^d f(t,y)\,dy$ for every t and obtain some function $A(t)$. Then we compute the definite integral $\int_a^b A(x)\,dx$. In other words:

$$\iint_R f(x,y)\,dx\,dy = \int_a^b \left(\int_c^d f(x,y)\,dy\right) dx\,.$$

For that reason $\iint_R f(x,y)\,dx\,dy$ is called *double integral*.

Obviously our arguments remain valid if we exchange the rôles of x and y. Thus

$$\iint_R f(x,y)\,dx\,dy = \int_c^d \left(\int_a^b f(x,y)\,dx\right) dy\,.$$

We summarize our results in the following theorem which we state without a formal proof.

Theorem 20.2

Fubini's theorem. Let $f\colon R = [a,b]\times[c,d] \subset \mathbb{R}^2 \to \mathbb{R}$ be a continuous function. Then

$$\iint_R f(x,y)\,dx\,dy = \int_a^b \left(\int_c^d f(x,y)\,dy\right) dx = \int_c^d \left(\int_a^b f(x,y)\,dx\right) dy\,.$$

By this theorem we have the following recipe to compute the double integral of a continuous function $f(x,y)$ defined on the rectangle $[a,b]\times[c,d]$.

(1) Keep y fixed and compute the inner integral w.r.t. x from $x = a$ to $x = b$. This gives $\int_a^b f(x,y)\,dx$, a function of y.

(2) Now integrate $\int_a^b f(x,y)\,dx$ w.r.t. y from $y = c$ to $y = d$ to obtain $\int_c^d \left(\int_a^b f(x,y)\,dx\right) dy$.

Of course we can reverse the order of integration, that is, we first compute $\int_c^d f(x,y)\,dy$ and obtain a function of x which is then integrated w.r.t. x and obtain $\int_a^b \left(\int_c^d f(x,y)\,dy\right) dx$. By Fubini's theorem the results of these two procedures coincide.

Example 20.3

Compute $\int_{-1}^{1}\int_{0}^{1}(1-x-y^2+xy^2)\,dx\,dy$.

SOLUTION. We have to integrate two times.

$$\int_{-1}^{1}\int_{0}^{1}(1-x-y^2+xy^2)\,dx\,dy=\int_{-1}^{1}\left(x-\frac{1}{2}x^2-xy^2+\frac{1}{2}x^2y^2\bigg|_0^1\right)dy$$
$$=\int_{-1}^{1}\left(\frac{1}{2}-\frac{1}{2}y^2\right)dy=\frac{1}{2}y-\frac{1}{6}y^3\bigg|_{-1}^{1}=\frac{1}{2}-\frac{1}{6}-\left(-\frac{1}{2}+\frac{1}{6}\right)=\frac{2}{3}\,.$$

We can also perform the integration in the reverse order.

$$\int_{0}^{1}\int_{-1}^{1}(1-x-y^2+xy^2)\,dy\,dx=\int_{0}^{1}\left(y-xy-\frac{1}{3}y^3+\frac{1}{3}xy^3\bigg|_{-1}^{1}\right)dx$$
$$=\int_{0}^{1}\left(1-x-\frac{1}{3}+\frac{1}{3}x-\left(-1+x+\frac{1}{3}-\frac{1}{3}x\right)\right)dx$$
$$=\int_{0}^{1}\left(\frac{4}{3}-\frac{4}{3}x\right)dx=\frac{4}{3}x-\frac{4}{6}x^2\bigg|_0^1=\frac{2}{3}\,.$$

We obtain the same result by both procedures. ♢

We can extend our results for multivariate functions. Let

$$\Omega=[a_1,b_1]\times\cdots\times[a_n,b_n]=\left\{(x_1,\ldots,x_n)\in\mathbb{R}^n: a_i\le x_i\le b_i,\ i=1,\ldots,n\right\}$$

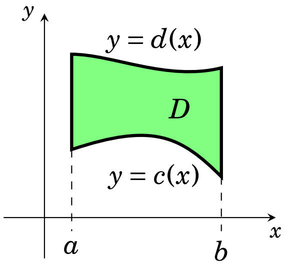
be the Cartesian product of closed intervals $[a_1,b_1],\ldots,[a_n,b_n]$. We call Ω an ***n*-dimensional rectangle**.

If $f\colon\Omega\to\mathbb{R}$ is a continuous function, then the **multiple integral** of f over Ω is defined as

$$\iint\cdots\int_{\Omega} f(x_1,\ldots,x_n)\,dx_1\ldots dx_n$$
$$=\int_{a_1}^{b_1}\left(\int_{a_2}^{b_2}\cdots\left(\int_{a_n}^{b_n} f(x_1,\ldots,x_n)\,dx_n\right)\ldots dx_2\right)dx_1\,.$$

It is important to note that the inner integrals are evaluated at first.

20.3 Double Integrals over General Domains



Consider now a domain $D\subseteq\mathbb{R}^2$ defined as

$$D=\{(x,y)\colon a\le x\le b,\ c(x)\le y\le d(x)\}$$

for two functions $c(x)$ and $d(x)$. Let $f(x,y)$ be a continuous function defined over D. As in the case of rectangular domains we can keep x fixed and compute the area

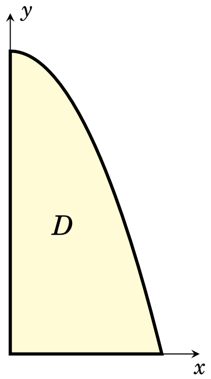
$$A(x)=\int_{c(x)}^{d(x)} f(x,y)\,dy\,.$$

We then can argue in the same way that the volume is given by

$$\iint_D f(x,y)\,dy\,dx=\int_a^b A(x)\,dx=\int_a^b\left(\int_{c(x)}^{d(x)} f(x,y)\,dy\right)dx\,.$$

Example 20.4

Let $D = \{(x,y)\colon 0 \le x \le 2,\ 0 \le y \le 4-x^2\}$ and let $f(x,y) = x^2 y$ be defined on D. Compute $\iint_D f(x,y)\,dy\,dx$.



SOLUTION.

$$\begin{aligned}
\iint_D f(x,y)\,dy\,dx &= \int_0^2 \int_0^{4-x^2} x^2 y\,dy\,dx = \int_0^2 \left(\int_0^{4-x^2} x^2 y\,dy\right) dx \\
&= \int_0^2 \left(\frac{1}{2} x^2 y^2 \Big|_0^{4-x^2}\right) dx = \int_0^2 \left(\frac{1}{2} x^2 (4-x^2)^2\right) dx \\
&= \int_0^2 \frac{1}{2}\left(x^6 - 8x^4 + 16x^2\right) dx \\
&= \frac{1}{14}x^7 - \frac{8}{10}x^5 + \frac{16}{6}x^3 \Big|_0^2 = \frac{512}{105}\,. \qquad \diamondsuit
\end{aligned}$$

It might be convenient if we partition the domain of integration D into two disjoint regions A and B, that is, $A \cup B = D$ and $A \cap B = \emptyset$. We then find

$$\iint_{A\cup B} f(x,y)\,dx\,dy = \iint_A f(x,y)\,dx\,dy + \iint_B f(x,y)\,dx\,dy$$

provided that all integrals exist. The formula extend the corresponding rule for univariate integrals in Table 19.12 on page 206. We can extend this formula to overlapping subsets A and B. We then find

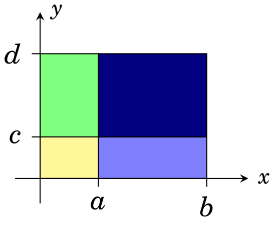
$$\begin{aligned}
&\iint_{A\cup B} f(x,y)\,dx\,dy = \\
&\qquad = \iint_A f(x,y)\,dx\,dy + \iint_B f(x,y)\,dx\,dy - \iint_{A\cap B} f(x,y)\,dx\,dy\,.
\end{aligned}$$

20.4 A “Double Indefinite” Integral

The Fundamental Theorem of Calculus tells us that we can compute a definite integral by the difference of the indefinite integral evaluated at the boundary of the domain of integration. In some sense an equivalent formula exists for double integrals.

