Mathematics 1b

Functions of several variables

Ole Christensen · Jakob Lemvig

DTU Compute 2026

Preface

Welcome to the course *Mathematics 1b: Functions of Several Variables*! This textbook builds on the *Compendium for the course Mat 1a* by Peter Beelen [Mathematics 1a], used in the preceding course, *Mathematics 1a: Logic and Linear Algebra* and continues much of the theory, notation, and terminology introduced in Mathematics 1a, and you are encouraged to consult the Mathematics 1a book whenever needed to revisit or clarify foundational concepts.

This textbook is designed to guide you through the core ideas of multi-variable calculus, a key area of mathematics that helps us analyze and solve problems in multiple dimensions. Calculus of several variables extends the familiar concepts of differentiation and integration to multiple variables, enabling the analysis of more complex systems that are commonly encountered in the natural and engineering sciences.

Differentiation in several variables allows us to determine the rate of change in multiple directions simultaneously, a important concept in fields like biology, where we might study how a plant's growth depends on environmental factors such as sunlight and water. Integration, on the other hand, offers a means to aggregate effects over a region — whether it is calculating the volume of a structure, the total energy in a system, or the cumulative rainfall over an area.

These mathematical tools are indispensable for solving real-world problems, from designing aerodynamic structures in engineering, to optimizing machine learning models in artificial intelligence, and understanding gravitational or electromagnetic force fields in physics. By bridging abstract mathematical principles with practical applications, this course lays the foundation for tackling complex problems across engineering disciplines.

Acknowledgements

This book is the result of collaboration between the two authors and draw on a long tradition of high-quality teaching material at DTU. We are grateful for the support and inspiration provided by our colleagues at DTU Compute, particularly Peter Beelen and Ulrik Engelund Pedersen, whose work on developing Mathematics 1a and Mathematics 1b has been invaluable. Jens Gravesen's lecture notes [2], along with earlier teaching material, e.g., [3], used in mathematics courses at DTU, have also been a significant source of guidance and inspiration.

We would like to thank Nikoline Mai Bøgely Rehn for her assistance with the LaTeX layout and the Ti*k*Z code used to create the figures in this book. The LaTeX layout is adapted from the MatNat Compendium, developed using the memoir class at the Faculty of Mathematics and Natural Sciences, University of Oslo.

We owe special thanks to the many students and colleagues who carefully reviewed the first draft of these notes, catching typos and providing invaluable feedback. Your contributions have greatly improved the quality of this text.

Kgs. Lyngby, January 2025

Ole Christensen and Jakob Lemvig

New to the 2026 edition are additional examples and exercises on the softmax function and neural networks (see Definition 1.3.2 on page 22), and on the role of multivariate calculus in machine learning, reflecting the growing importance of these topics in engineering and data science.

Kgs. Lyngby, December 2025

Ole Christensen and Jakob Lemvig

Contents

Preface ii

Acknowledgements iii

Contents v

List of Figures viii

0 Preliminaries 1
- 0.1 Scalars . . . 1
- 0.2 Sets . . . 5
- 0.3 Functions $f : A \to B$ . . . 6
- 0.4 Matrices and vectors . . . 8

1 Functions of Several Variables 12
- 1.1 Scalar functions of one variable . . . 13
- 1.2 Scalar functions of several variables . . . 15
- 1.3 Vector functions of several variables . . . 18
- 1.4 Visualizing functions . . . 26
- 1.5 What is multivariate calculus useful for? . . . 34

2 Inner Product Spaces and the Spectral Theorem 35
- 2.1 Inner product spaces . . . 36
- 2.2 Open and closed sets . . . 49
- 2.3 Projections onto a line . . . 53
- 2.4 Orthonormal basis . . . 55
- 2.5 The Gram-Schmidt process . . . 58
- 2.6 Unitary and orthogonal matrices . . . 62
- 2.7 Diagonalizable matrices . . . 66

2.8 The Spectral Theorem 67
2.9 Positive definite and semi-definite matrices 75
2.10 Projections onto a subspace 77

3 Continuity and Differentiability 80
3.1 Analysis of functions of one variable 80
3.2 Continuity of vector functions of several variables 87
3.3 Partial derivatives of first order and the gradient vector . 93
3.4 Directional derivatives 98
3.5 Partial derivatives of second order and the Hessian matrix 100
3.6 Differentiability of scalar functions of several variables . . 104
3.7 The chain rule for scalar functions 109
3.8 Differentiability of vector functions of several variables . . 112

4 Taylor Approximation 117
4.1 The tangent for a function of one variable 118
4.2 Taylor polynomials for functions of one variable 121
4.3 Taylor's formula for functions of one variable 124
4.4 The tangent plane for functions of several variables 128
4.5 Taylor polynomials for functions of several variables . . . 130
4.6 Taylor's formula for functions of several variables 134
4.7 Taylor polynomials for vector functions of several variables 136

5 Local and Global Extrema of Functions 138
5.1 The range of functions of one variable 139
5.2 The range of functions of several variables 143

6 Integration 151
6.1 The Riemann integral of functions of one variable 151
6.2 Anti-derivatives of functions of one variable 155
6.3 The Riemann integral of functions of two variables 160
6.4 Change of variables in $\mathbb{R}^2$ 167
6.5 Polar coordinates 172
6.6 The Riemann integral of functions of several variables . . 176
6.7 The Riemann integral of vector functions 182

7 Vector Fields 184
7.1 Parametric curves and surfaces 185
7.2 Line and surface integrals 190
7.3 Vector fields and gradient fields 195
7.4 Line integrals of vector fields and computing anti-derivatives 201

7.5 Surface integrals of vector fields 207

Appendices **210**

A Additional Proofs **211**
A.1 Proof of Taylor's theorem 211

B Symbols and terminology **215**
B.1 Short list of symbols 215
B.2 Short mathematical dictionary (English ↔ Danish) 216
B.3 Some well-known formulas 223

Bibliography **225**

Index **226**

List of Figures

1.1 Domain of the function in Example 1.2.2. 17
1.2 Sample image from the NIST Special Database 19 of handwritten digits. 25
1.3 Input-output view of a vector function of n variables. 26
1.4 The graph of the function $f(x_1, x_2) = x_1^2 + x_2^2 - 5$. 27
1.5 Illustration of the function $\boldsymbol{V}(x_1, x_2) = (-x_2/3, x_1/3)$. 29
1.6 The function $f(x_1, x_2) = x_1^2 + x_2^2 + 3$ and some level sets. . . . 30

2.1 Illustration of the triangle inequality (2.15) in $\mathbb{R}^2$. 41
2.2 Illustration of the Pythagorean theorem in $\mathbb{R}^2$. 47
2.3 Orthogonal projection in $\mathbb{R}^2$. 54
2.4 Illustration of the Gram-Schmidt process in $\mathbb{R}^2$. 62

3.1 The graph of the function $f(x) = \mathrm{ReLU}(x)$. 85

4.1 The function $f(x) = \sin(x)$ and its first-degree Taylor polynomial. 119
4.2 The function f in (4.4) and its first-degree Taylor polynomial. 120
4.3 The functions in (4.5). 121
4.4 The function $f(x) = \mathrm{e}^x$ and its fifth-degree Taylor polynomial. 123
4.5 The function f in (4.9) and its fourth-degree Taylor polynomial. 124
4.6 The function $f(x, y) = \sin(x^2 + y^2)$ and its Taylor polynomials of first degree and second degree. 135

6.1 Riemann sums of the function $f(x) = x^{-1}$ on $x \in [1/2, 1]$. . . . 153
6.2 The domain B in (6.48). 166
6.3 The domain B in (6.52). 168
6.4 The subset $B \subset \mathbb{R}^2$ given by Equation (6.62) on page 174. . . 174
6.5 Illustration of cylindrical and spherical coordinates. 181

7.1 The cylinder as in Equation (7.13) with $r = 1$ and $h = 4$. . . . 193

7.2 The graph surface associated with the function h in (7.17). . . 194

CHAPTER 0

Preliminaries

In this initial chapter, we recall some concepts from calculus and linear algebra that will be used throughout the book. The chapter is designed as a reference section to acquaint you with the essential notations and terminology, mainly from linear algebra, used throughout this book. It is meant to be consulted as needed, rather than read sequentially. Use it to clarify concepts and to familiarize yourself with the language of vector spaces and matrices.

0.1 Scalars

A field is a set $\mathbb{F}$ with two operations, addition $+$ and multiplication $\cdot$, and we refer to Definition 6.1.1 in Mathematics 1a for the formal definition of a field. The real numbers $\mathbb{R}$ and the complex numbers $\mathbb{C}$ are all examples of a field.

In this book, we use $\mathbb{F}$ as a common symbol for either $\mathbb{R}$ or $\mathbb{C}$. It is possible to consider other fields $\mathbb{F}$, e.g., the rational numbers $\mathbb{Q}$, but in this book $\mathbb{F}$ always denote $\mathbb{R}$ or $\mathbb{C}$. The numbers in $\mathbb{F}$ are called scalar. Hence, a *scalar* is always either a real or complex number.

In [Mathematics 1a] we introduced the complex numbers $\mathbb{C}$ from the real numbers $\mathbb{R}$ as the set of all numbers of the form $c_1 + ic_2$, where c_1 and c_2 are real numbers and i is the imaginary unit. Let $c = c_1 + ic_2$ be a complex scalar, where $c_1 = \operatorname{Re} c$ and $c_2 = \operatorname{Im} c$ are real numbers. Let us recall some rules (where d is another complex scalar) that will be useful later:

(i) $c\overline{c} = c_1^2 + c_2^2 = |c|^2$

(ii) $\overline{c + d} = \overline{c} + \overline{d}$ and $\overline{cd} = \overline{c}\,\overline{d}$

(iii) $c + \overline{c} = 2c_1 = 2 \operatorname{Re} c$

(iv) $(c-d)^2 = c^2 + d^2 - 2cd$ and $(c-d)(c+d) = c^2 - d^2$

Even though, we have introduced the complex numbers as an extension of the real numbers, you have probably never seen how the real numbers $\mathbb{R}$ are introduced. This is done in the next subsection.

The real numbers. We assume the reader is familiar with the rational numbers $\mathbb{Q}$ and their properties. The rational numbers are a *ordered* field, meaning that they satisfy the axioms of a field and that there is a *total order* relation $<$ on $\mathbb{Q}$ which is compatible with $\mathbb{Q}$ as a field. An order relation allows us to compare the size of two numbers, i.e., given two numbers, we can say that one of them is smaller than or equal to the other. The complex numbers $\mathbb{C}$ are not ordered, however, both $\mathbb{Q}$ and $\mathbb{R}$ are ordered fields.

The rational numbers are a useful set of numbers, but they have limitations. For example, the equation $x^2 = 2$ has no solution in $\mathbb{Q}$. To overcome this limitation, we introduce the real numbers $\mathbb{R}$ as an extension of the rational numbers. The real numbers are a *complete* ordered field, meaning that any nonempty subset of $\mathbb{R}$ which is bounded above has a *supremum* in $\mathbb{R}$.

Definition 0.1.1 Supremum

A nonempty subset $A \subset \mathbb{R}$ is *bounded above* if:

$$\exists M \in \mathbb{R}\ \forall x \in A : x \leq M.$$

Any such M is an *upper bound* for A. A real number L is a *least upper bound* for A (or a *supremum* for A) if:

(i) L is an upper bound for A, and

(ii) $\forall M \in \mathbb{R}:\ M$ is an upper bound for $A \implies L \leq M$.

We use the notation $\sup(A) := L$.

Similar definitions apply to subsets of $\mathbb{R}$ which are *bounded from below*:

(i) Any real number M with the property $\forall x \in A : x \geq M$ is a *lower bound* for A.

(ii) A real number L which is a lower bound for A and, in addition, satisfies:

$$\forall M \in \mathbb{R}:\ M \text{ is a lower bound for } A \implies L \geq M$$

is a *greatest lower bound* for A, or an *infimum* for A.

We use the notation $\inf(A) \coloneqq L$. If a set is bounded both above and below, we say it is *bounded*.

Remark 0.1.1

The supremum of A is unique. However, $\sup(A)$ is not necessarily an element of A; when it is, we say it is the *maximum* of A. The infimum of a set is also unique, and $\inf(A)$ may fail to be an element of A. When it is in A, we say it is the *minimum* of A.

We are now ready to define the real numbers $\mathbb{R}$.

Definition 0.1.2

The set of real numbers $\mathbb{R}$ is the unique[a] complete ordered field. This means that $\mathbb{R}$ satisfies the following axioms:

(i) (algebraic axioms) $(\mathbb{R}, +, \cdot, 0, 1)$ is a field. This means:

- $(\mathbb{R}, +, 0)$: Addition $(+)$ is an associative, commutative operation with neutral element 0. Any $x \in \mathbb{R}$ has a unique additive inverse $-x$ (meaning $x + (-x) = 0$).
- $(\mathbb{R}, \cdot, 1)$: Multiplication $(\cdot)$ is an associative, commutative operation with neutral element 1. Any $x \in \mathbb{R}$, $x \neq 0$ has a unique multiplicative inverse x^{-1} (meaning $x \cdot x^{-1} = 1$).
- Multiplication distributes over addition: $x \cdot (y + z) = x \cdot y + x \cdot z$.
- Nontriviality: $1 \neq 0$.

(ii) (order axioms) There exists a subset $\mathbb{R}^+ \subset \mathbb{R}$ with the following properties:

- Defining $x \leq y \iff y - x \in \mathbb{R}^+$ establishes a total order relation in $\mathbb{R}$.
- $x \leq y$ and $z \leq w \implies x + z \leq y + w$.
- $x \leq y$ and $z \geq 0 \implies x \cdot z \leq y \cdot z$.

(iii) (supremum axiom) Any nonempty subset $A \subset \mathbb{R}$ which is bounded above has a supremum $L \in \mathbb{R}$.

[a] To be precise, the uniqueness holds only up to so-called isomorphisms. Note that it is a *theorem* that any two complete ordered fields are isomorphic and that a complete ordered field exists. We *define* $\mathbb{R}$ to be one such field.

The set $\mathbb{R}^+$ in Definition 0.1.2 is the set of *positive real numbers*. Note that it follows from the supremum axiom that any nonempty subset $A \subset \mathbb{R}$ which is bounded from below has an infimum, hence there is no need to include this property in the axioms above.

Remark 0.1.2

Note that the set of rational numbers $\mathbb{Q}$ also satisfies the algebraic and order axioms. What distinguishes $\mathbb{R}$ from $\mathbb{Q}$ is the fact that $\mathbb{Q}$ does not satisfy the supremum axiom. That is, a nonempty subset of $\mathbb{Q}$ which is bounded above may fail to have a supremum in $\mathbb{Q}$. For example, the set $A = \{\, q \in \mathbb{Q} \mid q^2 < 2 \,\}$ is bounded above in $\mathbb{Q}$, but has no supremum in $\mathbb{Q}$.

We will not prove the existence of the real numbers here, that is, we will not prove that there exists a set, to be denoted $\mathbb{R}$, which satisfies the axioms in Definition 0.1.2. This proof is rather long and beyond the scope of this book. For a proper construction of the real numbers, we refer to Appendix A of the additional notes. For an alternative existence proof using so-called Didekind cuts, we refer to [4]. We do not expect the reader to be familiar with these proofs, only that the reader is familiar with the real numbers $\mathbb{R}$ and their properties.

However, we often need to find the supremum of a set $A \subset \mathbb{R}$. The following theorem provides a useful characterization of the supremum.

Theorem 0.1.1 Characterization of the supremum and infimum

Let $A \subset \mathbb{R}$ be nonempty and bounded above. Then $L = \sup(A)$ if and only if:

(i) L is an upper bound for A, and

(ii) $\forall \epsilon > 0 \; \exists a \in A : \; L - \epsilon < a \leq L$.

Analogously, if $A \subset \mathbb{R}$ is nonempty and bounded below, $L = \inf(A)$ if and only if:

(i) L is a lower bound for A, and

(ii) $\forall \epsilon > 0 \; \exists a \in A : L + \epsilon > a \geq L$.

Let us show the existence of $\sqrt{2}$ as a supremum in $\mathbb{R}$.

Example 0.1.1

Let $A = \{x \in \mathbb{R} \mid x > 0 \text{ and } x^2 < 2\}$. Let L denote the supremum

$L = \sup(A)$. We know that L exists since A is nonempty and bounded above, and since we have an axiom in $\mathbb{R}$ which guarantees the existence of the supremum. We will argue L satisfies $L^2 = 2$, that is, L is a square root of 2 which we write $L = \sqrt{2}$. Suppose, towards a contradiction, that $L^2 \neq 2$. Then either $L^2 < 2$ or $L^2 > 2$. Assume first that $L^2 < 2$. For any $0 < \epsilon < 1$ we have

$$(L+\epsilon)^2 = L^2 + 2\epsilon L + \epsilon^2 \leq L^2 + 4\epsilon + \epsilon = L^2 + 5\epsilon$$

since $L \leq 2$ and $\epsilon^2 \leq \epsilon$. Since $L^2 < 2$, we can choose a sufficiently small $\epsilon \in]0, 1[$ so that also $L^2 + 5\epsilon < 2$. Thus, it follows that $(L+\epsilon)^2 \leq L^2 + 5\epsilon < 2$. By definition of A, this means that $L + \epsilon \in A$ which contradicts the fact that L is an upper bound of A.

The case $L^2 > 2$ is handled similarly, and we leave it to the reader. Hence, we can conclude that $L^2 = 2$. Note that $L = \sqrt{2}$ is a supremum for A, but not a maximum of A as $L \notin A$.

0.2 Sets

Recall that a set A is a way to "bundle" elements together in one object. If we for example want to write down a set consisting of the numbers 0 and 1, we simply write $\{0, 1\}$ (or $\{1, 0\}$ or more complicated $\{1, 0, 1, 1\}$).

We often construct sets using set-builder notation; in fact, we already did this in Example 0.1.1 above. The set of all even integers, expressed in set-builder notation, is given by:

$$\{\, n \in \mathbb{Z} \mid \exists k \in \mathbb{Z},\, n = 2k \,\}, \tag{1}$$

where $\exists$ reads "there exists". In set-builder notation the vertical bar $\mid$ is a separator and is read "such that". Hence, Equation (1) is read "integers n such that there exists an integer k for which $n = 2k$". One often uses more complicated expressions left of the vertical bar $\mid$ to obtain simpler set-builder notation, e.g., the even integers can be written $\{\, 2k \mid k \in \mathbb{Z} \,\}$.

Sometimes the vertical $\mid$ is replaced with a colon $:$, but the meaning is the same and we will only use it if it increases the readability, e.g., in case the variable or the predicate contains other vertical bars such as:

$$\{\boldsymbol{x} \in \mathbb{R}^n : \|\boldsymbol{x}\| \leq 1\}$$

where $\|\boldsymbol{x}\|$ denotes the norm of a vector (it will be introduced in Equation (2.1) on page 36).

Intervals are subsets of the real numbers $\mathbb{R}$ defined by conditions on their elements. Let $a, b \in \mathbb{R}$, where $a \leq b$. A *closed interval* is written

$[a, b] = \{x \in \mathbb{R} \mid a \leq x \leq b\}$, where both endpoints are included. An *open interval* is written $]a, b[= \{x \in \mathbb{R} \mid a < x < b\}$, excluding the endpoints. Half-open intervals include one endpoint: $[a, b[= \{x \in \mathbb{R} \mid a \leq x < b\}$ and $]a, b] = \{x \in \mathbb{R} \mid a < x \leq b\}$. If $a = b$, the open and half-open intervals represent the empty set, while the closed interval is $\{a\}$. Unbounded intervals are intervals such as $]-\infty, b] = \{x \in \mathbb{R} \mid x \leq b\}$.

The set of all integers is $\mathbb{Z} = \{\ldots, -2, -1, 0, 1, 2, \ldots\}$. Positive integers are written $\mathbb{N} = \{1, 2, 3, \ldots\}$, and non-negative integers as $\mathbb{N}_0 = \{0, 1, 2, 3, \ldots\}$. In some references, $\mathbb{N}$ denotes the non-negative integers, however, we will use $\mathbb{N}_0$ for this purpose. To avoid confusion, one can use the notation $\mathbb{Z}_{>0}$ for the positive integers and $\mathbb{Z}_{\geq 0}$ for the non-negative integers.

0.3 Functions $f : A \to B$

Definition 0.3.1 Function

Let A and B denote two sets. A *function* is a correspondence, which to each element $x \in A$ in a unique way associates an element $y \in B$. Denoting the function by f, this will be written as

$$f : A \to B,\ y = f(x). \tag{2}$$

The set A is called the *domain* of the function f and wil be denoted by $\mathrm{dom}(f)$. The set B is called the *codomain* set and will denoted by $\text{co-dom}(f)$.

A function can be given in terms of an explicit expression, e.g., we can define the quadratic function by

$$f : \mathbb{R} \to \mathbb{R},\ \ f(x) = x^2. \tag{3}$$

However, a function can also be given in terms of, e.g., a table, which to each element $x \in A$ simply lists the corresponding element $y = f(x) \in B$.

Note that f is the *function* while $f(x)$ is the *function value*, i.e., a real number (scalar) if $B = \mathbb{R}$. A function f consists of a set of inputs (the domain $x \in A$), a set within all outputs occur (the codomain $y \in B$), and a rule for assigning each input to *exactly one* output, which we write as $x \mapsto y$. Hence, we can, alternatively, define the quadratic function as

$$f : \mathbb{R} \to \mathbb{R},\ \ f = x \mapsto x^2,$$

where $x \mapsto x^2$ reads "x (the input) maps to x^2 (the output)".

Two functions $f : A \to B$ and $g : C \to D$ are equal precisely if they have the same domain, the same codomain, *and* they assign the same values to each of the elements of their domain. Hence, if we define $g : [0, 1] \to \mathbb{R}$, $g = x \mapsto x^2$, then quadratic functions f and g are not the same function even though the expressions $f(x) = x^2 = g(x)$ agree. This is because $A = \mathbb{R}$ while $C = [0, 1]$ in this situation, in particular, $f(2) = 4$ while $g(2)$ is not defined.

The definition of a function does not require that A and B are sets of scalars — they can be arbitrary sets. For example, the correspondence which to each item in a shop associates its brand, is a function. Here the set A is the collection of all items in the shop, and B is the collection of all brands. Note that we indeed for each item in the shop can identity a unique brand. On the other hand, different items can very well be associated with the same brand.

Note that the concept of the codomain B should be clearly distinguished from the range of the function f. Indeed, the set B in Definition 0.3.1 is just a set, within which all function values occur! We do *not* claim that each of the elements in B occur as a function value. On the other hand, the *range* or *image* of a function f is defined as precisely the set of all function values $f(x)$, where x runs through the set A:

$$\mathrm{im}(f) = \{\, f(x) \mid x \in A \,\}. \tag{4}$$

The $\mathrm{im}(f)$ is sometimes written as $f(A)$. For the function f in (3), the codomain is $\mathbb{R}$, but the range is only the set of nonnegative real scalars, i.e., the interval $[0, \infty[$.

Definition 0.3.2 Surjective, Injective, Bijective

Consider a function $f : A \to B$.

(i) The function f is said to be *surjective* if the range of f equals the set B; that is, for each $y \in B$ there exists some $x \in A$ such that $f(x) = y$.

(ii) The function f is said to be *injective* if different elements in the set A get mapped to different elements in B, that is, if

$$x_1, x_2 \in A, x_1 \neq x_2 \Rightarrow f(x_1) \neq f(x_2).$$

(iii) The function f is said to be *bijective* if f is surjective and injective.

(iv) Let $A' \subseteq A$. The function f is said to be *injective on* the subset A' if

$$x_1, x_2 \in A', x_1 \neq x_2 \Rightarrow f(x_1) \neq f(x_2).$$

Example 0.3.1

Let us illustrate the concepts in Definition 0.3.2 with a number of examples.

(a) The function $f_1 : \mathbb{R} \to \mathbb{R}, f_1(x) = x^2$ is neither injective nor surjective. The function f_1 is not injective since, e.g., $f_1(-1) = f_1(1) = 1$ and not surjective since $f_1(x) \geq 0$ for all $x \in \mathbb{R}$. The function is injective *on* the subset $A' \coloneqq [0, \infty]$.

(b) The function $f_2 : \mathbb{R} \to \mathbb{R}, f_2(x) = x^3$ is injective and surjective, hence bijective.

(c) The function $f_3 : \mathbb{R} \to \mathbb{R}, f_3(x) = \mathrm{e}^x$ is injective, but not surjective. The function is not surjective since for, e.g., $y = -1$, the equation $f_3(x) = y$ has no solution x in $\mathbb{R}$. To see this, recall that $\mathrm{e}^x > 0$ for all $x \in \mathbb{R}$.

0.4 Matrices and vectors

An $m \times n$ *matrix* is an array of real or complex numbers of the form

$$A = [a_{ij}] = \begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,n} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m,1} & a_{m,2} & \cdots & a_{m,n} \end{bmatrix},$$

where $a_{i,j} \in \mathbb{R}$ or $\mathbb{C}$ is the (i, j) entry of the matrix A. We sometimes write a_{ij} (without a comma) instead of $a_{i,j}$. The set of all $m \times n$ matrices with entries in $\mathbb{F}$ is denoted by $\mathsf{M}_{m\times n}(\mathbb{F})$. Real $m \times n$ matrices are then denoted $\mathsf{M}_{m\times n}(\mathbb{R})$ and complex $m \times n$ matrices are denoted $\mathsf{M}_{m\times n}(\mathbb{C})$. The set of square matrices, i.e., $m = n$, are denoted $\mathsf{M}_n(\mathbb{F}) \coloneqq \mathsf{M}_{n\times n}(\mathbb{F})$.

The $m \times n$ zero matrix is denoted by $\mathbf{0}_{m\times n}$. The zero (column) vector $\mathbf{0}_{n\times 1}$ is denoted by $\mathbf{0}_n$ or simply $\mathbf{0}$. The $n \times n$ identity matrix is denoted I_n or just I.

Block matrices are matrices that are partitioned into smaller matrices. We have already seen such matrices before, e.g., the augmented matrix $[A \,|\, \boldsymbol{b}]$ in the context of linear systems. We will write such block matrices as $[A \,|\, \boldsymbol{b}]$, $[A, \boldsymbol{b}]$ or $[A \;\; \boldsymbol{b}]$. Thus, if $\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_n$ are column vectors in $\mathbb{F}^m$, then the matrix A whose columns are $\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_n$ can be written as $A = [\boldsymbol{a}_1 \,|\, \boldsymbol{a}_2 \,|\, \cdots \,|\, \boldsymbol{a}_n]$, $A = [\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_n]$ or

$$A = \begin{bmatrix} \boldsymbol{a}_1 & \boldsymbol{a}_2 & \cdots & \boldsymbol{a}_n \end{bmatrix}.$$

If $\boldsymbol{u}$ is a column vector in $\mathbb{F}^n$ and U_1 is an $n \times (m-1)$ matrix, then the $n \times m$-block matrix U whose first column is $\boldsymbol{u}$ and the remaining columns are the columns of U_1 is denoted by $[\boldsymbol{u} \,|\, U_1]$, $[\boldsymbol{u}, U_1]$ or $[\boldsymbol{u}\; U_1]$.

We recall some properties and terminology of matrix theory.

Inverse. We say that $A \in \mathsf{M}_n$ is an *invertible matrix* if there exists a $B \in \mathsf{M}_n$ such that

$$AB = I_n \quad \text{and} \quad BA = I_n. \tag{5}$$

Such a matrix B is an inverse of A. If A has no inverse, then A is non-invertible or singular. Either of the equalities in (5) implies the other, that is, for *square*[1] matrices $A, B \in \mathsf{M}_n$, we have $AB = I$ if and only if $BA = I$.

As we know not every square matrix has an inverse, but a matrix has at most one inverse. If A is invertible, we can therefore speak of *the* inverse of A. If A is invertible, then the inverse of A is denoted by A^{-1}. A matrix and its inverse satisfy:

$$AA^{-1} = I = A^{-1}A.$$

If A is invertible, we define $A^{-k} = (A^{-1})^k$ for $k = 1, 2, \ldots$.

Suppose A and B are invertible matrices. Let j, k be integers, and let c be a scalar. Then:

(i) $(A^k)^j = A^{jk} = (A^j)^k$.

(ii) $(A^{-1})^{-1} = A$.

(iii) If $c \neq 0$, then $(cA)^{-1} = c^{-1}A^{-1}$.

(iv) AB is invertible and $(AB)^{-1} = B^{-1}A^{-1}$.

Transpose. The transpose of $A = [a_{ij}] \in \mathsf{M}_{m\times n}$ is the matrix $A^T \in \mathsf{M}_{n\times m}$ whose (i, j) entry is a_{ji}. Let A and B be matrices of appropriate sizes and let c be a scalar.

(i) $(A^T)^T = A$.

(ii) $(A + B)^T = A^T + B^T$.

(iii) $(cA)^T = cA^T$.

(iv) $(AB)^T = B^T A^T$.

(v) If A is invertible, then $(A^T)^{-1} = (A^{-1})^T$. We write $(A^{-1})^T = A^{-T}$.

[1] Be careful! This statement is false for non-square matrices.

Conjugate. The conjugate of $A \in \mathsf{M}_{m\times n}$ is the matrix $\overline{A} \in \mathsf{M}_{m\times n}$ whose (i,j) entry is $\overline{a_{ij}}$, the complex conjugate of a_{ij}. Thus,

$$\overline{(\overline{A})} = A, \quad \overline{A+B} = \overline{A} + \overline{B}, \quad \overline{AB} = \overline{A}\,\overline{B}.$$

If A is a real matrix, then $\overline{A} = A$.

Adjoint (conjugate transpose). The adjoint of A is defined by $A^* = \overline{A^T}$, meaning that we take the transpose of the matrix, and then take the complex conjugate of each entry (or the other way around; the order does not matter which you can easily check). Hence, the (i,j) entry of A^* is $\overline{a_{j,i}}$. Note, that for a real matrix A taking transpose and adjoint is the same operation, i.e., $A^* = A^T$. The adjoint of a matrix is also called the conjugate transpose since this is what the operation does. Let A and B be matrices of appropriate sizes and let c be a scalar.

(i) $I^* = I_n$.

(ii) $(0_{m\times n})^* = 0_{n\times m}$.

(iii) $(A^*)^* = A$.

(iv) $(A+B)^* = A^* + B^*$.

(v) $(cA)^* = \overline{c}A^*$.

(vi) $(AB)^* = B^*A^*$.

(vii) If A is invertible, then $(A^*)^{-1} = (A^{-1})^*$. We write $(A^{-1})^* = A^{-*}$.

Special Types of Square Matrices. Let $A \in \mathsf{M}_n(\mathbb{C})$.

(i) If $A^* = A$, then A is *Hermitian*.

(ii) If $A^T = A$, then A is *symmetric*.

(iii) If $A^*A = I$, then A is *unitary*; if A is real and $A^TA = I$, then A is *real orthogonal*.

(iv) If $A^*A = AA^*$, then A is *normal* .

(v) If $A^2 = A$, then A is *idempotent*. An projection matrix is an idempotant matrix.

(vi) If A is Hermitian and has (strictly) positive eigenvalues, then A is *positive definite*.

(vii) If A is Hermitian and has non-negative eigenvalues, then A is *positive semi-definite*.

(viii) If A is idempotent and Hermitian, then A is an *orthogonal projection matrix*.