Let $f(x,y)$ be an continuous function defined on the rectangle $[a,b] \times [c,d]$. Suppose that $F(x,y)$ has the property that

$$\frac{\partial^2 F(x,y)}{\partial x \partial y} = f(x,y) \quad \text{for all } (x,y) \in [a,b]\times[c,d].$$



Then

$$\int_a^b \int_c^d f(x,y)\,dy\,dx = F(b,d) - F(a,d) - F(b,c) + F(a,c)\,.$$

20.5 Change of Variables

Integration by substitution (see Table 19.17 on page 208) can also be seen as a change of variables. Let $x = g(z)$ where g is a differentiable one-to-one function. Set $z_1 = g^{-1}(a)$ and $z_2 = g^{-1}(b)$. Then

$$\int_a^b f(x)\,dx = \int_{z_1}^{z_2} f(g(z))\cdot g'(z)\,dz\,.$$

That is, instead of expressing f as a function of variable x we introduce a new variable z and a transformation g such that $x = g(z)$. We then integrate $f \circ g$ with respect to z. However, we have to take into account that by this change of variable the domain of integration is deformed. Thus we need the correction factor $g'(z)$.

The same idea of changing variables also works for multivariate functions.

Theorem 20.5

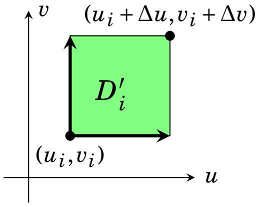
Change of variables in double integrals. Let $f(x, y)$ be a function defined on an open bounded domain $D \subset \mathbb{R}^2$. Suppose that

$$x = g(u, v), \quad y = h(u, v)$$

defines a one-to-one $\mathscr{C}^1$ transformation from an open bounded set D' onto D such that the Jacobian determinant $\frac{\partial(g,h)}{\partial(u,v)}$ is bounded and either strictly positive or strictly negative on D'. Then

$$\iint_D f(x, y)\,dx\,dy = \iint_{D'} f(g(u, v), h(u, v)) \left|\frac{\partial(g, h)}{\partial(u, v)}\right| du\,dv\,.$$

This theorem still holds if the set where $\frac{\partial(g,h)}{\partial(u,v)}$ is not bounded or vanishes is a null set, i.e., a set of area 0.



We only give a rough sketch of the proof for this formula. Let $\mathbf{g}$ denote our transformation $(u, v) \mapsto (g(u, v), h(u, v))$. Recall that

$$\iint_D f(x, y)\,dx\,dy = \lim_{\operatorname{diam}(D_i)\to 0} \sum_{i=1}^{n} f(\xi_i, \zeta_i) A(D_i)$$

where the subsets D_i are chosen as the images $\mathbf{g}(D'_i)$ of some paraxial rectangle D'_i with vertices

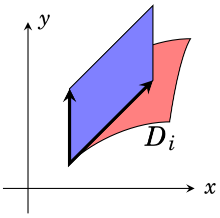
$$(u_i, v_i), \quad (u_i + \Delta u, v_i), \quad (u_i, v_i + \Delta v), \quad \text{and} \quad (u_i + \Delta u, v_i + \Delta v)$$

and $(\xi_i, \zeta_i) = \mathbf{g}(u_i, v_i) \in D_i$. Hence

$$\iint_D f(x, y)\,dx\,dy \approx \sum_{i=1}^{n} f(\xi_i, \zeta_i) A(D_i) = \sum_{i=1}^{n} f(\mathbf{g}(u_i, v_i)) A(\mathbf{g}(D'_i))\,.$$

If $\mathbf{g}(D'_i)$ were a parallelogram, then we could compute its area by means of the absolute value of the determinant

$$\begin{vmatrix} g(u_i + \Delta u, v_i) - g(u_i, v_i) & g(u_i, v_i + \Delta v) - g(u_i, v_i) \\ h(u_i + \Delta u, v_i) - h(u_i, v_i) & h(u_i, v_i + \Delta v) - h(u_i, v_i) \end{vmatrix}.$$



If $\mathbf{g}(D'_i)$ is not a parallelogram but Δu is small, then we may use this determinant as an approximation for the area $A(\mathbf{g}(D'_i))$. For small values of Δu we also have

$$g(u_i+\Delta u, v_i) - g(u_i, v_i) \approx \frac{\partial g(u_i,v_i)}{\partial u}\Delta u$$

and thus we find

$$A(\mathbf{g}(D'_i)) \approx \left| \begin{vmatrix} \frac{\partial g(u_i,v_i)}{\partial u} & \frac{\partial g(u_i,v_i)}{\partial v} \\ \frac{\partial h(u_i,v_i)}{\partial u} & \frac{\partial h(u_i,v_i)}{\partial v} \end{vmatrix} \right| \Delta u \Delta v = \left|\det(\mathbf{g}'(u_i,v_i))\right| \Delta u \Delta v \,.$$

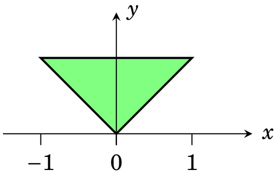
Notice that $\Delta u \Delta v = A(D'_i)$ and that we have used the symbol $\frac{\partial(g,h)}{\partial(u,v)}$ to denote the Jacobian determinant. Therefore

$$\begin{aligned}\iint_D f(x,y)\,dx\,dy &\approx \sum_{i=1}^n f(\mathbf{g}(u_i,v_i)) \left|\det(\mathbf{g}'(u_i,v_i))\right| A(D'_i) \\ &\approx \iint_{D'} f(g(u,v),h(u,v)) \left|\frac{\partial(g,h)}{\partial(u,v)}\right| du\,dv\,.\end{aligned}$$

When $\operatorname{diam}(D_i) \to 0$ the approximation errors also converge to 0 and we get the claimed identity. For a stringent proof of Theorem 20.5 the interested reader is referred to literature on advanced calculus.

Example 20.6

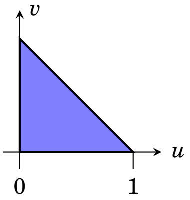
Let $D = \{(x,y)\colon -1 \le x \le 1, |x| \le y \le 1\}$ and $f(x,y) = x^2+y^2$ be defined on D. Compute $\iint_D f(x,y)\,dx\,dy$.