Trace The *trace* of $A = [a_{ij}] \in \mathsf{M}_n$ is the sum of the diagonal entries of A:

$$\operatorname{tr} A = \sum_{i=1}^{n} a_{ii}.$$

For $A, B \in \mathsf{M}_n$ and $c \in \mathbb{C}$, we have:

(i) $\operatorname{tr}(cA + B) = c \operatorname{tr} A + \operatorname{tr} B$.

(ii) $\operatorname{tr}(A^T) = \operatorname{tr} A$.

(iii) $\operatorname{tr}(cA) = c \operatorname{tr} A$.

(iv) $\operatorname{tr}(A^*) = \overline{\operatorname{tr} A}$.

CHAPTER 1

Functions of Several Variables

The concept of a function is probably the most fundamental topic in mathematical analysis. It provides the language for formulating basic laws in physics and engineering in a mathematical language and hereby calculate how we can obtain a desired effect in a given physical/mechanical system. Questions like

> *What is the necessary temperature of a gas in a tank that is necessary in order for the pressure in the gas to reach a certain value?*

can easily be phrased and solved using functions.

We expect the basic concepts of functions $f : A \to B$ in Section 0.3 on page 6 to be known to the reader. Standard calculus is concerned with the case where the domain of the function Equation (2) on page 6 is the set of real scalars, like in (3), or a subset hereof. Our goal in the current chapter is to generalize our knowledge of such functions to the case where A and B are sets of (column) vectors in $\mathbb{R}^n$ and $\mathbb{R}^k$, respectively. For example, this will allow us to consider real-valued functions of the form

$$f : \mathbb{R}^2 \to \mathbb{R}, \quad f(x, y) = x + y^2 \tag{1.1}$$

and vector-valued functions of the form

$$f : \mathbb{R} \to \mathbb{R}^3, \quad f(t) = \begin{bmatrix} \cos(t) \\ \sin(t) \\ t \end{bmatrix}.$$

In Section 1.2 we introduce scalar (also called real-valued) functions depending on several real variables, and in Section 1.3 we turn to the general case of vector-valued functions of several variable. In order to get a good visual understanding of a function of several variables, we need the so-called level sets which is the topic of Section 1.4.

1.1 Scalar functions of one variable

Functions depending on a single real variable are well-known from elementary calculus. In this section we will remind the reader about some of the most important concepts and results. We consider functions $f : I \to \mathbb{R}$, where either $I = \mathbb{R}$ or I is an open interval of the form $]-\infty, a[$, $]a, \infty[$ or $]a, b[$ for some $a, b \in \mathbb{R}$.

We already saw the first example of such a function in (3). In general, when dealing with a function, it is important to specify as well the expression for the function as its domain. For example, the natural logarithm $f(x) = \ln(x)$ is only defined for positive values of the variable x; thus, the function must be described as

$$f : \mathrm{dom}(f) = \{x \in \mathbb{R} \,\big|\, x > 0\} \to \mathbb{R}, \;\; f(x) = \ln(x).$$

From elementary calculus we are already familiar with several classes of functions, e.g., polynomials, exponential functions, and trigonometric functions.

Example 1.1.1

Functions also appear naturally within all branches of science and engineering, typically denoted by different letters than in the mathematical literature:

(a) Consider an (ideal) gas that is contained in a tank with volume V. Assuming that the gas contains n particles (measured in moles), the pressure P in the tank and the temperature T of the gas are related by $PV = nRT$, where R is the ideal gas constant (approximately 0.0821.) This can also be written as

$$P = \frac{nRT}{V}. \tag{1.2}$$

If the volume V of the tank is constant, (1.2) shows how the pressure in the tank varies with changing temperature. This means precisely that the pressure P is a function of the temperature T, i.e., we can write

$$P(T) = \frac{nRT}{V}.$$

(b) Consider an electric circuit, where a battery is connected with a resistant R. By Ohm's law, the voltage V delivered by the battery and the current I in the circuit are related by $V = RI$. Thus $I = V/R$. Considering the voltage V to be constant, this relationship shows how

the current I depends on the resistant R. In other words, the current is a function of R,

$$I(R) = \frac{V}{R}.$$

Remark 1.1.1

Note that when introducing the functions in Example 1.1.1, we have deliberately been sloppy. The problem is that we have only specified the function expression of, e.g., the pressure $P(T) = nRT/V$ in the tank, but not the domain and codomain of the function. Both the domain and codomain is part of the definition of the function so here we should specify $P : [0, \infty[\to \mathbb{R}$ (or perhaps we should even replace the domain $[0, \infty[$ with $[0, T_P]$ where $T_P \approx 1.416 \cdot 10^{32}$ is the Planck temperature measured in Kelvin).

Not all functions can be described explicitly by a formula or closed-form expression. Instead, some functions are defined implicitly as solutions to equations, or they are described algorithmically, or they are defined by a table of values, or by other means. Let us consider a couple of examples of such functions.

Example 1.1.2 Next prime function

Let $f : \mathbb{R} \to \mathbb{R}$ be the function that maps a real number x to the smallest prime number that is strictly larger than x. For example, $f(2) = 3$, $f(\pi) = 5$, and $f(13) = 17$. The function f is well-defined (meaning that the mapping satisfies the requirement of a function in Definition 0.3.1 on page 6) because for every real number x there is a unique prime number p that is strictly larger than x. However, there is no simple formula that can express $f(x)$ in terms of elementary functions like polynomials, exponentials, or trigonometric functions. The function f is known as the *next prime function*.

Example 1.1.3 A special function

Let the function $f :]0, \infty] \to \mathbb{R}$ be defined by $f(x) = y$, where y is the unique real number satisfying

$$ye^y = x, \quad x > 0.$$

This equation implicitly defines $f(x)$, but there is no simple formula that can express $f(x)$ in terms of elementary functions like polynomials, exponentials, or trigonometric functions.

The function $f(x)$ is well-defined because the equation $f(x)e^{f(x)} = x$ has *exactly one* real solution $f(x)$ for every positive input $x > 0$. However,

computing the value of $f(x)$ requires actually solving the equation, which is often done numerically. For example, $f(1) \approx 0.567143$. Of course, some function values can be computed exactly, e.g., $f(e) = 1$. As an exercise, can you find the value of $f(2e^2)$?

The function f is known as the *principal branch of the Lambert W function*. Special functions like this one arise naturally in many areas of science and engineering.

1.2 Scalar functions of several variables

In this section we will take the first steps to generalize the theory for functions of one variable to functions of several variables. The desire to do this comes from science and engineering, where we often face the need of considering several changing parameters simultaneously. Let us illustrate this with an example.

Example 1.2.1

We consider again the two examples in Example 1.1.1 on page 13:

(a) The pressure P in the tank is expressed by the formula (1.2). If we imagine that as well the volume V of the tank as the temperature T can vary, the pressure becomes a function of two variables:

$$P(T, V) = \frac{nRT}{V}.$$

(b) If we replace the battery by a power supply with an adjustable output, the current I in the circuit becomes a function of two variables, namely, the output V of the power supply and the resistant R,

$$I(R, V) = \frac{V}{R}.$$

We will now consider functions that are defined either on $\mathbb{R}^n$ for some $n \in \mathbb{N}$, or a subset hereof. We saw the first example of such a function in (1.1). We will typically denote the variables by $x_1, x_2, \ldots, x_n$; however, for functions of just two variables we will also apply the shorter names x and y for the variables.

For functions of n variables the vector notation from linear algebra is very useful. For example, instead of writing a function value as $f(x_1, x_2, \ldots, x_n)$, we can define $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ and then write the function value as $f(\boldsymbol{x})$.

Visually, this is similar to our standard writing $f(x)$ for a function of one variable, except that now the boldface $\boldsymbol{x}$ indicates that the input is a vector. Note that strictly speaking $(x_1, x_2, \ldots, x_n)$ is a n-tuple real numbers. We will, however, most often consider $\boldsymbol{x}$ as a column vector in $\mathbb{R}^n$:

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

or (less often) a row vector $\boldsymbol{x} = [x_1, \ldots, x_n]$ in $\mathbb{R}^n$. Even when we consider $\boldsymbol{x}$ as a vector, we will also, when there is no danger of confusion, use the tuple notation $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$. Unless stated otherwise, we consider vectors in $\mathbb{R}^n$ as column vectors.

The vector notation allows us to formulate several definitions and key results for functions of n variable in a very similar way as for functions of one variable (compare, e.g., the definition of continuity in Definition 3.1.1 on page 81 with the subsequent Definition 3.2.1 on page 88).

The following class of scalar functions of n variables called *quadratic forms* will play an important role throughout the text.

Definition 1.2.1 Quadratic form

Let $A \in \mathsf{M}_{n\times n}(\mathbb{R})$ be a square, non-zero matrix, let $\boldsymbol{b} \in \mathbb{R}^n$ be a column vector, and $c \in \mathbb{R}$ a scalar. The scalar function $q : \mathbb{R}^n \to \mathbb{R}$ given by

$$q(\boldsymbol{x}) = \boldsymbol{x}^T A\boldsymbol{x} + \boldsymbol{x}^T\boldsymbol{b} + c \tag{1.3}$$

where $\boldsymbol{x}$ is considered as a column vector in $\mathbb{R}^n$, is called a *quadratic form.*[a]

[a]The term *quadratic form* is often used to denote the quadratic term $\boldsymbol{x}^T A\boldsymbol{x}$ (without the linear $\boldsymbol{x}^T\boldsymbol{b}$ and constant terms c).

Note that if $n = 1$, the quadratic form is a polynomial function of degree two since (1.3) reduces to $q(x) = Ax^2 + bx + c$. Indeed, a quadratic form is a multivariate polynomial of degree at most two.

Determining a domain. As for functions of a single variable, it is important to be able to describe the domain for a function of n variables. Typically this is more involved than for functions of one variable as illustrated by the following example.

Example 1.2.2 Largest domain

We would like to describe the domain for the function expression:

$$f(x_1, x_2) = \ln\left(x_1^2 - 2x_1 + x_2^2\right). \tag{1.4}$$

The natural logarithm is only defined on positive scalars so the function f is defined on the set of $(x_1, x_2) \in \mathbb{R}^2$ for which $x_1^2 - 2x_1 + x_2^2 > 0$. Now, by a direct calculation,

$$x_1^2 - 2x_1 + x_2^2 = (x_1 - 1)^2 - 1 + x_2^2,$$

so the condition $x_1^2 - 2x_1 + x_2^2 > 0$ means precisely that

$$(x_1 - 1)^2 + x_2^2 > 1.$$

Thus the domain for the function f is

$$\mathrm{dom}(f) = \{\, (x_1, x_2) \in \mathbb{R}^2 \mid (x_1 - 1)^2 + x_2^2 > 1 \,\}.$$

Note that $\mathrm{dom}(f)$ has a geometrical description, see Figure 1.1. It is the set of points $(x_1, x_2) \in \mathbb{R}^2$ that are located *outside* the circle in $\mathbb{R}^2$ centered at the point $(1, 0)$ and with radius 1.

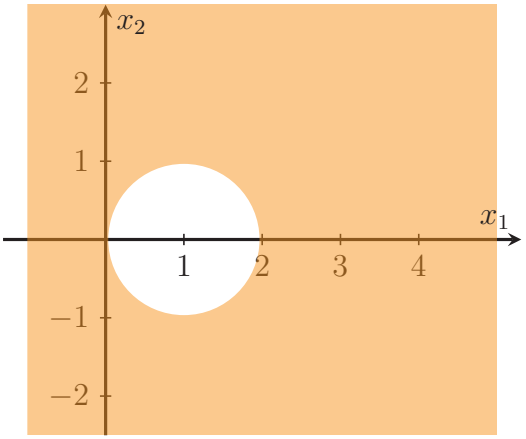


Figure 1.1: Domain of the function in Example 1.2.2 is everything strictly *outside* the circle of radius one centered at $(1, 0)$.

Note that the domain $\mathrm{dom}(f)$ found in Example 1.2.2 is the largest possible domain of the real-valued function expression (1.4). It is perfectly

fine to introduced *another* function with a smaller domain, e.g.,

$$g :]2, \infty[\times \mathbb{R} \to \mathbb{R}, \quad g(x_1, x_2) = \ln\left(x_1^2 - 2x_1 + x_2^2\right).$$

1.3 Vector functions of several variables

Using the vector notation from linear algebra introduced above, we consider a vector function as a function that takes column vectors in $\mathbb{R}^n$ as input and outputs column vectors in $\mathbb{R}^k$:

Definition 1.3.1 Vector function

Let $n, k \in \mathbb{N}$. A *vector function of several variables* is a function of the form

$$\boldsymbol{f} \colon \operatorname{dom}(\boldsymbol{f}) \to \mathbb{R}^k, \text{ where the domain } \operatorname{dom}(\boldsymbol{f}) \text{ is a subset of } \mathbb{R}^n.$$

Since $\boldsymbol{f}(\boldsymbol{x})$ is a vector in $\mathbb{R}^k$ for each $\boldsymbol{x} \in \operatorname{dom}(\boldsymbol{f})$, we will write $\boldsymbol{f} = (f_1, \ldots, f_k)$ or

$$\boldsymbol{f}(\boldsymbol{x}) = \begin{bmatrix} f_1(\boldsymbol{x}) \\ \vdots \\ f_k(\boldsymbol{x}) \end{bmatrix}, \tag{1.5}$$

where $f_i \colon \operatorname{dom} f \to \mathbb{R}$ is the ith *coordinate function* of $\boldsymbol{f}$. In other words, the ith coordinate function is defined *from* $\boldsymbol{f}$ by setting $f_i(\boldsymbol{x}) \coloneqq (\boldsymbol{f}(\boldsymbol{x}))_i$, where $(\boldsymbol{f}(\boldsymbol{x}))_i$ is the ith coordinate of the vector $\boldsymbol{f}(\boldsymbol{x})$. The coordinate functions are *scalar* functions of n variables, and they all have domain $\operatorname{dom}(\boldsymbol{f})$. A vector function $\boldsymbol{f} = (f_1, \ldots, f_m) : \operatorname{dom}(\boldsymbol{f}) \to \mathbb{R}^k$ can often be analyzed by considering the scalar-valued coordinate functions $f_i \colon \operatorname{dom}(f) \to \mathbb{R}$ separately.

We already know examples of vector functions from basic linear algebra, namely, *linear mapping* from $\mathbb{R}^n$ to $\mathbb{R}^k$.

Example 1.3.1 Linear and affine maps

Let $A \in \mathsf{M}_{k\times n}(\mathbb{R})$ be a rectangular matrix and define the linear map $L_A : \mathbb{R}^n \to \mathbb{R}^k$ by $L_A = \boldsymbol{x} \mapsto A\boldsymbol{x}$, i.e., $L_A\boldsymbol{x} = A\boldsymbol{x}$. Then L_A is a vector function of n variables, and its image is the column space of the matrix A, i.e., $\operatorname{im}(L_A) = \operatorname{col}(A)$. The ith coordinate function of L_A is given by $(L_A)_i(\boldsymbol{x}) = a_{i,1}x_1 + a_{i,2}x_2 + \cdots + a_{i,n}x_n$ for $i = 1, \ldots, k$.

Linear mappings L_A is a subclass of *affine mappings* $T_{A,\boldsymbol{b}} : \mathbb{R}^n \to \mathbb{R}^k$ which are vector functions of the form:

$$T_{A,\boldsymbol{b}}(\boldsymbol{x}) = A\boldsymbol{x} + \boldsymbol{b} \quad \text{for } \boldsymbol{x} \in \mathbb{R}^n, \tag{1.6}$$

where $\boldsymbol{b} \in \mathbb{R}^k$.

Example 1.3.2 Non-linear activation functions

In machine learning, so-called neural networks are used to learn complex relationships between input and output data. A key ingredient in neural networks is the use of non-linear *activation functions*. In fact, it is crucial for the success of neural networks that these activation functions are *non-linear*, as this allows the network to learn and represent complex patterns in the data.

One of the most popular activation function is the so-called *rectified linear unit* function:

$$\mathrm{ReLU} : \mathbb{R} \to \mathbb{R}, \ \mathrm{ReLU}(x) = \max(x, 0). \tag{1.7}$$

We can write the function on explicit form as

$$\mathrm{ReLU}(x) = \begin{cases} x & \text{for } x \geq 0, \\ 0 & \text{for } x < 0. \end{cases}$$

This function is indeed non-linear as $\mathrm{ReLU}(-1) = 0 \neq -1 = -1\,\mathrm{ReLU}(1)$. The ReLU function can easily be generalized to a *vector function* $\mathrm{ReLU} : \mathbb{R}^n \to \mathbb{R}^n$ by applying the scalar ReLU function coordinate-wise:

$$\mathrm{ReLU}(\boldsymbol{x}) = \begin{bmatrix} \mathrm{ReLU}(x_1) \\ \mathrm{ReLU}(x_2) \\ \vdots \\ \mathrm{ReLU}(x_n) \end{bmatrix} = \begin{bmatrix} \max(x_1, 0) \\ \max(x_2, 0) \\ \vdots \\ \max(x_n, 0) \end{bmatrix}. \tag{1.8}$$

Strictly speaking, we should write $\mathbf{ReLU}(\boldsymbol{x})$ instead of $\mathrm{ReLU}(\boldsymbol{x})$ to indicate that we are now considering the vector function; however, it is common practice to use the same notation for both the scalar and vector versions of the function.

The ReLU function is widely used for its simplicity as it is highly computationally efficient to evaluate and differentiate. There are, however, many more activation functions used in deep learning, e.g., the sigmoid function $x \mapsto (1 + \mathrm{e}^{-x})^{-1}$ or the well-known hyperbolic tangent function tanh. These scalar functions on $\mathbb{R}$ are also extended to vector functions from $\mathbb{R}^n$ to $\mathbb{R}^n$ by applying the scalar function coordinate-wise. It is common to denote *activation functions* by the Greek letter $\boldsymbol{\sigma}$, where $\boldsymbol{\sigma} : \mathbb{R}^n \to \mathbb{R}^n$.

Example 1.3.3 The softmax activation function

A popular non-linear activation function used in transformer models for natural language processing, is the so-called *softmax function*. Unlike the activation functions discussed in Example 1.3.2 (ReLU, sigmoid, tanh), which are applied coordinate-wise to extend scalar functions to vectors, the softmax function is fundamentally a vector function that depends on all components of the input vector simultaneously. The softmax function maps a vector of real numbers to a probability vector using the exponential function — recall that a probability vector is just a vector with non-negative entries that sum to one so that it can be interpreted as discrete probabilities over n possible outcomes. The softmax function is defined as follows:

$$\text{softmax} : \mathbb{R}^n \to \mathbb{R}^n, \quad \text{softmax}(\boldsymbol{x})_i = \frac{\mathrm{e}^{x_i}}{\sum_{j=1}^{n} \mathrm{e}^{x_j}} \text{ for } i = 1, \ldots, n. \tag{1.9}$$

The softmax function is often used in the final layer of a neural network in cases where the output is to be interpreted as probabilities, e.g., it is a simple way to convert vectors of arbitrary real numbers into a discrete probability distribution. The softmax function is closely related to the argmax function which returns the index of the largest entry in a vector. In fact, the softmax should be called soft*arg*max as it approximates the argmax function in a differentiable way, and therefore is easier to work with in optimization problems.

Let us recap some terminology:

(a) $f\colon \mathbb{R}^n \to \mathbb{R}$ is a *scalar* (or real) function[1] of n variables. This is just a vector function $f\colon \mathbb{R}^n \to \mathbb{R}^k$ where $k = 1$.

(b) $\boldsymbol{f}\colon \mathbb{R}^n \to \mathbb{R}^k$ is a *vector* function of n variables.

Let us also introduce some new terminology:

(c) If $k = n$, the vector function $\boldsymbol{f}\colon \text{dom}(\boldsymbol{f}) \to \mathbb{R}^n$ where $\text{dom}(\boldsymbol{f}) \subseteq \mathbb{R}^n$ is called a *vector field*.

(d) If $k = 1$, the scalar function $f\colon \mathbb{R}^n \to \mathbb{R}$ is also called a scalar field (often the scalar will be a physical quantity with units).

(e) If $n = 1$ and the (vector) function is continuous, we obtain *curves* in $\mathbb{R}^k$. Strictly speaking, the curve is the image $\text{im}(\boldsymbol{f})$ of the vector

[1] also called scalar-valued or real-valued functions

function, not the function itself. If $n = 1$ and $k \geq 3$, we obtain *space curves*, and one could think of a particle path in $\mathbb{R}^k$, e.g., a helix in $\mathbb{R}^3$:

$$\boldsymbol{r}(t) = [\cos(t), \sin(t), t]^T \quad \text{for } t \in I \tag{1.10}$$

where I is usually a finite interval $[a, b]$, but could be infinite, e.g., $\mathbb{R}$. The vector function $\boldsymbol{r} : I \to \mathbb{R}^n$ is called a *parametrization* of the curve $\mathrm{im}(\boldsymbol{r})$.

(f) If $n = 2$ and the (vector) function is continuous, we obtain *surfaces* in $\mathbb{R}^k$. A vector function of the form $\boldsymbol{r} : I \times J \to \mathbb{R}^n$, where I, J are intervals, is called a *parametrization* of the surface $\mathrm{im}(\boldsymbol{r})$. An example is the parametrization of the (unit) sphere in $\mathbb{R}^3$ by

$$\boldsymbol{r}(u, v) = [\cos(u)\cos(v), \sin(u)\cos(v), \sin(v)]^T \tag{1.11}$$

where $u \in [0, 2\pi[$ is the longitude and $v \in [0, \pi]$ is the latitude.

We will see that parametrizations of geometrical objects become crucial for the development of integration in higher dimensions, e.g., the computation of a flux through a surface.

Remark 1.3.1 Complex functions

In this book we only consider functions with domain $\mathrm{dom}(\boldsymbol{f}) \subseteq \mathbb{R}^n$ and with codomain $\text{co-dom}(\boldsymbol{f}) \subseteq \mathbb{R}^k$. We could without too much effort also allow for complex vector-valued functions, e.g., $\text{co-dom}(\boldsymbol{f}) \subseteq \mathbb{C}^k$. For $k = 1$ such functions are known from [Mathematics 1a] as *complex-valued functions*. Given a complex vector-valued function $f : \mathrm{dom}(\boldsymbol{f}) \to \mathbb{C}^k$, one can for any $\boldsymbol{x} \in \mathbb{R}^n$, write $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{f}_1(\boldsymbol{x}) + i\boldsymbol{f}_2(\boldsymbol{x})$, where $\boldsymbol{f}_1(\boldsymbol{x}) = \mathrm{Re}(\boldsymbol{f}(\boldsymbol{x}))$ is the coordinate-wise real part of the vector $\boldsymbol{f}(\boldsymbol{x})$ and $\boldsymbol{f}_2(\boldsymbol{x}) = \mathrm{Im}(\boldsymbol{f}(\boldsymbol{x}))$ is the coordinate-wise imaginary part of vector $\boldsymbol{f}(\boldsymbol{x})$. In this way, any complex vector-valued function $\boldsymbol{f} : \mathrm{dom}(\boldsymbol{f}) \to \mathbb{C}^k$ gives rise to two real valued-functions $\mathrm{Re}(\boldsymbol{f}) : \mathrm{dom}(\boldsymbol{f}) \to \mathbb{R}^k$ defined as $\boldsymbol{x} \mapsto \mathrm{Re}(\boldsymbol{f}(\boldsymbol{x}))$ and $\mathrm{Im}(\boldsymbol{f}) : \mathrm{dom}(\boldsymbol{f}) \to \mathbb{R}^k$ defined as $\boldsymbol{x} \mapsto \mathrm{Im}(\boldsymbol{f}(\boldsymbol{x}))$. Hence, we can reduce the analysis of a complex vector-valued function to two functions taking values in $\mathbb{R}^k$.

Let us end this section by introducing neural networks, a remarkably flexible and powerful class of functions that has become fundamental to modern machine learning and artificial intelligence.

In applications, we often build complicated functions from simple building blocks through two main operations: summation (linear combinations of functions) and composition. Recall that the composition of two functions

$\boldsymbol{g} \circ \boldsymbol{f}$ is given by $\boldsymbol{x} \mapsto \boldsymbol{g}(\boldsymbol{f}(\boldsymbol{x}))$ whenever co-dom$(\boldsymbol{f}) \subseteq$ dom$(\boldsymbol{g})$. Neural networks exploit this idea of composition systematically. They construct very complicated functions through repeated composition of two simple building blocks: affine transformations $T_{A,\boldsymbol{b}} : \mathbb{R}^n \to \mathbb{R}^k$ given by $T_{A,\boldsymbol{b}}(\boldsymbol{x}) = A\boldsymbol{x} + \boldsymbol{b}$, and non-linear activation functions $\boldsymbol{\sigma} : \mathbb{R}^n \to \mathbb{R}^n$ such as those discussed in Example 1.3.2. By alternating these two operations in multiple layers, neural networks can approximate a vast class of functions, making them extraordinarily powerful tools for learning from data.

Definition 1.3.2 Feed-forward neural network

A *feedforward neural network* or simply a *neural network*[a] is a vector function $\boldsymbol{\Phi} : \mathbb{R}^{n_0} \to \mathbb{R}^{n_L}$ constructed by composing L layers, where each layer $\ell = 1, 2, \ldots, L$ consists of an affine transformation followed by a activation function. More precisely, the network is defined recursively as:

$$\begin{aligned} \boldsymbol{\Phi}(\boldsymbol{x}) &= \boldsymbol{f}_L \circ \boldsymbol{f}_{L-1} \circ \cdots \circ \boldsymbol{f}_1(\boldsymbol{x}), \\ \text{where } \boldsymbol{f}_\ell(\boldsymbol{z}) &= \boldsymbol{\sigma}_\ell(A_\ell \boldsymbol{z} + \boldsymbol{b}_\ell) \end{aligned}$$

for $\ell = 1, \ldots, L$. Here $A_\ell \in \mathsf{M}_{n_\ell \times n_{\ell-1}}(\mathbb{R})$ is the *weight matrix* for layer ℓ, $\boldsymbol{b}_\ell \in \mathbb{R}^{n_\ell}$ is the *bias vector*, and $\boldsymbol{\sigma}_\ell : \mathbb{R}^{n_\ell} \to \mathbb{R}^{n_\ell}$ is the activation function[b]. The dimensions $n_0, n_1, \ldots, n_L \in \mathbb{N}$ are called the *layer sizes*, where n_0 is the input dimension and n_L is the output dimension.

The intermediate vectors $\boldsymbol{z}_\ell = \boldsymbol{f}_\ell \circ \cdots \circ \boldsymbol{f}_1(\boldsymbol{x}) \in \mathbb{R}^{n_\ell}$ for $\ell = 1, \ldots, L-1$ are called the *hidden layer activations*.

[a] In this text, the term neural network will refer to a feedforward neural network, which is the most common architecture. Other types exist, such as recurrent neural networks. A multilayer perceptron (MLP) is a particular kind of feedforward neural network.

[b] At least one activation function $\boldsymbol{\sigma}_\ell$ should be non-linear.

The final activation function $\boldsymbol{\sigma}_L$ is often left out so that $\boldsymbol{f}_L(\boldsymbol{z}) = A_L\boldsymbol{z} + \boldsymbol{b}_L$, i.e., $\boldsymbol{\sigma}_L(\boldsymbol{x}) = \boldsymbol{x}$, or it is specially designed to suit the specific task, e.g., a softmax function for classification tasks (see Example 1.3.5). When $L > 3$, a neural network is called a *deep neural network*, which is the basis for the term *deep learning* in machine learning, while for $L = 2$, the network is called a *shallow neural network*.

The parameters of the neural network are the weight matrices A_ℓ and bias vectors $\boldsymbol{b}_\ell$ for each layer ℓ. Recall that the parameters of the well-known linear function $f(x) = ax + b$ are a and b and that adjusting these parameters changes the behavior of the function. In the neural network, each weight matrix A_ℓ has $n_\ell \cdot n_{\ell-1}$ parameters, and each bias vector $\boldsymbol{b}_\ell$ has n_ℓ parameters.

There are thus a total of

$$N = \sum_{\ell=1}^{L} (n_\ell \cdot n_{\ell-1} + n_\ell) \tag{1.12}$$

parameters that can be adjusted to change the behavior of the function $\boldsymbol{\Phi}$. Sometimes, we write $\boldsymbol{\Phi}_\theta$ to explicitly indicate the dependence of the neural network function on the parameters $\theta = \{A_\ell, \boldsymbol{b}_\ell\}_{\ell=1}^{L}$.

The definition of a (feedforward) neural network functions may look complicated at first, but it is simply a systematic way of building complicated functions by composing affine transformations and non-linear activations, e.g., a three-layer neural network with ReLU activations in the hidden layers and softmax activation in the output layer can be expressed as:

$$\boldsymbol{\Phi}(\boldsymbol{x}) = \operatorname{softmax}(A_3 \operatorname{ReLU}(A_2 \operatorname{ReLU}(A_1 \boldsymbol{x} + \boldsymbol{b}_1) + \boldsymbol{b}_2) + \boldsymbol{b}_3).$$

Example 1.3.4 A shallow neural network

A *shallow neural network* is a neural network with one hidden layer, i.e., $L = 2$. Let the input dimension be $n_0 = 2$, the hidden layer size be $n_1 = 4$, and the output dimension be $n_2 = 1$. The network function $\boldsymbol{\Phi} : \mathbb{R}^2 \to \mathbb{R}$ is then given by:

$$\boldsymbol{\Phi}(\boldsymbol{x}) = A_2 \boldsymbol{\sigma}_1 (A_1 \boldsymbol{x} + \boldsymbol{b}_1) + \boldsymbol{b}_2,$$

where $A_1 \in \mathsf{M}_{4\times 2}(\mathbb{R})$, $\boldsymbol{b}_1 \in \mathbb{R}^4$, $A_2 \in \mathsf{M}_{1\times 4}(\mathbb{R})$, and $\boldsymbol{b}_2 \in \mathbb{R}$. The activation function $\boldsymbol{\sigma}_1 : \mathbb{R}^4 \to \mathbb{R}^4$ could be, for example, the ReLU function applied coordinate-wise. This network has a total of $(4 \cdot 2 + 4) + (1 \cdot 4 + 1) = 17$ parameters.

Let us make specific choices for the parameters:

$$A_1 = \begin{bmatrix} 1 & -1 \\ 0 & 1 \\ -1 & 0 \\ 1 & 1 \end{bmatrix}, \quad \boldsymbol{b}_1 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ -1 \end{bmatrix},$$
$$A_2 = \begin{bmatrix} 1 & 1 & 1 & 1 \end{bmatrix}, \quad \boldsymbol{b}_2 = \begin{bmatrix} 0 \end{bmatrix}.$$

Then, for an input vector $\boldsymbol{x} = [x_1, x_2]^T \in \mathbb{R}^2$, the neural network computes:

$$\boldsymbol{\Phi}(\boldsymbol{x}) = A_2 \boldsymbol{\sigma}_1 \left(\begin{bmatrix} 1 & -1 \\ 0 & 1 \\ -1 & 0 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 0 \\ -1 \end{bmatrix} \right) + \begin{bmatrix} 0 \end{bmatrix}$$

$$
\begin{aligned}
&= A_2 \boldsymbol{\sigma}_1 \left(\begin{bmatrix} x_1 - x_2 \\ x_2 \\ -x_1 \\ x_1 + x_2 - 1 \end{bmatrix} \right) \\
&= A_2 \begin{bmatrix} \max(x_1 - x_2, 0) \\ \max(x_2, 0) \\ \max(-x_1, 0) \\ \max(x_1 + x_2 - 1, 0) \end{bmatrix} \\
&= \max(x_1 - x_2, 0) + \max(x_2, 0) + \max(-x_1, 0) + \max(x_1 + x_2 - 1, 0).
\end{aligned}
$$

Hence, for, e.g., input $\boldsymbol{x} = [2, 0]^T$, the output is:

$$
\begin{aligned}
\boldsymbol{\Phi}(\boldsymbol{x}) &= \max(2 - 0, 0) + \max(0, 0) + \max(-2, 0) + \max(2 + 0 - 1, 0) \\
&= 2 + 0 + 0 + 1 = 3.
\end{aligned}
$$

Let us look at a more advanced example where a deep neural network is used for image classification. We will only describe how such a network is constructed and not go into details about finding all the parameters in the network through training or optimization. The purpose is to illustrate how vector functions can be designed for complex tasks.