SOLUTION. We directly can compute this integral as

$$\begin{aligned}\iint_D f(x,y)\,dy\,dx &= \int_{-1}^1 \int_{|x|}^1 x^2+y^2\,dy\,dx \\ &= \int_{-1}^0 \int_{-x}^1 x^2+y^2\,dy\,dx + \int_0^1 \int_x^1 x^2+y^2\,dy\,dx \\ &= \int_{-1}^0 \left(x^2y+\frac{1}{3}y^3\right)\Big|_{y=-x}^1 dx + \int_0^1 \left(x^2y+\frac{1}{3}y^3\right)\Big|_{y=x}^1 dx \\ &= \int_{-1}^0 x^2+\frac{1}{3}+x^3+\frac{1}{3}x^3\,dx + \int_0^1 x^2+\frac{1}{3}-x^3-\frac{1}{3}x^3\,dx \\ &= \left(\frac{1}{3}x^3+\frac{1}{3}x+\frac{1}{3}x^4\right)\Big|_{-1}^0 + \left(\frac{1}{3}x^3+\frac{1}{3}x-\frac{1}{3}x^4\right)\Big|_0^1 \\ &= \frac{1}{3}+\frac{1}{3}-\frac{1}{3}+\frac{1}{3}+\frac{1}{3}-\frac{1}{3} = \frac{2}{3}\,.\end{aligned}$$

We also can first change variables. Let



$$\mathbf{g}(u,v) = \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix} \cdot \begin{pmatrix} u \\ v \end{pmatrix}$$

and $D' = \{(u,v)\colon 0 \le u \le 1, 0 \le v \le 1-u\}$. Then $\mathbf{g}(D') = D$ and

$$\left|\mathbf{g}'(u,v)\right| = \begin{vmatrix} 1 & -1 \\ 1 & 1 \end{vmatrix} = 2$$

which is constant and thus bounded and strictly positive. Thus we find

$$\begin{aligned}
\iint_D f(x,y)\,dy\,dx &= \iint_{D'} f(\mathbf{g}(u,v))\,|\mathbf{g}'(u,v)|\,dv\,du \\
&= \int_0^1 \int_0^{1-u} \left((u-v)^2+(u+v)^2\right)2\,dv\,du \\
&= 4\int_0^1 \int_0^{1-u} \left(u^2+v^2\right)\,dv\,du \\
&= 4\int_0^1 \left(u^2v+\frac{1}{3}v^3\right)\Big|_{v=0}^{1-u}\,du \\
&= 4\int_0^1 \left(-\frac{4}{3}u^3+2u^2-u+\frac{1}{3}\right)du \\
&= 4\left(-\frac{1}{3}u^4+\frac{2}{3}u^3-\frac{1}{2}u^2+\frac{1}{3}u\right)\Big|_0^1 \\
&= 4\left(-\frac{1}{3}+\frac{2}{3}-\frac{1}{2}+\frac{1}{3}\right)=\frac{2}{3}
\end{aligned}$$

which gives (of course) the same result. ♢

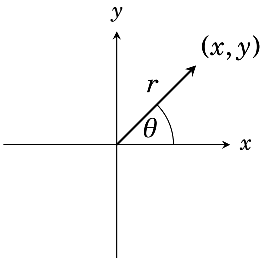
Theorem 20.7

Change of variables in multiple integrals. Let $f(\mathbf{x})$ be a function defined on an open bounded domain $D \subset \mathbb{R}^n$. Suppose that $\mathbf{x}=\mathbf{g}(\mathbf{z})$ defines a one-to-one $\mathscr{C}^1$ transformation from an open bounded set $D' \subset \mathbb{R}^n$ onto D such that the Jacobian determinant $\frac{\partial(g_1,\dots,g_h)}{\partial(z_1,\dots,z_n)}$ is bounded and either strictly positive or strictly negative on D'. Then

$$\iint_D f(\mathbf{x})d\mathbf{x} = \iint_{D'} f(\mathbf{g}(\mathbf{z}))\left|\frac{\partial(g_1,\dots,g_n)}{\partial(z_1,\dots,z_n)}\right| d\mathbf{z}\,.$$

We also may state this rule analogously to the rule for integration by substitution (Table 19.17)

$$\iint_D f(\mathbf{x})\,d\mathbf{x} = \iint_{D'} f(\mathbf{g}(\mathbf{z}))\left|\det(\mathbf{g}'(\mathbf{z}))\right|\,d\mathbf{z}\,.$$



Polar coordinates are very convenient when we have to deal with circular functions. Thus we represent a point by its distant r from the origin and the angle enclosed by the corresponding vector and the positive x-axis. The corresponding transformation is given by

$$\begin{pmatrix} x \\ y \end{pmatrix} = \mathbf{g}(r,\theta) = \begin{pmatrix} r\cos(\theta) \\ r\sin(\theta) \end{pmatrix}$$

where $(r,\theta) \in [0,\infty)\times[0,2\pi)$. It is a $\mathscr{C}^1$ function and its Jacobian determinant is given by

$$\left|\mathbf{g}'(r,\theta)\right| = \begin{vmatrix} \cos(\theta) & -r\sin(\theta) \\ \sin(\theta) & r\cos(\theta) \end{vmatrix} = r\left(\cos^2(\theta)+\sin^2(\theta)\right) = r$$

which is bounded on every bounded domain and it is strictly positive except for the null set $\{(0,\theta)\colon 0 \le \theta < 2\pi\}$.

Let $f(x,y) = 1 - x^2 - y^2$ be a function defined on $D = \{(x,y)\colon x^2 + y^2 \le 1\}$. Compute $\iint_D f(x,y)\,dx\,dy$. Example 20.8

SOLUTION. A direct computation of this integral is cumbersome:

$$\iint_D (1 - x^2 - y^2)\,dx\,dy = \int_0^1 \int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} (1 - x^2 - y^2)\,dy\,dx\,.$$

Thus we change to polar coordinates. Then $D' = \{(r,\theta)\colon 0 \le r \le 1,\ 0 \le \theta < 2\pi\}$ and we find

$$\begin{aligned}\iint_D (1 - x^2 - y^2)\,dx\,dy &= \int_0^1 \int_0^{2\pi} (1 - r^2) r\,d\theta\,dr = 2\pi \int_0^1 (r - r^3)\,dr \\ &= 2\pi \left(\frac{1}{2} r^2 - \frac{1}{4} r^4\right)\Bigg|_0^1 = \frac{\pi}{2}\,. \qquad \diamondsuit\end{aligned}$$

20.6 Improper Multiple Integrals

In Section 19.5 we have extended the concept of integral to unbounded functions or functions with unbounded domains. Using Fubini's theorem the definition of such improper integrals is straight forward by means of limits.

Compute $\displaystyle\int_0^\infty \int_0^\infty e^{-x^2 - y^2}\,dx\,dy$. Example 20.9

SOLUTION. We switch to polar coordinates. $f(x,y) = e^{-x^2-y^2}$ is defined on $D = \{(x,y)\colon x \ge 0,\ y \ge 0\}$. Then $D' = \{(r,\theta)\colon r \ge 0,\ 0 \le \theta < \pi/2\}$ and we find

$$\begin{aligned}\int_0^\infty \int_0^\infty e^{-x^2-y^2}\,dx\,dy &= \int_0^\infty \int_0^{\pi/2} e^{-r^2} r\,d\theta\,dr = \frac{\pi}{2} \int_0^\infty e^{-r^2} r\,dr \\ &= \lim_{t\to\infty} \frac{\pi}{2} \int_0^t e^{-r^2} r\,dr = \lim_{t\to\infty} \left(-\frac{\pi}{4} e^{-r^2}\right)\Bigg|_0^t \\ &= \lim_{t\to\infty} \left(-\frac{\pi}{4}\left(e^{-t^2} - 1\right)\right) = \frac{\pi}{4}\,. \qquad \diamondsuit\end{aligned}$$

Exercises

20.1 Evaluate the following double integrals

(a) $\displaystyle\int_0^2 \int_0^1 (2x+3y+4)\,dx\,dy$ (b) $\displaystyle\int_0^a \int_0^b (x-a)(y-b)\,dx\,dy$

(c) $\displaystyle\int_1^3 \int_1^2 \frac{x-y}{x+y}\,dx\,dy$ (d) $\displaystyle\int_0^{1/2} \int_0^{2\pi} y^3 \sin(xy^2)\,dx\,dy$

20.2 Compute

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-x^2-y^2}\,dx\,dy\,.$$

Problems

20.3 Prove the formula from Section 20.4:

$$\int_a^b \int_c^d f(x,y)\,dy\,dx = F(b,d)-F(a,d)-F(b,c)+F(a,c)\,.$$

HINT: $\int f(x,y)\,dy = \frac{\partial F(x,y)}{\partial x}$

where

$$\frac{\partial^2 F(x,y)}{\partial x \partial y} = f(x,y) \quad \text{for all } (x,y) \in [a,b]\times[c,d].$$

20.4 Let $\Phi(x)$ denote the cumulative distribution function of the (univariate) standard normal distribution. Let

$$f(x,y) = \frac{\sqrt{6}}{\pi}\exp(-2x^2-3y^2)$$

be the probability density function of a bivariate normal distribution.

HINT: Show that $\iint_{\mathbb{R}^2} f(x,y)\,dx\,dy = 1$.

(a) Show that $f(x,y)$ is indeed a probability function.

(b) Compute the cumulative distribution function and express the results by means of Φ.
HINT: $F(x,y) = \int_{-\infty}^{x}\int_{-\infty}^{y} \frac{\sqrt{6}}{\pi}\exp(-2s^2-3t^2)\,ds\,dt = \int_{-\infty}^{x}\int_{-\infty}^{y} \frac{\sqrt{2}}{\sqrt{\pi}}\exp(-2s^2)\cdot \frac{\sqrt{3}}{\sqrt{\pi}}\exp(-3t^2)\,ds\,dt$.

20.5 Compute

$$\iint_{\mathbb{R}^2} \exp(-q(x,y))\,dx\,dy$$

where

$$q(x,y) = 2x^2 - 2xy + 2y^2$$

HINT: Observe, that q is a quadratic form with matrix $\mathbf{A} = \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}$. So change the variables with respect to eigenvectors of $\mathbf{A}$.

Solutions

4.1 (a) $\mathbf{A}+\mathbf{B} = \begin{pmatrix} 2 & -2 & 8 \\ 10 & 1 & -1 \end{pmatrix}$; (b) not possible since the number of columns of $\mathbf{A}$ does not coincide with the number of rows of $\mathbf{B}$; (c) $3\mathbf{A}' = \begin{pmatrix} 3 & 6 \\ -18 & 3 \\ 15 & -9 \end{pmatrix}$; (d) $\mathbf{A}\cdot\mathbf{B}' = \begin{pmatrix} -8 & 18 \\ -3 & 10 \end{pmatrix}$; (e) $\mathbf{B}'\cdot\mathbf{A} = \begin{pmatrix} 17 & 2 & -19 \\ 4 & -24 & 20 \\ 7 & -16 & 9 \end{pmatrix}$; (f) not possible; (g) $\mathbf{C}\cdot\mathbf{A}+\mathbf{C}\cdot\mathbf{B} = \mathbf{C}\cdot(\mathbf{A}+\mathbf{B}) = \begin{pmatrix} -8 & -3 & 9 \\ 22 & 0 & 6 \end{pmatrix}$; (h) $\mathbf{C}^2 = \mathbf{C}\cdot\mathbf{C} = \begin{pmatrix} 0 & -3 \\ 3 & 3 \end{pmatrix}$.

4.2 $\mathbf{A}\cdot\mathbf{B} = \begin{pmatrix} 4 & 2 \\ 1 & 2 \end{pmatrix} \neq \mathbf{B}\cdot\mathbf{A} = \begin{pmatrix} 5 & 1 \\ -1 & 1 \end{pmatrix}$.

4.3 $\mathbf{x}'\mathbf{x} = 21$, $\mathbf{x}\mathbf{x}' = \begin{pmatrix} 1 & -2 & 4 \\ -2 & 4 & -8 \\ 4 & -8 & 16 \end{pmatrix}$, $\mathbf{x}'\mathbf{y} = -1$, $\mathbf{y}'\mathbf{x} = -1$,

$\mathbf{x}\mathbf{y}' = \begin{pmatrix} -3 & -1 & 0 \\ 6 & 2 & 0 \\ -12 & -4 & 0 \end{pmatrix}$, $\mathbf{y}\mathbf{x}' = \begin{pmatrix} -3 & 6 & -12 \\ -1 & 2 & -4 \\ 0 & 0 & 0 \end{pmatrix}$.

4.4 $\mathbf{B}$ must be a 2 × 4 matrix. $\mathbf{A}\cdot\mathbf{B}\cdot\mathbf{C}$ is then a 3 × 3 matrix.

4.5 (a) $\mathbf{X} = (\mathbf{A}+\mathbf{B}-\mathbf{C})^{-1}$; (b) $\mathbf{X} = \mathbf{A}^{-1}\mathbf{C}$; (c) $\mathbf{X} = \mathbf{A}^{-1}\mathbf{B}\mathbf{A}$; (d) $\mathbf{X} = \mathbf{C}\mathbf{B}^{-1}\mathbf{A}^{-1} = \mathbf{C}(\mathbf{A}\mathbf{B})^{-1}$.

4.6 (a) $\mathbf{A}^{-1} = \begin{pmatrix} 1 & 0 & 0 & -\frac{1}{4} \\ 0 & 1 & 0 & -\frac{2}{4} \\ 0 & 0 & 1 & -\frac{3}{4} \\ 0 & 0 & 0 & \frac{1}{4} \end{pmatrix}$; (b) $\mathbf{B}^{-1} = \begin{pmatrix} 1 & 0 & -\frac{5}{3} & -\frac{3}{2} \\ 0 & \frac{1}{2} & 0 & -\frac{7}{8} \\ 0 & 0 & \frac{1}{3} & 0 \\ 0 & 0 & 0 & \frac{1}{4} \end{pmatrix}$.

5.1 For example: (a) $2\mathbf{x}_1 + 0\mathbf{x}_2 = \begin{pmatrix} 2 \\ 4 \end{pmatrix}$; (b) $3\mathbf{x}_1 - 2\mathbf{x}_2 = \begin{pmatrix} 4 \\ -2 \\ 3 \end{pmatrix}$.

6.1 (a) $\ker(\phi) = \operatorname{span}(\{1\})$; (b) $\operatorname{Im}(\phi) = \operatorname{span}(\{1, x\})$; (c) $\mathbf{D} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 2 \\ 0 & 0 & 0 \end{pmatrix}$;

(d) $\mathbf{U}_\ell^{-1} = \begin{pmatrix} 1 & 1 & 1 \\ 0 & -1 & -2 \\ 0 & 0 & \frac{1}{2} \end{pmatrix}$, $\mathbf{U}_\ell = \begin{pmatrix} 1 & 1 & 2 \\ 0 & -1 & -4 \\ 0 & 0 & 2 \end{pmatrix}$;

(e) $\mathbf{D}_\ell = \mathbf{U}_\ell \mathbf{D} \mathbf{U}_\ell^{-1} = \begin{pmatrix} 0 & -1 & -1 \\ 0 & 0 & -1 \\ 0 & 0 & 0 \end{pmatrix}$.

7.1 Row reduced echelon form $\mathbf{R} = \begin{pmatrix} 1 & 0 & -1 \\ 0 & 1 & 2 \\ 0 & 0 & 0 \end{pmatrix}$. $\operatorname{Im}(\mathbf{A}) = \operatorname{span}\{(1,4,7)',(2,5,8)'\}$, $\ker(\mathbf{A}) = \operatorname{span}\{(1,-2,1)'\}$, $\operatorname{rank}(\mathbf{A}) = 2$.

10.1 (a) -3; (b) -9; (c) 8; (d) 0; (e) -40; (f) -10; (g) 48; (h) -49; (i) 0.

10.2 See Exercise 10.1.

10.3 All matrices except those in Exercise 10.1(d) and (i) are regular and thus invertible and have linear independent column vectors.

Ranks of the matrices: (a)–(d) rank 2; (e)–(f) rank 3; (g)–(h) rank 4; (i) rank 1.