Example 1.3.5 A function on hand-written digits

The MNIST database is a collection of 70,000 images of handwritten digits (0-9), each represented as a 28×28 grayscale image, which we represent as points in $\mathbb{R}^{28\times 28}$. We want to build a neural network that can classify these images into the correct digit category. Thus, our function will take an image as input. However, since vector functions are defined on $\mathbb{R}^n$, we need to convert each image from a matrix in $\mathsf{M}_{28\times 28}(\mathbb{R})$ into a vector in $\mathbb{R}^{784}$ (784 is $28 \cdot 28$ so the dimensions match). We do this by selecting a basis for the vector space $\mathsf{M}_{28\times 28}(\mathbb{R})$ and representing each image as a coordinate vector with respect to that basis. A simple approach is to "flatten" each 28×28 image into a vector of length 784 by stacking its columns (or rows) in sequence[a]. Hence, each image $X \in \mathbb{R}^{28\times 28}$ is transformed into a vector $\boldsymbol{x} \in \mathbb{R}^{784}$.

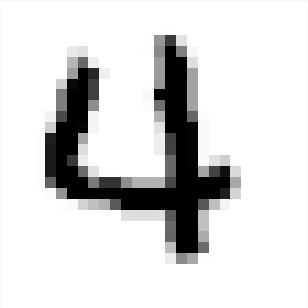


Figure 1.2: Sample image from the NIST Special Database 19 of handwritten digits https://www.nist.gov/srd/nist-special-database-19. The image shown is a 28×28 grayscale image, which can be flattened into a vector in $\mathbb{R}^{784}$.

A neural network for classifying these digits can be constructed as a function

$$\boldsymbol{\Phi} : \mathbb{R}^{784} \to \mathbb{R}^{10},$$

where the *output* vector $\boldsymbol{\Phi}(\boldsymbol{x}) = [p_0, p_1, \ldots, p_9]^T$ represents the predicted probabilities that the input image $\boldsymbol{x}$ corresponds to each digit 0–9. The digit with the highest probability is taken as the network's prediction. For example, the output $\boldsymbol{\Phi}(\boldsymbol{x}) = [0.1, 0.02, 0.05, 0.05, 0.7, 0.03, 0.02, 0.01, 0.01, 0.01]^T$ indicates a 70% probability that the image $\boldsymbol{x}$ is the digit 4. This is a desirable output for the input image shown in Figure 1.2, as it indeed depicts the digit 4.

The network is typically trained on a labeled dataset of images, where the true digit for each image is known. During training, the parameters of the neural network (weights and biases) are adjusted to minimize the difference between the predicted probabilities and the true labels.

An architecture might use $L = 3$ layers with dimensions $n_0 = 784$, $n_1 = 128$, $n_2 = 64$, and $n_3 = 10$. Here, $n_0 = 784$ and $n_3 = 10$ are fixed by the input and output dimensions, respectively, while n_1 and n_2 are numbers we can choose. The network function is given by:

$$\boldsymbol{\Phi}(\boldsymbol{x}) = \operatorname{softmax}(A_3\boldsymbol{\sigma}_2(A_2\boldsymbol{\sigma}_1(A_1\boldsymbol{x} + \boldsymbol{b}_1) + \boldsymbol{b}_2) + \boldsymbol{b}_3),$$

where $A_1 \in \mathsf{M}_{128\times 784}(\mathbb{R})$, $\boldsymbol{b}_1 \in \mathbb{R}^{128}$, $A_2 \in \mathsf{M}_{64\times 128}(\mathbb{R})$, $\boldsymbol{b}_2 \in \mathbb{R}^{64}$, $A_3 \in \mathsf{M}_{10\times 64}(\mathbb{R})$, and $\boldsymbol{b}_3 \in \mathbb{R}^{10}$.

The first two layers might use ReLU activation functions ($\boldsymbol{\sigma}_1 = \boldsymbol{\sigma}_2 =$ ReLU), while the final layer uses the softmax activation function ($\boldsymbol{\sigma}_3 =$ softmax) to ensure the outputs can be interpreted as probabilities. Such a network would have $784 \cdot 128 + 128 + 128 \cdot 64 + 64 + 64 \cdot 10 + 10 = 109\,386$ parameters (weights and biases) that need to be determined during training.

[a]This flattening corresponds to choosing the standard matrix basis E_{ij} for $\mathsf{M}_{28\times 28}(\mathbb{R})$, see Example 2.1.8.

A word on variables and parameters: In the context of functions, *variables* refer to the inputs of the function, which can change and are often denoted by symbols like x, y, or $\boldsymbol{x}$. In contrast, *parameters* are fixed values that define the specific behavior of the function but do not change during its evaluation. For example, in a neural network, the weights and biases are parameters that are learned during training, while the input data points (like an image) are the variables. Sometimes, however, the role of parameters and variables can be interchanged as it is the case during the training of a neural network, where the parameters are considered as variables of a so-called loss function that are optimized using a training algorithm (see Example 3.7.2 on page 111).

1.4 Visualizing functions

It is often helpful to have a certain "geometric understanding" of a function $\boldsymbol{f}\colon \operatorname{dom}(\boldsymbol{f}) \to \mathbb{R}^k$ where $\operatorname{dom}(\boldsymbol{f}) \subseteq \mathbb{R}^n$. In Figure 1.3 we see the "input vectors" of the function $\boldsymbol{f}$ on the left and the "output vectors" on the right.

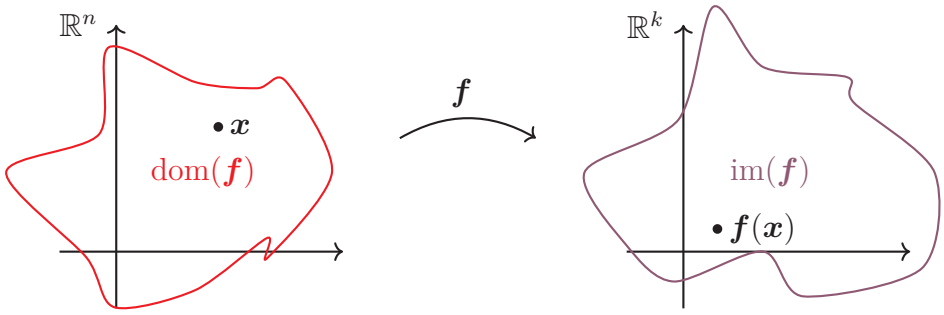


Figure 1.3: Input-output view of a vector function of n variables. The function $\boldsymbol{f}$ takes $\boldsymbol{x} \in \mathbb{R}^n$ as input and outputs $\boldsymbol{f}(\boldsymbol{x}) \in \mathbb{R}^k$. We have plotted both $\mathbb{R}^n$ and $\mathbb{R}^k$ as $\mathbb{R}^2$, but we should allow ourselves to mentally think of n and k taking on larger values.

In general, it is not possible to visualize functions when k and n are large. However, for small values of k and n it is useful to study the graph and the level sets of a function.

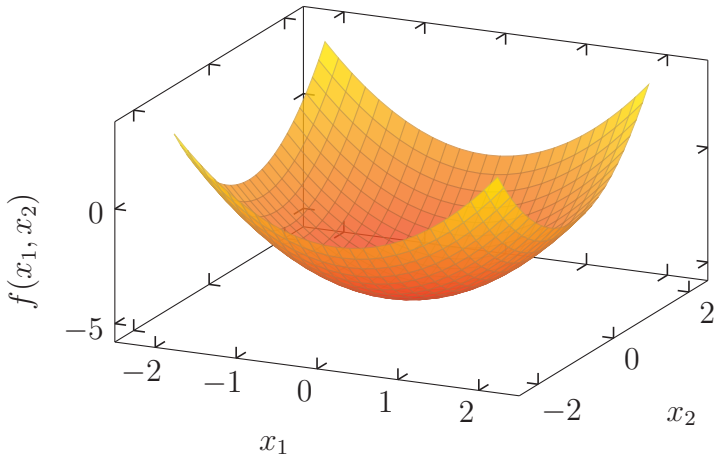


Figure 1.4: The graph of the function $f(x_1, x_2) = x_1^2 + x_2^2 - 5$ with domain $\mathrm{dom}(f) = [-3, 3] \times [-3, 3]$.

Graph of a function. From standard calculus we are used to extract information on a function $f : \mathbb{R} \to \mathbb{R}$ by drawing or plotting its graph. Formally, the *graph* of a scalar function of one variable consists of all points in $\mathbb{R}^2$ of the form $(x, f(x))$, where x belongs to the domain $\mathrm{dom}(f)$ of f.

Similarly, for a vector function $\boldsymbol{f} : \mathrm{dom}(\boldsymbol{f}) \to \mathbb{R}^k$, where $\mathrm{dom}(\boldsymbol{f})$ is a subset of $\mathbb{R}^n$, the *graph* consists of all points in $\mathbb{R}^n \times \mathbb{R}^k = \mathbb{R}^{n+k}$ of the form $(\boldsymbol{x}, \boldsymbol{f}(\boldsymbol{x}))$, where $\boldsymbol{x} \in \mathrm{dom}(\boldsymbol{f})$.

Definition 1.4.1 Graph of a function

Consider a vector function $\boldsymbol{f} : \mathrm{dom}(\boldsymbol{f}) \to \mathbb{R}^k$, where $\mathrm{dom}(\boldsymbol{f})) \subseteq \mathbb{R}^n$. The *graph* of the function is the subset of $\mathbb{R}^{n+k} = \mathbb{R}^n \times \mathbb{R}^k$ given by

$$\{\, (\boldsymbol{x}, \boldsymbol{f}(\boldsymbol{x})) \in \mathbb{R}^{n+k} \mid \boldsymbol{x} \in \mathrm{dom}(\boldsymbol{f}) \,\}.$$

We can only illustrate such a set graphically for the cases where $n+k \leq 3$. Hence, it is a very limited tool for visualizing and understanding functions of several variables unless n and k are small. Examples of such functions are *scalar* functions of *two* variables, i.e., $n = 2$ and $k = 1$ in Definition 1.3.1.

Example 1.4.1

We consider the scalar function of two variables:

$$f : [-3, 3] \times [-3, 3] \to \mathbb{R}, \;\; f(x_1, x_2) = x_1^2 + x_2^2 - 5.$$

The graph of this function is given by

$$\{\,(x_1, x_2, x_1^2 + x_2^2 - 5) \mid (x_1, x_2) \in [-3, 3] \times [-3, 3]\,\} \subset \mathbb{R}^3$$

See Figure 1.4 for an illustration of the graph.

Exercise 1.4.2

Consider the shallow neural network $\mathbf{\Phi} : \mathbb{R}^2 \to \mathbb{R}$ from Example 1.3.4.

(a) Write down the expression for the graph of the function $\mathbf{\Phi}$.

(b) Use a computer to plot the graph of the function $\mathbf{\Phi}$ over the domain $[-2, 2] \times [-2, 2]$.

(c) Try varying the number of neurons in the hidden layer (i.e., change n_1) and see how this affects the shape of the graph.

(d) The graph is pieced together from a number of flat (linear) surfaces. How does the number of these flat surfaces relate to the number of neurons n_1 in the hidden layer?

Visualizing vector fields Vector fields $\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^n$ $(n = 2, 3)$ can be visualized by plotting the "output" vectors on a regular grid corresponding to the coordinates of the "input" vectors. Similar to the graph of a function, this idea mixes the domain and codomain of a function into one object. Let us show the idea by an example.

Example 1.4.3

Consider the function

$$\boldsymbol{V} : \mathbb{R}^2 \to \mathbb{R}^2, \quad \boldsymbol{V}(x_1, x_2) = (-x_2/3, x_1/3).$$

This is an example of a *vector-field*, which maps a point $\boldsymbol{x} \in \mathbb{R}^2$ to a point in $\boldsymbol{V}(\boldsymbol{x}) \in \mathbb{R}^2$. The new point $\boldsymbol{V}(\boldsymbol{x})$ is then interpreted as a vector originating from the point $\boldsymbol{x}$. This is illustrated in Figure 1.5. E.g., the vector-field $\boldsymbol{V}$ evaluated at $(x_1, x_2) = (0, 2)$ is the vector $(0, 2/3)$, whereas at $(x_1, x_2) = (1, 4)$, it is $(-4/3, 1/3)$.

Level sets of a function. It can be difficult to get a good impression of a function $f : \mathbb{R}^2 \to \mathbb{R}$ from its graph. In such cases it is often useful to

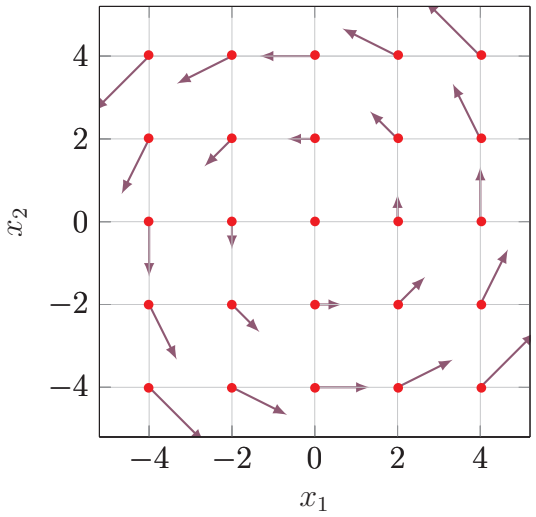


Figure 1.5: Illustration of the function $\boldsymbol{V}(x_1, x_2) = (-x_2/3, x_1/3)$ on $[-4, 4] \times [-4, 4]$.

consider the so-called *level sets*. They consist of points in the domain that are mapped to a pre-defined fixed vector:

Definition 1.4.2 Level set

Consider a vector function $\boldsymbol{f} : \mathrm{dom}(\boldsymbol{f}) \to \mathbb{R}^k$, where $\mathrm{dom}(\boldsymbol{f})) \subseteq \mathbb{R}^n$. Given a vector $\boldsymbol{c} \in \mathbb{R}^k$, the corresponding level set is the subset of $\mathbb{R}^n$ given by

$$\{\, \boldsymbol{x} \in \mathrm{dom}(\boldsymbol{f}) \mid \boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{c} \,\}.$$

Example 1.4.4

Consider the scalar function f

$$f : \mathbb{R}^2 \to \mathbb{R}, \;\; f(x, y) = x_1^2 + x_2^2 + 3.$$

For $c \in \mathbb{R}$ the level sets are given by

$$\{\, (x_1, x_2) \in \mathrm{dom}(f) \mid f(x_1, x_2) = c \,\}.$$

Since $x_1^2 + x_2^2 \geq 0$ for all (x, y), the level set associated with any $c < 3$ is empty. For $c = 3$, the level set consists of the point $(x_1, x_2) = (0, 0)$, and for $c > 3$, it consists of the $(x_1, x_2) \in \mathbb{R}^2$ for which $x_1^2 + x_2^2 = c - 3$. Thus, for $c > 3$, the level set is a circle centered at $(0, 0)$ with radius $\sqrt{c - 3}$.