10.4 (a) $\det(\mathbf{A}) = 3$; (b) $\det(5\mathbf{A}) = 5^3 \det(\mathbf{A}) = 375$; (c) $\det(\mathbf{B}) = 2\det(\mathbf{A}) = 6$; (d) $\det(\mathbf{A}') = \det(\mathbf{A}) = 3$; (e) $\det(\mathbf{C}) = \det(\mathbf{A}) = 3$; (f) $\det(\mathbf{A}^{-1}) = \frac{1}{\det(\mathbf{A})} = \frac{1}{3}$; (g) $\det(\mathbf{A}\cdot\mathbf{C}) = \det(\mathbf{A})\cdot\det(\mathbf{C}) = 3\cdot 3 = 9$; (h) $\det(\mathbf{I}) = 1$.

10.5 $|\mathbf{A}'\cdot\mathbf{A}| = 0$; $|\mathbf{A}\cdot\mathbf{A}'|$ depends on matrix $\mathbf{A}$.

10.6 (a) 9; (b) 9; (c) 40; (e) 40.

10.7 $\mathbf{A}^{-1} = \frac{1}{|\mathbf{A}|}\mathbf{A}^{*\prime}$.

(a) $\mathbf{A}^* = \begin{pmatrix} 1 & -2 \\ -2 & 1 \end{pmatrix}$, $\mathbf{A}^{*\prime} = \begin{pmatrix} 1 & -2 \\ -2 & 1 \end{pmatrix}$, $|\mathbf{A}| = -3$;

(b) $\mathbf{A}^* = \begin{pmatrix} 3 & -1 \\ -3 & -2 \end{pmatrix}$, $\mathbf{A}^{*\prime} = \begin{pmatrix} 3 & -3 \\ -1 & -2 \end{pmatrix}$, $|\mathbf{A}| = -9$;

(c) $\mathbf{A}^* = \begin{pmatrix} 2 & 0 \\ 3 & 4 \end{pmatrix}$, $\mathbf{A}^{*\prime} = \begin{pmatrix} 2 & 3 \\ 0 & 4 \end{pmatrix}$, $|\mathbf{A}| = 8$;

(d) $\mathbf{A}^* = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 3 & -6 \\ -1 & 0 & 3 \end{pmatrix}$, $\mathbf{A}^{*\prime} = \begin{pmatrix} 1 & 1 & -1 \\ 0 & 3 & 0 \\ 0 & -6 & 3 \end{pmatrix}$, $|\mathbf{A}| = 3$;

(e) $\mathbf{A}^{*\prime} = \begin{pmatrix} -20 & -12 & 8 \\ 20 & 4 & -16 \\ 5 & -5 & 0 \end{pmatrix}$, $|\mathbf{A}| = -40$;

(f) $\mathbf{A}^{*\prime} = \begin{pmatrix} 9 & 3 & -4 \\ -2 & -4 & 2 \\ -14 & -8 & 4 \end{pmatrix}$, $|\mathbf{A}| = -10$.

10.8 (a) $\mathbf{A}^{-1} = \frac{1}{ad-bc}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$; (b) $\mathbf{A}^{-1} = \frac{1}{x_1y_2-x_2y_1}\begin{pmatrix} y_2 & -y_1 \\ -x_2 & x_1 \end{pmatrix}$;

(c) $\mathbf{A}^{-1} = \frac{1}{\alpha\beta^2-\alpha^2\beta}\begin{pmatrix} \beta^2 & -\beta \\ -\alpha^2 & \alpha \end{pmatrix}$.

10.9 (a) $\mathbf{x} = (1,0)'$; (b) $\mathbf{x} = (1/3, 5/9)'$; (c) $\mathbf{x} = (1,1)'$; (d) $\mathbf{x} = (0,2,-1)'$; (e) $\mathbf{x} = (1/2, 1/2, 1/8)'$; (f) $\mathbf{x} = (-3/10, 2/5, 9/5)'$.

11.1 (a) $\lambda_1 = 7$, $\mathbf{v}_1 = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$; $\lambda_2 = 2$, $\mathbf{v}_2 = \begin{pmatrix} -2 \\ 1 \end{pmatrix}$; (b) $\lambda_1 = 14$, $\mathbf{v}_1 = \begin{pmatrix} 1 \\ 4 \end{pmatrix}$; $\lambda_2 = 1$, $\mathbf{v}_2 = \begin{pmatrix} -3 \\ 1 \end{pmatrix}$; (c) $\lambda_1 = -6$, $\mathbf{v}_1 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$; $\lambda_2 = 4$, $\mathbf{v}_2 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$.

11.2 (a) $\lambda_1 = 0$, $\mathbf{x}_1 = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}$; $\lambda_2 = 2$, $\mathbf{x}_2 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$; $\lambda_3 = 2$, $\mathbf{x}_3 = \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix}$.

(b) $\lambda_1 = 1$, $\mathbf{x}_1 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$; $\lambda_2 = 2$, $\mathbf{x}_2 = \begin{pmatrix} -1 \\ 2 \\ 2 \end{pmatrix}$; $\lambda_3 = 3$, $\mathbf{x}_3 = \begin{pmatrix} -1 \\ 1 \\ 1 \end{pmatrix}$.

(c) $\lambda_1 = 1$, $\mathbf{x}_1 = \begin{pmatrix} 2 \\ -1 \\ 1 \end{pmatrix}$; $\lambda_2 = 3$, $\mathbf{x}_2 = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}$; $\lambda_3 = 3$, $\mathbf{x}_3 = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}$.

(d) $\lambda_1 = -3$, $\mathbf{x}_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$; $\lambda_2 = -5$, $\mathbf{x}_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$; $\lambda_3 = -9$, $\mathbf{x}_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$.

(e) $\lambda_1 = 0$, $\mathbf{x}_1 = \begin{pmatrix} 1 \\ 0 \\ -3 \end{pmatrix}$; $\lambda_2 = 1$, $\mathbf{x}_2 = \begin{pmatrix} 2 \\ -3 \\ -1 \end{pmatrix}$; $\lambda_3 = 4$, $\mathbf{x}_3 = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}$.

(f) $\lambda_1 = 0$, $\mathbf{x}_1 = \begin{pmatrix} -2 \\ 2 \\ 1 \end{pmatrix}$; $\lambda_2 = 27$, $\mathbf{x}_2 = \begin{pmatrix} 2 \\ 1 \\ 2 \end{pmatrix}$; $\lambda_3 = -9$, $\mathbf{x}_3 = \begin{pmatrix} -1 \\ -2 \\ 2 \end{pmatrix}$.

11.3 (a) $\lambda_1 = \lambda_2 = \lambda_3 = 1$, $\mathbf{x}_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$, $\mathbf{x}_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$, $\mathbf{x}_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$; (b) $\lambda_1 = \lambda_2 = \lambda_3 = 1$, $\mathbf{x}_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$.

11.4 11.1a: positiv definit, 11.1c: indefinit, 11.2a: positiv semidefinit, 11.2d: negativ definit, 11.2f: indefinit, 11.3a: positiv definit.
The other matrices are not symmetric. So our criteria cannot be applied.

11.5 $q_{\mathbf{A}}(\mathbf{x}) = 3x_1^2 + 4x_1x_2 + 2x_1x_3 - 2x_2^2 - x_3^2$.

11.6 $\mathbf{A} = \begin{pmatrix} 5 & 3 & -1 \\ 3 & 1 & -2 \\ -1 & -2 & 1 \end{pmatrix}$.

11.7 Eigenspace corresponding to eigenvalue $\lambda_1 = 0$: span $\left\{ \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} \right\}$;

Eigenspace corresponding to eigenvalues $\lambda_2 = \lambda_3 = 2$: span $\left\{ \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix} \right\}$.

11.8 Give examples.

11.9 11.1a: $H_1 = 3$, $H_2 = 14$, positive definite; 11.1c: $H_1 = -1$, $H_2 = -24$, indefinite; 11.2a: $H_1 = 1$, $H_2 = 0$, $H_3 = 0$, cannot be applied; 11.2d: $H_1 = -3$, $H_2 = 15$, $H_3 = -135$, negative definite; 11.2f: $H_1 = 11$, $H_2 = -27$, $H_3 = 0$, cannot be applied; 11.3a: $H_1 = 1$, $H_2 = 1$, $H_3 = 1$, positive definite.
All other matrices are not symmetric.

11.10 11.1a: $M_1 = 3$, $M_2 = 6$, $M_{1,2} = 14$, positive definite; 11.1c: $M_1 = -1$, $M_2 = -1$, $M_{1,2} = -24$, indefinite; 11.2a: $M_1 = 1$, $M_2 = 1$, $M_3 = 2$, $M_{1,2} = 0$, $M_{1,3} = 2$, $M_{2,3} = 2$, $M_{1,2,3} = 0$, positive semidefinite. 11.2d: $M_1 = -3$, $M_2 = -5$, $M_3 = -9$, $M_{1,2} = 15$, $M_{1,3} = 27$, $M_{2,3} = 45$, $M_{1,2,3} = -135$, negative definite. 11.2f: $M_1 = 11$, $M_2 = -1$, $M_3 = 8$, $M_{1,2} = -27$, $M_{1,3} = -108$, $M_{2,3} = -108$, $M_{1,2,3} = 0$, indefinite.

11.11

$$\mathbf{A} = \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}.$$

11.12

$$\sqrt{\mathbf{A}} = \begin{pmatrix} \frac{1+\sqrt{3}}{2} & \frac{1-\sqrt{3}}{2} \\ \frac{1-\sqrt{3}}{2} & \frac{1+\sqrt{3}}{2} \end{pmatrix}.$$

11.13 Matrix **A** has eigenvalues $\lambda_1 = 2$ and $\lambda_2 = -4$ with corresponding eigenvectors $\mathbf{v}_1 = (1,1)'$ and $\mathbf{v}_2 = (-1,1)'$. Then

$$e^{\mathbf{A}} = \begin{pmatrix} \frac{e^2+e^{-4}}{2} & \frac{e^2-e^{-4}}{2} \\ \frac{e^2-e^{-4}}{2} & \frac{e^2+e^{-4}}{2} \end{pmatrix}.$$

12.1 (a) 7; (b) $\frac{2}{7}$; (c) 0; (d) divergent with $\lim_{n\to\infty} \frac{n^2+1}{n+1} = \infty$; (e) divergent; (f) $\frac{29}{6}$.

12.2 (a) divergent; (b) 0; (c) $e^2 \approx 7{,}38906$; (d) $e^{-2} \approx 0.135335$; (e) 0; (f) 1; (g) divergent with $\lim_{n\to\infty} \frac{n}{n+1} + \sqrt{n} = \infty$; (h) 0.

12.3 (a) e^x; (b) e^x; (c) $e^{1/x}$.

12.11 By Lemma 12.20 we find $\sum_{k=1}^{\infty} q^n = q\sum_{k=0}^{\infty} q^n = \frac{q}{1-q}$.

14.1 (a) 0, (b) 0, (c) ∞, (d) $-\infty$, (e) 1.

14.2 The functions are continuous in
(a) D, (b) D, (c) D, (d) D, (e) D, (f) $\mathbb{R}\setminus\mathbb{Z}$, (g) $\mathbb{R}\setminus\{2\}$.

14.3 (a) $6x - 5\sin(x)$; (b) $6x^2 + 2x$; (c) $1 + \ln(x)$; (d) $-2x^{-2} - 2x^{-3}$; (e) $\frac{3x^2+6x+1}{(x+1)^2}$; (f) 1; (g) $18x-6$; (h) $6x\cos(3x^2)$; (i) $\ln(2)\cdot 2^x$; (j) $4x-1$; (k) $6e^{3x+1}(5x^2+1)^2 + 40e^{3x+1}(5x^2+1)x + \frac{3(x-1)(x+1)^2-(x+1)^3}{(x-1)^2} - 2$.

14.4

	$f'(x)$	$f''(x)$	$f'''(x)$
(a)	$-x e^{-\frac{x^2}{2}}$	$(x^2-1)e^{-\frac{x^2}{2}}$	$(3x-x^3)e^{-\frac{x^2}{2}}$
(b)	$\frac{-2}{(x-1)^2}$	$\frac{4}{(x-1)^3}$	$\frac{-12}{(x-1)^4}$
(c)	$3x^2-4x+3$	$6x-4$	6

14.5 Derivatives:

	(a)	(b)	(c)	(d)	(e)	(f)
f_x	1	y	$2x$	$2xy^2$	$\alpha x^{\alpha-1} y^{\beta}$	$x(x^2+y^2)^{-1/2}$
f_y	1	x	$2y$	$2x^2y$	$\beta x^{\alpha} y^{\beta-1}$	$y(x^2+y^2)^{-1/2}$
f_{xx}	0	0	2	$2y^2$	$\alpha(\alpha-1)x^{\alpha-2} y^{\beta}$	$(x^2+y^2)^{-1/2} - x^2(x^2+y^2)^{-3/2}$
$f_{xy} = f_{yx}$	0	1	0	$4xy$	$\alpha\beta x^{\alpha-1} y^{\beta-1}$	$-xy(x^2+y^2)^{-3/2}$
f_{yy}	0	0	2	$2x^2$	$\beta(\beta-1)x^{\alpha} y^{\beta-2}$	$(x^2+y^2)^{-1/2} - y^2(x^2+y^2)^{-3/2}$

Derivatives at $(1,1)$:

	(a)	(b)	(c)	(d)	(e)	(f)
f_x	1	1	2	2	α	$\sqrt{2}/2$
f_y	1	1	2	2	β	$\sqrt{2}/2$
f_{xx}	0	0	2	2	$\alpha(\alpha-1)$	$\sqrt{2}/4$
$f_{xy} = f_{yx}$	0	1	0	4	$\alpha\beta$	$-\sqrt{2}/4$
f_{yy}	0	0	2	2	$\beta(\beta-1)$	$\sqrt{2}/4$

14.6 (a) $f'(1,1) = (1,1)$, $f''(1,1) = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$;

(b) $f'(1,1) = (1,1)$, $f''(1,1) = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$;

(c) $f'(1,1)=(2,2)$, $f''(1,1)=\begin{pmatrix}2 & 0\\ 0 & 2\end{pmatrix}$;

(d) $f'(1,1)=(2,2)$, $f''(1,1)=\begin{pmatrix}2 & 4\\ 4 & 2\end{pmatrix}$;

(e) $f'(1,1)=(\alpha,\beta)$, $f''(1,1)=\begin{pmatrix}\alpha(\alpha-1) & \alpha\beta\\ \alpha\beta & \beta(\beta-1)\end{pmatrix}$;

(f) $f'(1,1)=(\sqrt{2}/2,\sqrt{2}/2)$, $f''(1,1)=\begin{pmatrix}\sqrt{2}/4 & -\sqrt{2}/4\\ -\sqrt{2}/4 & \sqrt{2}/4\end{pmatrix}$.

14.7 $\frac{\partial f}{\partial \mathbf{a}}=2\mathbf{x}'\cdot\mathbf{a}$.

14.8 $\nabla f(0,0)=(4/\sqrt{10},12/\sqrt{10})$.

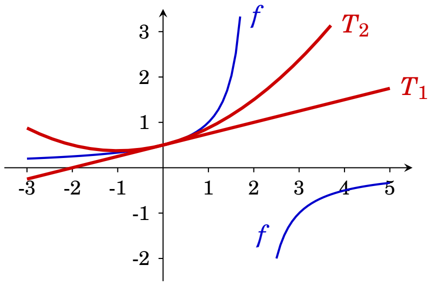
14.9 $D(f\circ g)(t)=2t+4t^3$; $D(g\circ f)(x,y)=\begin{pmatrix}2x & 2y\\ 4x^3+4xy^2 & 4x^2y+3y^3\end{pmatrix}$.