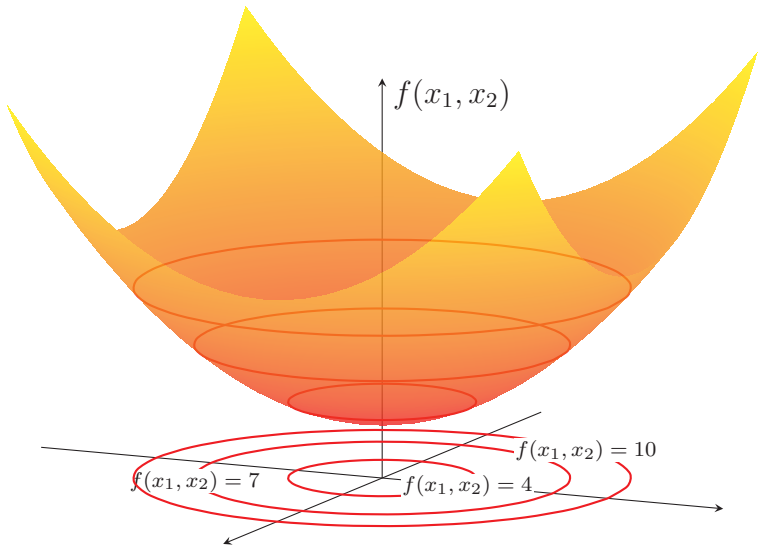


Figure 1.6: The graph of the function $f(x_1, x_2) = x_1^2 + x_2^2 + 3$ and some of its level sets (only shown for ; $(x, y) \in [-3, 3] \times [-3, 3]$.)

Level sets are particularly useful if they correspond to "recognizable" geometric shapes as in Example 1.4.4. In the following example we remind the reader of the mathematical equations that characterize some of the most well-known geometric shapes.

Example 1.4.5

Fix a point $(c_1, c_2) \in \mathbb{R}^2$.

(a) (circle) A circle in $\mathbb{R}^2$ centered at the point (c_1, c_2) and with radius r consists of precisely the $(x, y) \in \mathbb{R}^2$ for which

$$(x - c_1)^2 + (y - c_2)^2 = r^2. \tag{1.13}$$

(b) (ellipsoid) Given $a, b > 0$, the set of points $(x, y) \in \mathbb{R}^2$ for which

$$\frac{(x - c_1)^2}{a^2} + \frac{(y - c_2)^2}{b^2} = 1, \tag{1.14}$$

forms an ellipsoid centered at (c_1, c_2). The ellipsoid has symmetry axes $x = c_1, y = c_2$ and half-axes $a > 0$, $b > 0$.

(c) (hyperbola) Given $a, b > 0$, the set of points $(x, y) \in \mathbb{R}^2$ for which

$$\frac{(x - c_1)^2}{a^2} - \frac{(y - c_2)^2}{b^2} = 1, \tag{1.15}$$

forms a hyperbola centered at (c_1, c_2). The hyperbola has symmetry axes $x = c_1, y = c_2$. Similarly, a hyperbola with symmetry axes $y = c_1, x = c_2$, is given by the equation

$$\frac{(y-c_2)^2}{b^2} - \frac{(x-c_1)^2}{a^2} = 1. \tag{1.16}$$

Example 1.4.6 Parabola

The graph of a second-order polynomial function of one variable is of the form

$$y = ax^2 + bx + c.$$

Hence, the graph forms a parabola with symmetry axis $x = -\frac{b}{2a}$.

Note that we can also consider higher-dimensional versions of the sets in Example 1.4.5.

Example 1.4.7 Sphere

Fixing a point $(c_1, c_2, c_3) \in \mathbb{R}^3$, the sphere in $\mathbb{R}^3$ centered at the point (c_1, c_2, c_3) and with radius $r > 0$ is given by

$$(x-c_1)^2 + (y-c_2)^2 + (z-c_3)^2 = r^2. \tag{1.17}$$

In the rest of this section we will consider the level sets for quadratic forms, see Definition 1.2.1, of a particular simple form, namely,

$$q : \mathbb{R}^2 \to \mathbb{R},\ q(x, y) = \lambda_1 x^2 + \lambda_2 y^2 + b_1 x + b_2 y + c, \tag{1.18}$$

where $\lambda_1, \lambda_2, b_1, b_2, c \in \mathbb{R}$. This function is indeed a quadratic form according to (1.3) with the choices

$$A = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}, \quad \text{and} \quad \boldsymbol{b} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}.$$

For such functions we are able to give explicit descriptions of the level sets in terms of the geometric shapes discussed in Examples 1.4.5 and 1.4.6. Note that the expression for the analyzed function q in (1.18) does not involve the product xy since A was assumed to be a diagonal matrix. The case where the product xy appears is significantly more complicated and will be handled in Example 2.8.6 on page 74.

Example 1.4.8

For the function q in (1.18), the level set associated with a given $k \in \mathbb{R}$ consists precisely of the $(x, y) \in \mathbb{R}^2$ for which

$$\lambda_1 x^2 + \lambda_2 y^2 + b_1 x + b_2 y + c = k,$$

or,

$$\lambda_1 x^2 + \lambda_2 y^2 + b_1 x + b_2 y = k - c. \tag{1.19}$$

For some (not all) values of the parameters $\lambda_1, \lambda_2, b_1, b_2, c$ we will now describe the geometric shape of the level set.

(a) If $\lambda_1 = \lambda_2 = 0$ and $b_2 \neq 0$, the equation (1.19) takes the form

$$b_1 x + b_2 y = k - c,$$

or

$$y = -\frac{b_1}{b_2} x + \frac{k - c}{b_2}.$$

Thus, the level set is a straight line with slope $-b_1/b_2$.

(b) If $\lambda_2 = 0$ and $\lambda_1 \neq 0, b_2 \neq 0$, (1.19) can be written as

$$y = -\frac{\lambda_1}{b_2} x^2 - \frac{b_1}{b_2} x + \frac{k - c}{b_2}.$$

Thus, the level set is a parabola. Similarly, the case $\lambda_1 = 0, \lambda_2 \neq 0, b_1 \neq 0$ yields a parabola.

(c) If $\lambda_1 \neq 0, \lambda_2 \neq 0$, the level set can be identified as one of the sets in Example 1.4.5. The technique to do so is always the same; let us illustrate it on the equation

$$4x^2 + y^2 + 8x - 6y - 3 = 0. \tag{1.20}$$

Let us rewrite (1.20) as

$$4(x^2 + 2x) + y^2 - 6y = 3. \tag{1.21}$$

We will now consider the terms involving x and y separately. Thus, we first look at the expression $x^2 + 2x$. Notice that all the definitions in Example 1.4.5 involve terms of the form

$$(x - c_1)^2 = x^2 - 2c_1 x + c_1^2$$

for some $c_1 \in \mathbb{R}$. In order to rewrite $x^2 + 2x$ such that a term of this form occur, we see that it is necessary that we choose c_1 such that $-2c_1 = 2$, i.e., $c_1 = -1$. With this choice of c_1, we have that

$$(x - c_1)^2 = (x + 1)^2 = x^2 + 2x + 1 = (x^2 + 2x) + 1;$$

therefore,

$$x^2 + 2x = (x + 1)^2 - 1. \tag{1.22}$$

Thus, we see that we do not obtain that the term $x^2 + 2x$ can be written exactly as $(x - c_1)^2$, but we will see what to do with the occurring number -1 soon. Before doing so, we will now run the same procedure on the term $y^2 - 6y$. Exactly as before, we see that in order to introduce a term

$$(y - c_2)^2 = y^2 - 2c_2y + c_2^2,$$

it is necessary to take c_2 such that $-2c_2 = -6$, i.e., $c_2 = 3$. With this choice,

$$(y - c_2)^2 = (y - 3)^2 = y^2 - 6y + 9 = (y^2 + 6y) + 9,$$

i.e.,

$$y^2 - 6y = (y - 3)^2 - 9, \tag{1.23}$$

Inserting the expressions (1.22) and (1.23) in (1.21) yields that

$$4\Big((x + 1)^2 - 1\Big) + (y - 3)^2 - 9 = 3,$$

or,

$$4(x + 1)^2 + (y - 3)^2 = 16.$$

Finally, this equation can be rewritten as

$$\frac{(x + 1)^2}{4} + \frac{(y - 3)^2}{16} = 1;$$

comparing with the characterizations in Example 1.4.5 shows that the set is an ellipsoid centered at the point $(-1, 3)$, with symmetry axes $x = -1, y = 3$ and half-axes $a = 2$, $b = 4$.

1.5 What is multivariate calculus useful for?

Understanding multivariate calculus is crucial for making sense of and solving many problems in the modern world. The question posed in the title is very broad and cannot be fully answered in this short section, but we here aim to give a flavor of its importance. One striking example is its use in the training of artificial intelligence models, see Example 3.7.2 on page 111. During the training process, the model adjusts its parameters (i.e., the weights and biases) to minimize a scalar overall loss function that quantifies how far its predictions are from the true or desired values. This process involves calculating the derivative of the loss function with respect to each parameter. These derivatives form a vector that points in the direction of the steepest increase of the loss function (see Example 3.4.2 on page 99). To minimize the loss, the parameters are adjusted in the opposite direction. Since modern networks often have millions or even billions of parameters, the derivatives are computed in very high dimensional spaces.

But what does it mean to take the derivative of a function with several (or billions of!) variables? The derivative quantifies the sensitivity of change of a function's output with respect to its input. From standard calculus we know that the derivative of a scalar function $f : \mathbb{R} \to \mathbb{R}$ at a point $x = x_0$ is, whenever it exists, the slope of the tangent line to the graph of the function at $(x_0, f(x_0))$. It is not at all clear what the derivative of a vector function of several variables should be. In fact, as we will see in Chapter 3, a proper understanding of differentiability of functions of several variables truly requires a multi-dimensional analysis of the problem. To give an example, the derivative of the linear mapping L_A in Example 1.3.1 is A, that is, the derivative is a $k \times n$ matrix, and that the derivative of the quadratic form considered in Definition 1.2.1 is $(A + A^T)\boldsymbol{x} + \boldsymbol{b}$ which is a vector in $\mathbb{R}^n$ and, considered as a function of $\boldsymbol{x}$, the derivative is an affine function from $\mathbb{R}^n$ to $\mathbb{R}^n$.

For the loss function, how can the derivatives be efficiently computed and interpreted to update the model parameters? Moreover, how can we ensure that the loss function even is differentiable and that the derivatives are well-defined? Multivariate calculus provides the mathematical toolbox to answer these questions, making it a cornerstone of modern techniques in, e.g., machine learning and optimization.

CHAPTER 2

Inner Product Spaces and the Spectral Theorem

In this chapter we will develop the theory of linear algebra that will be needed in later chapters. Continuing from the textbook [Mathematics 1a], we will introduce new central aspects of linear algebra that has its origins from plane geometry, e.g., lengths and angles in the two-dimensional real plane. This leads to the development of axioms for inner products and norms, extending the concepts of length and perpendicularity (or orthogonality as we will say) to abstract vector spaces. From the concept of a norm, we will be able to introduce open and closed sets in a vector space, Section 2.2. That section provides the necessary background for studying continuity and differentiability of functions of several real variables.

The main result of the chapter is the *spectral theorem* that shows us that symmetric matrices have a basis of eigenvectors that are orthogonal to each other. The name *spectral theory* was introduced by David Hilbert, and it concerns theories extending the eigenvalue problem of a single square matrix to classes of matrices and linear transformations. The set of eigenvalues of a square matrix is called the *spectrum* and the matrix classes we will study are real symmetric, Hermitian and normal matrices.

Abstract vectors spaces can be considered over arbitrary scalar fields $\mathbb{F}$, however, we will restrict our attention to either real or complex vector spaces. That is, we will always assume that $\mathbb{F} = \mathbb{R}$ or $\mathbb{F} = \mathbb{C}$. We are mainly interested in the real case, however, as even real matrices can have a complex spectrum (i.e., complex eigenvalues), it is natural to study spectral theory in the setting of both real and complex vector spaces. We will sometimes write formulas both for real and complex case, but we will in some cases only state the complex formulas as they also apply to the real case.

While we consider abstract vector spaces, we will focus on the space of

real column vectors of size $n \times 1$, that is, the vector space $\mathbb{R}^n$ over the reals $\mathbb{R}$, as well as the space of complex column vectors of size $n \times 1$, that is, the vector space $\mathbb{C}^n$ over the field $\mathbb{C}$ of complex scalars.

2.1 Inner product spaces

Inner product and norm in $\mathbb{R}^n$. In dimensions two and three, the length of a vector, i.e., the distance from the origin to its endpoint, can be computed by the Pythagorean rule. For example in $\mathbb{R}^2$, the length of the vector $\boldsymbol{x} = [x_1, x_2]^T$ is given by

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

It is natural to generalize this formula for all $n \in \mathbb{N}$, and we define the length of the vector $\boldsymbol{x} \in \mathbb{R}^n$ as

$$\|\boldsymbol{x}\| = \sqrt{x_1^2 + x_2^2 + \ldots + x_n^2}. \tag{2.1}$$

Instead of the word "length" we will say that the non-negative real number $\|\boldsymbol{x}\|$ is the *norm* of $\boldsymbol{x}$.

The dot product of two vectors $\boldsymbol{x} = [x_1, x_2, \ldots, x_n]^T$, $\boldsymbol{y} = [y_1, y_2, \ldots, y_n]^T$ in $\mathbb{R}^n$ is defined as

$$\boldsymbol{x} \cdot \boldsymbol{y} = x_1 y_1 + x_2 y_2 + \ldots + x_n y_n. \tag{2.2}$$

In dimensions two and three, this definition reduces to the well-known formula, e.g., the dot product in $\mathbb{R}^2$ is $\boldsymbol{x} \cdot \boldsymbol{y} = x_1 y_1 + x_2 y_2$, where $\boldsymbol{x} = [x_1, x_2]^T$ and $\boldsymbol{y} = [y_1, y_2]^T$. While the notation $\boldsymbol{x} \cdot \boldsymbol{y}$ and the term "dot product" is often used for the inner product in (2.2), we often prefer the notation $\langle \boldsymbol{x}, \boldsymbol{y} \rangle$ and call this the *standard inner product* on $\mathbb{R}^n$.

The norm of $\boldsymbol{x}$ can be computed from the dot product of $\boldsymbol{x}$ with itself:

$$\|\boldsymbol{x}\| = \sqrt{\boldsymbol{x} \cdot \boldsymbol{x}}. \tag{2.3}$$

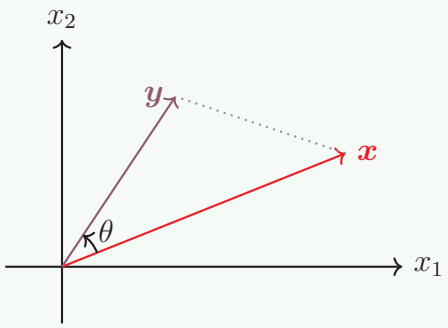
With the dot product and the norm the vector space $\mathbb{R}^n$ is sometimes said to have euclidean structure or geometry.

Using the notion of matrix transpose, one can write the standard inner product in $\mathbb{R}^n$ by $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \boldsymbol{y}^T \boldsymbol{x}$. Note that this is a slight abuse of notation because $\langle \boldsymbol{x}, \boldsymbol{y} \rangle$ is a real scalar while $\boldsymbol{x}^T \boldsymbol{y}$ is a 1×1 matrix as it is the matrix-matrix multiplication of a $1 \times n$ and $n \times 1$ matrix, respectively. Note also that $\boldsymbol{y}^T \boldsymbol{x} = \boldsymbol{x}^T \boldsymbol{y}$, and we use the notation $\boldsymbol{y}^T \boldsymbol{x}$ only to be consistent with the formula in the complex case.

The angle between two vectors is closely related to the norm and inner product as the following example shows.

Example 2.1.1

Consider the triangle below, whose vertices are given by the origin $\mathbf{0} = [0, 0]^T$ and the real, non-zero vectors $\boldsymbol{x} = [x_1, x_2]^T$ and $\boldsymbol{y} = [y_1, y_2]^T$.