14.10 $D(\mathbf{f}\circ\mathbf{g})(\mathbf{x})=\begin{pmatrix}-1 & 6x_2^5\\ -3x_1^2 & 2x_2\end{pmatrix}$; $D(\mathbf{g}\circ\mathbf{f})(\mathbf{x})=\begin{pmatrix}2(x_1-x_2^3) & 6(-x_1x_2^2+x_2^5)\\ 3x_2^2 & -1\end{pmatrix}$.

14.11 $\frac{\partial x_i}{\partial b_j}=(-1)^{i+j}M_{ji}/|\mathbf{A}|$ and thus $D\mathbf{x}(\mathbf{b})=\mathbf{A}^{-1}$.

14.12 $\frac{d}{dt}F(K(t),L(t),t)=F_K(K,L,t)K'(t)+F_L(K,L,t)L'(t)+F_t(K,L,t)$.

15.1



(a) $f(x)\approx T_1(x)=\frac{1}{2}+\frac{1}{4}x$,

(b) $f(x)\approx T_2(x)=\frac{1}{2}+\frac{1}{4}x+\frac{1}{8}x^2$.

radius of convergence $\rho=2$.

15.2 $T_{f,0,3}(x)=1+\frac{1}{2}x-\frac{1}{8}x^2+\frac{1}{16}x^3$.

15.3 $T_{f,0,30}(x)=x^{10}-\frac{1}{6}x^{30}$.

15.4 $T_{f,0,4}(x)\approx 0.959+0.284x^2-0.479x^4$.

15.5 $f(x)=\sum_{n=0}^{\infty}(-1)^n x^{2n}$; $\rho=1$.

15.6 $f(x)=\sum_{n=0}^{\infty}(-\frac{1}{2})^n\frac{1}{n!}x^{2n}$; $\rho=\infty$.

15.7 $f(x,y)=1+x^2+y^2+O(\|(x,y)\|^3)$.

15.8 $|R_n(1)|\le\frac{e}{(n+1)!}$; $|R_n(1)|<10^{-16}$ if $n\ge 18$.

16.1 (a) $D\mathbf{f}(\mathbf{x})=\begin{pmatrix}1-x_2 & -x_1\\ x_2 & x_1\end{pmatrix}$, $\frac{\partial(y_1,y_2)}{\partial(x_1,x_2)}=x_1$;

(b) for all images of points (x_1,x_2) with $x_1\neq 0$;

(c) $D(\mathbf{f}^{-1})(\mathbf{y})=(D\mathbf{f}(\mathbf{x}))^{-1}=\begin{pmatrix}1-x_2 & -x_1\\ x_2 & x_1\end{pmatrix}^{-1}=\frac{1}{x_1}\begin{pmatrix}x_1 & x_1\\ -x_2 & 1-x_2\end{pmatrix}$;

(d) in order to get the inverse function we have to solve equation $\mathbf{f}(\mathbf{x})=\mathbf{y}$: $x_1=y_1+y_2$ and $x_2=y_2/(y_1+y_2)$, if $y_1+y_2\neq 0$.

16.2 T is the linear map given by the matrix $\mathbf{T} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$. Hence its Jacobian matrix is just det($\mathbf{T}$). If det($\mathbf{T}$) = 0, then the columns of $\mathbf{T}$ are linearly dependent. Since the constants are non-zero, $\mathbf{T}$ has rank 1 and thus the image is a linear subspace of dimension 1, i.e., a straight line through the origin.

16.3 Let $J = \begin{vmatrix} \frac{\partial f}{\partial x} & \frac{\partial f}{\partial y} \\ \frac{\partial g}{\partial x} & \frac{\partial g}{\partial y} \end{vmatrix}$ be the Jacobian determinant of this function. Then the equation can be solved locally if $J \neq 0$. We then have $\frac{\partial F}{\partial u} = \frac{1}{J}\frac{\partial g}{\partial y}$ and $\frac{\partial G}{\partial u} = -\frac{1}{J}\frac{\partial g}{\partial x}$.

16.4 (a) $F_y = 3y^2 + 1 \neq 0$, $y' = -F_x/F_y = 3x^2/(3y^2+1) = 0$ for $x = 0$;
(b) $F_y = 1 + x\cos(xy) = 1 \neq 0$ for $x = 0$, $y'(0) = 0$.

16.5 $\frac{dy}{dx} = -\frac{2x}{3y^2}$, $y = f(x)$ exists locally in an open rectangle around $\mathbf{x}_0 = (x_0, y_0)$ if $y_0 \neq 0$; $x = g(y)$ exists locally if $x_0 \neq 0$.

16.6 (a) $z = g(x,y)$ can be locally expressed since $F_z = 3z^2 - xy$ and $F_z(0,0,1) = 3 \neq 0$; $\frac{\partial g}{\partial x} = -\frac{F_x}{F_z} = -\frac{3x^2-yz}{3z^2-xy} = -\frac{0}{3} = 0$ for $(x_0, y_0, z_0) = (0,0,1)$; $\frac{\partial g}{\partial y} = -\frac{F_y}{F_z} = -\frac{3y^2-xz}{3z^2-xy} = -\frac{0}{3} = 0$.
(b) $z = g(x,y)$ can be locally expressed since $F_z = \exp(z) - 2z$ and $F_z(1,0,0) = 1 \neq 0$; $\frac{\partial g}{\partial x} = -\frac{F_x}{F_z} = -\frac{-2x}{\exp(z)-2z} = 2$ for $(x_0, y_0, z_0) = (1,0,0)$; $\frac{\partial g}{\partial y} = -\frac{F_y}{F_z} = -\frac{-2y}{\exp(z)-2z} = 0$ for $(x_0, y_0, z_0) = (1,0,0)$.

16.7 $\frac{dK}{dL} = -\frac{\beta K}{\alpha L}$.

16.8 (a) $\frac{dx_i}{dx_j} = -\frac{u_{x_j}}{u_{x_i}} = -\frac{\left(x_1^{\frac{1}{2}} + x_2^{\frac{1}{2}}\right) x_j^{-\frac{1}{2}}}{\left(x_1^{\frac{1}{2}} + x_2^{\frac{1}{2}}\right) x_i^{-\frac{1}{2}}} = -\frac{x_i^{\frac{1}{2}}}{x_j^{\frac{1}{2}}}$;

(b) $\frac{dx_i}{dx_j} = -\frac{u_{x_j}}{u_{x_i}} = -\frac{\frac{\theta}{\theta-1}\left(\sum_{i=1}^n x_i^{\frac{\theta-1}{\theta}}\right)^{\frac{1}{\theta-1}} \frac{\theta-1}{\theta} x_j^{-\frac{1}{\theta}}}{\frac{\theta}{\theta-1}\left(\sum_{i=1}^n x_i^{\frac{\theta-1}{\theta}}\right)^{\frac{1}{\theta-1}} \frac{\theta-1}{\theta} x_i^{-\frac{1}{\theta}}} = -\frac{x_i^{\frac{1}{\theta}}}{x_j^{\frac{1}{\theta}}}$.

17.1 (a) decreasing in $(-\infty, -4] \cup [0,3]$, increasing in $[-4,0] \cup [3,\infty)$; (b) concave in $[-2-\sqrt{148})/6, -2+\sqrt{148})/6]$, convex otherwise.

17.2 (a) log-concave; (b) not log-concave; (c) not log-concave; (d) log-concave on $(-1,1)$.

17.3 (a) concave; (b) concave.