Let θ denote the angle between the vectors $\boldsymbol{x}$ and $\boldsymbol{y}$. One can show (e.g., using the law of cosines) that

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \|\boldsymbol{x}\| \|\boldsymbol{y}\| \cos\theta.$$

If $\theta = \pm 90°$, we say that $\boldsymbol{x}$ and $\boldsymbol{y}$ are orthogonal to each other. Note that $\boldsymbol{x}$ and $\boldsymbol{y}$ are *orthogonal* if and only if $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = 0$.

From the formula

$$\cos\theta = \frac{\langle \boldsymbol{x}, \boldsymbol{y} \rangle}{\|\boldsymbol{x}\| \|\boldsymbol{y}\|}$$

we see that the angle θ between two vectors can be found from the values of inner product $\langle \boldsymbol{x}, \boldsymbol{y} \rangle$ and the norms $\|\boldsymbol{x}\|$ and $\|\boldsymbol{y}\|$.

Inner product and norm in $\mathbb{C}^n$. Let us now define norm and inner product for the vector space $\mathbb{C}^n$ over the scalar field $\mathbb{C}$. We cannot directly use (2.3) as for $\mathbb{R}^n$ since the expression inside the square root might be complex if $\boldsymbol{x}$ is a complex vector; recall that x^2 is rarely a real number if x is complex (in fact, x^2 is real only if x is real or if x is purely imaginary). However, for a complex number $x = a + ib$, $a, b \in \mathbb{R}$, we have $|x|^2 = x\overline{x} = a^2 + b^2 \geq 0$. Hence, if $\boldsymbol{x} \in \mathbb{C}^n$ is given by

$$\boldsymbol{x} = \begin{bmatrix} a_1 + ib_1 \\ a_2 + ib_2 \\ \vdots \\ a_n + ib_n \end{bmatrix},$$

it is natural to define its norm $\|\boldsymbol{x}\|$ by

$$\|\boldsymbol{x}\| = \sqrt{\sum_{k=1}^{n}(a_k^2 + b_k^2)} = \sqrt{\sum_{k=1}^{n}|x_k|^2}. \tag{2.4}$$

This definition guarantees that $\|\boldsymbol{x}\|$ is a *non-negative* real number which is what we would expect from a quantity that is the "length" of a vector $\boldsymbol{x} \in \mathbb{C}^n$.

Since we want to keep the relation $\|\boldsymbol{x}\|^2 = \langle \boldsymbol{x}, \boldsymbol{x}\rangle$ as for $\mathbb{R}^n$, we can define the *standard inner product on* $\mathbb{C}^n$ by

$$\langle \boldsymbol{x}, \boldsymbol{y}\rangle = x_1\overline{y_1} + x_2\overline{y_2} + \ldots + x_n\overline{y_n} = \sum_{k=1}^{n} x_k\overline{y}_k;$$

it follows that $\langle \boldsymbol{x}, \boldsymbol{x}\rangle \geq 0$ since $x_k\overline{x_k} = |x_k|^2 \geq 0$ for all $k = 1, \ldots, n$.

Using the notion of matrix adjoints, one can write the standard inner product in $\mathbb{C}^n$ by $\langle \boldsymbol{x}, \boldsymbol{y}\rangle = \boldsymbol{y}^*\boldsymbol{x}$. Again, this is an abuse of notation (that we will ignore) because $\langle \boldsymbol{x}, \boldsymbol{y}\rangle$ is a complex scalar while $\boldsymbol{y}^*\boldsymbol{x}$ is a 1×1 matrix.

Inner products and norms in abstract spaces. Let V be a vector space over $\mathbb{F}$. If $\mathbb{F} = \mathbb{R}$, we say that V is a real vector space, and if $\mathbb{F} = \mathbb{C}$, we say that V is a complex vector space.

A norm on a vector space V is a function that assigns to each vector $\boldsymbol{x} \in V$ a non-negative real number $\|\boldsymbol{x}\| \in [0, \infty[$ that we call the norm of $\boldsymbol{x}$ and consider as the length of $\boldsymbol{x}$. The norm function has to satisfy the following properties:

Definition 2.1.1 Norm

Let V be a vector space over $\mathbb{F} = \mathbb{R}$ or $\mathbb{C}$. A *norm* on V is a function

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

that satisfies for any $\boldsymbol{x}, \boldsymbol{y} \in V$ and $c \in \mathbb{F}$ the following properties:

(i) non-negativity: $\|\boldsymbol{x}\|$ is (real and) non-negative,

(ii) non-degeneracy: $\|\boldsymbol{x}\| = 0$ if and only if $\boldsymbol{x} = 0$,

(iii) scaling: $\|c\boldsymbol{x}\| = |c|\,\|\boldsymbol{x}\|$,

(iv) triangle inequality: $\|\boldsymbol{x} + \boldsymbol{y}\| \leq \|\boldsymbol{x}\| + \|\boldsymbol{y}\|$.

A vector space V with a norm is called a *normed vector space* or simply a *normed space*.

The norm is a function that behaves like the length of vectors in the plane $\mathbb{R}^2$, e.g., the length of (-5) times the vector is 5 times the length of the original vector. This last property is a special case of Item (iii) in the definition above.

Example 2.1.2 *p-norms on $\mathbb{C}^n$*

Consider $V = \mathbb{C}^n$ as a vector space over $\mathbb{C}$. For $p \in [1, \infty]$, the *p-norm* of a vector $\boldsymbol{x} = [x_1, x_2, \ldots, x_n]^T$ is defined by

$$\|\boldsymbol{x}\|_p = \left(\sum_{k=1}^{n} |x_k|^p \right)^{1/p} \qquad \text{for } p \in [1, \infty[$$

and

$$\|\boldsymbol{x}\|_\infty = \max_{k=1,\ldots,n} |x_k|.$$

The p-norms are norms on $\mathbb{C}^n$ for $p \in [1, \infty]$. Since these norms are different from each other, we have added a subscript in $\|\boldsymbol{x}\|_p$ to distinguish them. The *2-norm* $\|\boldsymbol{x}\|_2$ is the standard norm on $\mathbb{C}^n$ that we considered above in Equation (2.4). The *1-norm* $\|\boldsymbol{x}\|_1$ is the sum of the absolute values of the entries of $\boldsymbol{x}$, and the *∞-norm* $\|\boldsymbol{x}\|_\infty$ is the maximum absolute value of the entries of $\boldsymbol{x}$. The p-norms are used in many applications, e.g., in statistics and signal processing and are often called ℓ^p-norms.

An inner product is a function that assigns to each *pair* of vectors $(\boldsymbol{x}, \boldsymbol{y})$ a scalar in $\mathbb{F}$, denoted by $\langle \boldsymbol{x}, \boldsymbol{y} \rangle$. The inner product function $(\boldsymbol{x}, \boldsymbol{y}) \mapsto \langle \boldsymbol{x}, \boldsymbol{y} \rangle$ has to satisfy the following properties:

Definition 2.1.2 Inner product

Let $\mathbb{F} = \mathbb{R}$ or $\mathbb{C}$. An *inner product* on V is a function

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

that satisfies for any $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in V$ and $c, d \in \mathbb{F}$ the following properties:

(i) non-negativity: $\langle \boldsymbol{x}, \boldsymbol{x} \rangle \geq 0$,

(ii) non-degeneracy: $\langle \boldsymbol{x}, \boldsymbol{x} \rangle = 0$ if and only if $\boldsymbol{x} = 0$,

(iii) (conjugate) symmetry: $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \overline{\langle \boldsymbol{y}, \boldsymbol{x} \rangle}$; note, that for a real vector space, this property is just symmetry, $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \langle \boldsymbol{y}, \boldsymbol{x} \rangle$,

(iv) linearity: $\langle c\boldsymbol{x} + d\boldsymbol{y}, \boldsymbol{z} \rangle = c\langle \boldsymbol{x}, \boldsymbol{z} \rangle + d\langle \boldsymbol{y}, \boldsymbol{z} \rangle$.

A vector space V with an inner product is called an *inner product space*.

It is important to remember that the quantity $\langle \boldsymbol{x}, \boldsymbol{y} \rangle$ is always just a scalar: it is a real number for real vector spaces and a (possible) complex number for complex vector spaces. For complex vector space the inner product is conjugate linear in the second entry (and linear in the first as required in Item (iv)). Conjugate linearity means that

$$\langle \boldsymbol{x}, c\boldsymbol{y} + d\boldsymbol{z} \rangle = \overline{c}\langle \boldsymbol{x}, \boldsymbol{y} \rangle + \overline{d}\langle \boldsymbol{x}, \boldsymbol{z} \rangle$$

which follows from (iii) and (iv) in Definition 2.1.2. Loosely speaking, we are allowed to move scalars "out" from inside the inner product, but we have to remember to do a complex conjugation if the scalar comes from the second entry of the inner product. Finally, we mention that the inner products with the zero vector are always zero, i.e.,

$$\langle \boldsymbol{0}, \boldsymbol{y} \rangle = 0 \text{ and } \langle \boldsymbol{x}, \boldsymbol{0} \rangle = 0.$$

To see this, we use linearity of the inner product to compute $\langle \boldsymbol{0}, \boldsymbol{y} \rangle = \langle 0\boldsymbol{x}, \boldsymbol{y} \rangle = 0 \cdot \langle \boldsymbol{x}, \boldsymbol{y} \rangle = 0$.

Given an inner product space V, we can always construct a norm on V. From an inner product $\langle \cdot, \cdot \rangle$ on V, we define the *derived norm* on V by

$$\| \cdot \| : V \to [0, \infty[, \quad \|\boldsymbol{x}\| = \sqrt{\langle \boldsymbol{x}, \boldsymbol{x} \rangle} \tag{2.5}$$

Note that by the non-negativity property of the inner product $\langle \boldsymbol{x}, \boldsymbol{x} \rangle \geq 0$, the expression inside the square root in (2.5) satisfies $\langle \boldsymbol{x}, \boldsymbol{x} \rangle \geq 0$ hence $\|\boldsymbol{x}\| \geq 0$.

Theorem 2.1.1 Induced or Derived Norm

Suppose V is an inner product space over $\mathbb{F}$ with inner product $\langle \cdot, \cdot \rangle$. Then the derived norm $\| \cdot \|$ given by (2.5) is indeed a norm.

Proof. (i): Non-negativity follows from non-negativity of the inner product in Definition 2.1.2.

(ii): The equation $\|\boldsymbol{x}\| = 0$ holds if and only if $\langle \boldsymbol{x}, \boldsymbol{x} \rangle = 0$ holds which, by Definition 2.1.2, holds if and only if $\boldsymbol{x} = 0$.

(iii): The scaling law follows from (conjugate) linearity and conjugate symmetry of the inner product:

$$\|c\boldsymbol{x}\| = \langle c\boldsymbol{x}, c\boldsymbol{x} \rangle^{1/2} = \left(c\overline{c}\,\langle \boldsymbol{x}, \boldsymbol{x} \rangle\right)^{1/2} = \left(|c|^2 \langle \boldsymbol{x}, \boldsymbol{x} \rangle\right)^{1/2} = |c|\,\|\boldsymbol{x}\|.$$

(iv): The inequality is illustrated in Figure 2.1. We postpone the proof of the triangle inequality to Theorem 2.1.5 on page 48.

■

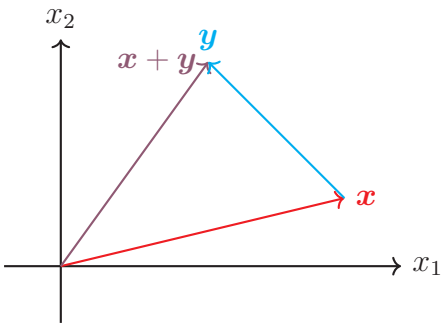


Figure 2.1: Illustration of the triangle inequality (2.15) in $\mathbb{R}^2$.

The standard inner products on $\mathbb{R}^n$ and on $\mathbb{C}^n$ introduced above are in fact inner products, that is, they satisfy the requirements in Definition 2.1.2, see Exercise 2.1.3. Let $\mathbb{F}^n$ denote either $\mathbb{R}^n$ and on $\mathbb{C}^n$. Then the standard inner products can be written as:

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle = x_1\overline{y_1} + x_2\overline{y_2} + \ldots + x_n\overline{y_n} = \sum_{k=1}^{n} x_k\overline{y}_k, \tag{2.6}$$

since this formula reduces to the dot product $\boldsymbol{x} \cdot \boldsymbol{y}$ if $\boldsymbol{y}$ is a real vector. The norm induced by the standard inner product is then the so-called 2-norm given by:

$$\|\boldsymbol{x}\| = \sqrt{\langle \boldsymbol{x}, \boldsymbol{x} \rangle} = \sqrt{|x_1|^2 + |x_2|^2 + \ldots + |x_n|^2} = \sqrt{\sum_{k=1}^{n} |x_k|^2}. \tag{2.7}$$

The standard inner product and norm on $\mathbb{F}^n$ are the ones we will use unless stated otherwise.

The standard inner product on $\mathbb{F}^n$ has the following useful property when taking adjoints (or transposes) of a matrix $A \in \mathsf{M}_n(\mathbb{F})$:

$$\langle A\boldsymbol{x}, \boldsymbol{y} \rangle = \langle \boldsymbol{x}, A^*\boldsymbol{y} \rangle \quad \text{and} \quad \langle \boldsymbol{x}, A\boldsymbol{y} \rangle = \langle A^*\boldsymbol{x}, \boldsymbol{y} \rangle. \tag{2.8}$$

for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{F}^n$. If $\mathbb{F} = \mathbb{R}$ one can write A^T in place of A^* in (2.8), e.g., $A\boldsymbol{x} \cdot \boldsymbol{y} = \boldsymbol{x} \cdot A^*\boldsymbol{y}$.

Exercise 2.1.3

Show that the standard inner product on $\mathbb{F}^n$ in (2.6) satisfies the assumptions of Definition 2.1.2.

There can be many different norms and inner products on a vector space, thus, also on $\mathbb{F}^n$. Hence, one sometimes adds subscripts to distinguish

them, e.g., one writes $\langle \boldsymbol{x}, \boldsymbol{y} \rangle_{\mathbb{F}^n}$ for the standard inner product in (2.6). The *standard* inner product is also called the *canonical* inner product. However, this particular inner product is nothing special, it is just the one we will always use unless stated otherwise. As an example of another inner product on $\mathbb{F}^n$, we can define $\langle \boldsymbol{x}, \boldsymbol{y} \rangle_A := \langle A\boldsymbol{x}, A\boldsymbol{y} \rangle_{\mathbb{F}^n}$ where $A \in \mathsf{M}_n(\mathbb{F})$ is assumed to be an invertible matrix.

Exercise 2.1.4

Show that $\langle \cdot\,, \cdot \rangle_A$ is not an inner product on $\mathbb{F}^n$ if A is singular. (Hint: The non-degeneracy condition of Definition 2.1.2 fails.)

Inner products and norms in other spaces. Let us equip some of the vector spaces from Section 10.1 in Mathematics 1a with an inner product.

Example 2.1.5 Frobenius inner product and norm

Recall the vector space of matrices $\mathbb{F}^{m\times n}$ from Example 10.2.2 in Mathematics 1a. In this book we also use the notation $\mathsf{M}_{m\times n}(\mathbb{F})$ as an alternative to $\mathbb{F}^{m\times n}$. This vector space becomes an inner product space under the *Frobenius inner product*

$$\langle A, B \rangle_F = \operatorname{trace}(B^*A).$$

The Frobenius norm is then

$$\|A\|_F = \sqrt{\langle A, A \rangle_F} = \sqrt{\operatorname{trace}(A^*A)} = \left(\sum_{i=1}^{m} \sum_{j=1}^{n} |a_{i,j}|^2 \right)^{1/2}.$$

The Frobenius norm is often used in machine learning and many other areas of applied mathematics.