18.1 (a) global minimum at $x = 3$ ($f''(x) \geq 0$ for all $x \in \mathbb{R}$), no local maximum;
(b) local minimum at $x = 1$, local maximum at $x = -1$, no global extrema.

18.2 (a) global minimum in $x = 1$, no local maximum;
(b) global maximum in $x = \frac{1}{4}$, no local minimum;
(c) global minimum in $x = 0$, no local maximum.

18.3 (a) stationary point: $\mathbf{p}_0 = (0,0)$, $\mathbf{H}_f = \begin{pmatrix} -2 & 1 \\ 1 & 2 \end{pmatrix}$,
$H_2 = -5 < 0$, $\Rightarrow$ $\mathbf{p}_0$ is a saddle point;
(b) stationary point: $\mathbf{p}_0 = (e,0)$, $\mathbf{H}_f(\mathbf{p}_0) = \begin{pmatrix} -e^{-3} & 0 \\ 0 & -2 \end{pmatrix}$,
$H_1 = -e^{-3} < 0$, $H_2 = 2e^{-3} > 0$, $\Rightarrow$ $\mathbf{p}_0$ is local maximum;

(c) stationary point: $\mathbf{p}_0 = (1,1)$, $\mathbf{H}_f(\mathbf{p}_0) = \begin{pmatrix} 802 & -400 \\ -400 & 200 \end{pmatrix}$,
$H_1 = 802 > 0$, $H_2 = 400 > 0$, $\Rightarrow$ $\mathbf{p}_0$ is local minimum;
(d) stationary point: $\mathbf{p}_0 = (\ln(3), \ln(4))$, $\mathbf{H}_f = \begin{pmatrix} -e^{x_1} & 0 \\ 0 & -e^{x_2} \end{pmatrix}$,
$H_1 = -e^{x_1} < 0$, $H_2 = e^{x_1} \cdot e^{x_2} > 0$, $\Rightarrow$ local maximum in $\mathbf{p}_0 = (\ln(3), \ln(4))$.

18.4 stationary points: $\mathbf{p}_1 = (0,0,0)$, $\mathbf{p}_2 = (1,0,0)$, $\mathbf{p}_3 = (-1,0,0)$,
$\mathbf{H}_f = \begin{pmatrix} 6x_1x_2 & 3x_1^2 - 1 & 0 \\ 3x_1^2 - 1 & 0 & \\ 0 & 0 & 2 \end{pmatrix}$,
leading principle minors: $H_1 = 6x_1x_2 = 0$, $H_2 = -(3x_1^2 - 1)^2 < 0$ (da $x_1 \in \{0, -1, 1\}$), $H_3 = -2(3x_1^2 - 1)^2 < 0$,
$\Rightarrow$ all three stationary points are saddle points. The function is neither convex nor concave.

18.5 (b) Lagrange function: $\mathscr{L}(x, y; \lambda) = x^2 y + \lambda(3 - x - y)$,
stationary points $\mathbf{x}_1 = (2, 1; 4)$ and $\mathbf{x}_2 = (0, 3; 0)$,
(c) bordered Hessian: $\bar{\mathbf{H}} = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 2y & 2x \\ 1 & 2x & 0 \end{pmatrix}$,
$\bar{\mathbf{H}}(\mathbf{x}_1) = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & 4 & 0 \end{pmatrix}$, $\det(\bar{\mathbf{H}}(\mathbf{x}_1)) = 6 > 0$, $\Rightarrow$ $\mathbf{x}_1$ is a local maximum,
$\bar{\mathbf{H}}(\mathbf{x}_2) = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 6 & 0 \\ 1 & 0 & 0 \end{pmatrix}$, $\det(\bar{\mathbf{H}}(\mathbf{x}_2)) = -6$ $\Rightarrow$ $\mathbf{x}_2$ is a local minimum.

18.6 Lagrange function: $\mathscr{L}(x_1, x_2, x_3; \lambda_1, \lambda_2) = f(x_1, x_2, x_3) = \frac{1}{3}(x_1 - 3)^3 + x_2 x_3 + \lambda_1(4 - x_1 - x_2) + \lambda_2(5 - x_1 - x_3)$,
stationary points: $\mathbf{x}_1 = (0, 4, 5; 5, 4)$ and $\mathbf{x}_2 = (4, 0, 1; 1, 0)$.

18.7 (a) $x_1 = \alpha \frac{m}{p_1}$, $x_2 = (1 - \alpha)\frac{m}{p_2}$ and $\lambda = \frac{1}{m}$, (c) marginal change for optimum: $\frac{1}{m}$.

18.8 Kuhn-Tucker theorem: $\mathscr{L}(x, y; \lambda) = -(x - 2)^2 - y + \lambda(1 - x - y)$, $x = 1$, $y = 0$, $\lambda = 2$.

19.1 (a) integration by parts (P): $\frac{1}{4}x^2(2\ln x - 1) + c$;
(b) 2×P: $2\cos(x) - x^2\cos(x) + 2x\sin(x) + c$;
(c) by substitution (S), $z = x^2 + 6$: $\frac{2}{3}\left(x^2 + 6\right)^{\frac{3}{2}} + c$;
(d) S, $z = x^2$: $\frac{1}{2}e^{x^2} + c$;
(e) S, $z = 3x^2 + 4$: $\frac{1}{6}\ln(4 + 3x^2) + c$;
(f) P or S, $z = x + 1$: $\frac{2}{5}(x + 1)^{\frac{5}{2}} - \frac{2}{3}(x + 1)^{\frac{3}{2}} + c$;
(g) $= \int 3x + \frac{4}{x}\,dx = \frac{3}{2}x^2 + 4\ln(x) + c$; S not suitable;
(h) S, $z = \ln(x)$: $\frac{1}{2}(\ln(x))^2 + c$.

19.2 (a) 39, (b) $3e^2 - 3 \approx 19.17$, (c) 93, (d) $-\frac{1}{6}$ (use radiant instead of degree), (e) $\frac{1}{2}\ln(8) \approx 1.0397$

19.3 (a) $\int_0^\infty -e^{-3x}\,dx = \lim_{t\to\infty}\int_0^t -e^{-3x}\,dx = \lim_{t\to\infty}\frac{1}{3}e^{-3t} - \frac{1}{3} = -\frac{1}{3}$;
(b) $\int_0^1 \frac{2}{\sqrt[4]{x^3}}\,dx = \lim_{t\to 0}\int_t^1 \frac{2}{\sqrt[4]{x^3}}\,dx = \lim_{t\to 0} 8 - 8t^{\frac{1}{4}} = 8$;
(c) $= \lim_{t\to\infty}\int_0^t \frac{x}{x^2+1}\,dx = \lim_{t\to\infty}\frac{1}{2}\int_2^{t^2+1}\frac{1}{z}\,dz = \lim_{t\to\infty}\frac{1}{2}(\ln(t^2 + 1) - \ln(2)) = \infty$,
the improper integral does not exist.

19.4 We need the antiderivative $C(x)$ of $C'(x) = 30 - 0.05x$ with $C(0) = 2000$: $C(x) = 2000 + 30x - 0{,}025x^2$.

19.5 $\mathbb{E}(X) = \sqrt{\frac{2}{\pi}}$.

19.6 $\mathbb{E}(X) = -\sqrt{\frac{2}{\pi}} + \sqrt{\frac{2}{\pi}} = 0$.

19.7 (a) The improper integral exists if and only if $\alpha > -1$;
(b) the improper integral exists if and only if $\alpha < -1$;
(c) the improper integral always converges.

20.1 (a) 16; (b) $\frac{a^2b^2}{4}$; (c) $-5\ln(5) + 8\ln(4) - 3\ln(3) \approx -0.2527$; (d) $\frac{\pi-2}{8\pi}$.

20.2 π.

Index

Entries in *italics* indicate lemmata and theorems.