We leave it as an exercise for the reader to prove that Frobenius inner product indeed is an inner product.

Example 2.1.6

Let P_n be the complex vector space of polynomials of degree at most n. Given a finite, nonempty real interval $[a, b]$, we consider the vector space $P_n([a, b])$ of degree n complex polynomial functions restricted to $[a, b]$. For polynomials $p, q \in V$ we define

$$\langle p, q \rangle_{L^2} = \int_a^b p(x)\overline{q(x)}\,\mathrm{d}x \tag{2.9}$$

This is known as the L^2 *inner product* on P_n over the interval $[a, b]$, and the

inner product plays an important role in many areas of mathematics and physics, e.g., in signal processing and quantum mechanics.

Unit and orthogonal vectors. Suppose V is an inner product space over $\mathbb{F}$ with inner product $\langle \cdot, \cdot \rangle$ and derived norm $\| \cdot \|$. A vector $\boldsymbol{x} \in V$ with the property that

$$\|\boldsymbol{x}\| = 1 \tag{2.10}$$

is called a *unit norm vector* or simply a *unit vector*. We can always create a unit norm vector from a non-zero vector $\boldsymbol{y} \neq \boldsymbol{0}$ that points in the same direction as $\boldsymbol{y}$. In fact, we just define $\boldsymbol{u}$ by:

$$\boldsymbol{u} = \frac{\boldsymbol{y}}{\|\boldsymbol{y}\|} = \frac{1}{\|\boldsymbol{y}\|}\,\boldsymbol{y}$$

and see that

$$\|\boldsymbol{u}\| = \Big\|\frac{1}{\|\boldsymbol{y}\|}\,\boldsymbol{y}\Big\| = \frac{1}{\|\boldsymbol{y}\|}\,\|\boldsymbol{y}\| = 1,$$

where we have used the scaling law in Theorem 2.1.1. Since $\boldsymbol{u}$ is a (very simple) linear combination of $\boldsymbol{y}$, it follows that $\boldsymbol{u} \in \operatorname{span}\{\boldsymbol{y}\}$ and thus $\boldsymbol{u}$ and $\boldsymbol{y}$ point in the "same direction". We have added quotation marks here since V is an abstract vector space, e.g., $\boldsymbol{u}$ and $\boldsymbol{y}$ could be polynomials, so we take "same direction" to mean exactly $\boldsymbol{u} = c\boldsymbol{y}$ where c is a non-negative real number. If $c < 0$ in the relation $\boldsymbol{x} = c\boldsymbol{y}$, one can say that $\boldsymbol{x}$ and $\boldsymbol{y}$ has "opposite" direction, but if c is complex and non-real this picture does not work. (In fact, in complex vector space it is usually sufficient to know if two vectors span the same subspace, i.e., whether $\operatorname{span}\{\boldsymbol{u}\} = \operatorname{span}\{\boldsymbol{y}\}$).

With Example 2.1.1 on page 37 in mind, the following definition of orthogonality is rather natural.

Definition 2.1.3 Orthogonality

Let V be an inner product space over $\mathbb{F}$ with inner product $\langle \cdot, \cdot \rangle$.

(i) Two vectors $\boldsymbol{x}, \boldsymbol{y} \in V$ are *orthogonal* if $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = 0$. A list/set of vectors is orthogonal if the vectors in the list/set are pairwise orthogonal.

(ii) Two vectors $\boldsymbol{u}, \boldsymbol{v} \in V$ are *orthonormal* if they are both unit vectors and orthogonal. A list/set of vectors is orthonormal if the vectors in the list/set are unit vectors and pairwise orthogonal.

(iii) Let S denote a (nonempty) list/set of vectors in V. Then the orthogonal complement of S in V consists of vectors that are orthogonal

to all vectors in S. It is denoted $S^\perp$ and given by

$$S^\perp = \{\boldsymbol{x} \in V : \langle \boldsymbol{x}, \boldsymbol{s} \rangle = 0 \text{ for all } \boldsymbol{s} \in S\}$$

If $S = \emptyset$, then $S^\perp := V$

Note that also $\langle \boldsymbol{y}, \boldsymbol{x} \rangle = 0$ if $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = 0$ by conjugate symmetry of the inner product. We write $\boldsymbol{x} \perp \boldsymbol{y}$ if $\boldsymbol{x}$ and $\boldsymbol{y}$ are orthogonal. It follows from Definition 2.1.3 that a list of vectors $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n$ is orthonormal if and only if

$$\langle \boldsymbol{u}_j, \boldsymbol{u}_k \rangle = \begin{cases} 1 & \text{if } j = k, \\ 0 & \text{if } j \neq k. \end{cases} \tag{2.11}$$

Example 2.1.7

We consider $\mathbb{R}^3$ as an inner product space over the reals $\mathbb{R}$ with the standard inner product (2.6) (which is the same as the dot product (2.2)). Let $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^3$ be given by

$$\boldsymbol{x} = \begin{bmatrix} 1 \\ 2 \\ 2 \end{bmatrix}, \quad \boldsymbol{y} = \begin{bmatrix} -2/3 \\ 2/3 \\ -1/3 \end{bmatrix}$$

The vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ are *orthogonal*, written $\boldsymbol{x} \perp \boldsymbol{y}$, since

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle = 1\,(-2/3) + 2\,(2/3) + 2\,(-1/3) = 0$$

However, the two vectors are *not orthonormal* as only $\boldsymbol{y}$ is a unit vector. Let us check:

$$\|\boldsymbol{x}\| = \sqrt{\langle \boldsymbol{x}, \boldsymbol{x} \rangle} = \sqrt{1^2 + 2^2 + 2^2} = \sqrt{5} \neq 1$$

and

$$\|\boldsymbol{y}\| = \sqrt{\langle \boldsymbol{y}, \boldsymbol{y} \rangle} = \sqrt{(-2/3)^2 + (2/3)^2 + (-1/3)^2} \\ = \sqrt{4/9 + 4/9 + 1/9} = \sqrt{1} = 1$$

Example 2.1.8

We consider the vector space $\mathsf{M}_{m\times n}(\mathbb{F})$ of $n \times m$ matrices equipped with the Frobenius inner product introduced in Example 2.1.5. In Example 10.2.2 in Mathematics 1a the standard basis of $E^{(i,j)}$ matrices was introduced. The

matrices $E^{(i,j)}$ are defined as $n \times m$ matrices of zeros except at index (i,j), where the matrix has a one. The list containing the nm matrices $E^{(i,j)}$ is orthonormal, i.e.,

$$\langle E^{(i,j)}, E^{(i',j')} \rangle_F = \operatorname{trace}\left((E^{(i',j')})^* E^{(i,j)}\right) = \begin{cases} 1 & \text{if } j = j' \text{ and } i = i', \\ 0 & \text{if } j \neq j' \text{ or } i \neq i'. \end{cases}$$

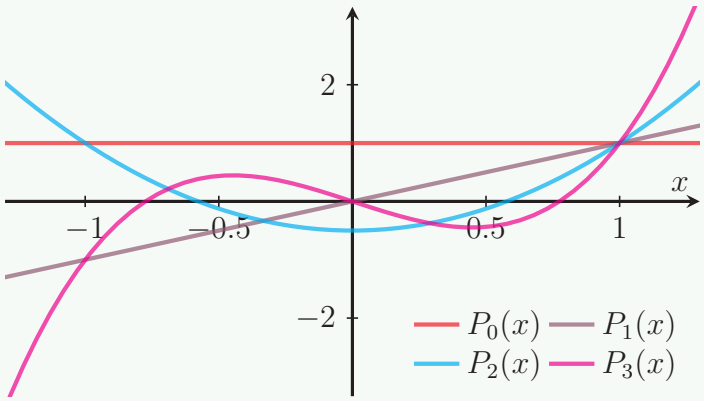
This can be verified by computing the diagonal of $(E^{(i',j')})^* E^{(i,j)}$. Note that we are only interested in the diagonal of this matrix product as the Frobenius inner product only takes the trace of the product.

Example 2.1.9

The Legendre polynomials are a family of orthogonal polynomials with respect to the L^2-inner product introduced in (2.9) on the interval $[-1,1]$. The first four Legendre polynomials are given by:

$$P_0(x) = 1, \quad P_1(x) = x, \quad P_2(x) = \tfrac{1}{2}\left(3x^2 - 1\right), \quad P_3(x) = \tfrac{1}{2}\left(5x^3 - 3x\right)$$

and plotted here (note we are only interested in the polynomial inside the interval $[-1,1]$):



It is straightforward to check that P_0, P_1, P_2, P_3 is an orthogonal list in $P_3([-1,1])$. Let us check one of these orthogonalities:

$$\begin{aligned} \langle P_0, P_2 \rangle &= \int_{-1}^{1} P_0(x) P_2(x)\, \mathrm{d}x = \int_{-1}^{1} \tfrac{1}{2}\left(3x^2 - 1\right) \mathrm{d}x \\ &= \tfrac{1}{2}[x^3 - x]_{x=-1}^{1} = \tfrac{1}{2}\Big((1^3 - 1) - ((-1)^3 - (-1))\Big) = \tfrac{1}{2}(0 - 0) = 0 \end{aligned}$$

hence $P_0 \perp P_2$. Note that the list is not ortho*normal* as the polynomials are not normalized, e.g.,

$$\|P_1\|^2 = \int_{-1}^{1} |P_1(x)|^2 \, \mathrm{d}x = \int_{-1}^{1} x^2 \, \mathrm{d}x = \left[\tfrac{1}{3}x^3\right]_{x=-1}^{1} = \tfrac{2}{3} \neq 1$$

The only vector that is orthogonal to all other vectors is the zero vector. Let us state and prove this simple but very useful fact; it is an essential property in the so-called weak formulation of partial differential equations used in, e.g., the finite element method.

Lemma 2.1.2

Let V be an inner product space over $\mathbb{F}$ with inner product $\langle \cdot, \cdot \rangle$. If $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \langle \boldsymbol{x}, \boldsymbol{z} \rangle$ for all $\boldsymbol{x} \in V$, then $\boldsymbol{y} = \boldsymbol{z}$. In particular, if $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = 0$ for all $\boldsymbol{x} \in V$, then $\boldsymbol{y} = \boldsymbol{0}$.

Proof. Suppose $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \langle \boldsymbol{x}, \boldsymbol{z} \rangle$ for all $\boldsymbol{x} \in V$. Then $\langle \boldsymbol{x}, \boldsymbol{y} - \boldsymbol{z} \rangle = 0$ for all $\boldsymbol{x} \in V$. Taking $\boldsymbol{x} = \boldsymbol{y} - \boldsymbol{z}$ yields $\langle \boldsymbol{y} - \boldsymbol{z}, \boldsymbol{y} - \boldsymbol{z} \rangle = 0$. By non-degeneracy of the inner product it follows that $\boldsymbol{y} - \boldsymbol{z} = \boldsymbol{0}$, i.e., $\boldsymbol{y} = \boldsymbol{z}$. ■

The Pythagorean theorem, known from two-dimensional and three-dimensional euclidean spaces[1], just think of $\mathbb{R}^2$ and $\mathbb{R}^3$, can also be generalized to inner product spaces.

Theorem 2.1.3 Pythagoras

Let V be an inner product space over $\mathbb{F}$ with inner product $\langle \cdot, \cdot \rangle$ and derived norm $\| \cdot \|$. If $\boldsymbol{x}$ and $\boldsymbol{y}$ are orthogonal, then

$$\|\boldsymbol{x} + \boldsymbol{y}\|^2 = \|\boldsymbol{x}\|^2 + \|\boldsymbol{y}\|^2 \tag{2.12}$$

Proof. Assume $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = 0$. Then $\langle \boldsymbol{y}, \boldsymbol{x} \rangle = \overline{\langle \boldsymbol{x}, \boldsymbol{y} \rangle} = \bar{0} = 0$. Now just compute:

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

[1] Strictly speaking, a euclidean vector is a geometric vector, often represented by a directed line segment (an "arrow"), that has a length and a direction. In physics this representation is often used, e.g., as the velocity or force vector. Here, we will think of euclidean vectors as elements of $\boldsymbol{x} \in \mathbb{R}^n$ (an "arrow" from $\boldsymbol{0}$ to $\boldsymbol{x}$) equipped with the standard inner product.

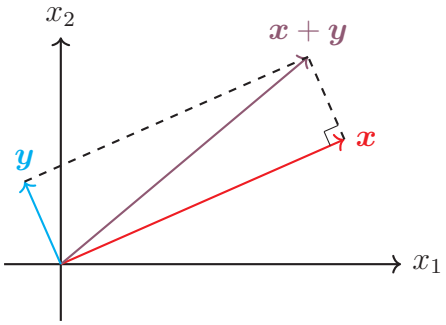


Figure 2.2: Illustration of the Pythagorean theorem in $\mathbb{R}^2$.

$$
\begin{aligned}
&= \langle \boldsymbol{x}, \boldsymbol{x} \rangle + \langle \boldsymbol{x}, \boldsymbol{y} \rangle + \langle \boldsymbol{y}, \boldsymbol{x} \rangle + \langle \boldsymbol{y}, \boldsymbol{y} \rangle \\
&= \|\boldsymbol{x}\|^2 + 0 + 0 + \|\boldsymbol{y}\|^2 \\
&= \|\boldsymbol{x}\|^2 + \|\boldsymbol{y}\|^2.
\end{aligned}
$$

■

Two important inequalities for the inner product and norm. The inner product and the derived norm satisfies the following useful inequality:

Theorem 2.1.4 Cauchy/Schwarz' inequality

Let V be an inner product space over $\mathbb{F}$ with inner product $\langle \cdot, \cdot \rangle$ and derived norm $\|\cdot\|$. Then

$$|\langle \boldsymbol{x}, \boldsymbol{y} \rangle| \leq \|\boldsymbol{x}\| \|\boldsymbol{y}\|. \tag{2.13}$$

Proof. Let $\boldsymbol{x}, \boldsymbol{y} \in V$. If $\boldsymbol{y}$ is the zero vector, then both sides of the inequality are equal to zero, and the inequality is trivially satisfied. Thus, we can assume that $\boldsymbol{y} \neq \boldsymbol{0}$.

Suppose first that $\langle \boldsymbol{x}, \boldsymbol{y} \rangle$ is real. We know that, for all $t \in \mathbb{R}$,

$$\|\boldsymbol{x} + t\boldsymbol{y}\|^2 \geq 0. \tag{2.14}$$

since the norm is always non-negative. Expanding the left-hand side using the properties of the norm, we have

$$
\begin{aligned}
\|\boldsymbol{x} + t\boldsymbol{y}\|^2 &= \langle \boldsymbol{x} + t\boldsymbol{y}, \boldsymbol{x} + t\boldsymbol{y} \rangle = \langle \boldsymbol{x}, \boldsymbol{x} \rangle + \langle \boldsymbol{x}, t\boldsymbol{y} \rangle + \langle t\boldsymbol{y}, \boldsymbol{x} \rangle + \langle t\boldsymbol{y}, t\boldsymbol{y} \rangle \\
&= \|\boldsymbol{x}\|^2 + 2t\langle \boldsymbol{x}, \boldsymbol{y} \rangle + t^2 \|\boldsymbol{y}\|^2,
\end{aligned}
$$

where we have used that $\langle \boldsymbol{x}, \boldsymbol{y} \rangle$ is real. The expression above is a quadratic polynomial in t with the leading coefficient $\|\boldsymbol{y}\|^2 > 0$, which means it opens upwards. Therefore, since the polynomial is non-negative for all $t \in \mathbb{R}$, the discriminant of the polynomial must be zero or negative. Indeed, if the discriminant is positive, the polynomial has two real, distinct roots and therefore taking negative values for t between the two roots, contradicting (2.14). Hence, the discriminant is non-positive, that is,

$$D = (2\langle \boldsymbol{x}, \boldsymbol{y} \rangle)^2 - 4\|\boldsymbol{x}\|^2\|\boldsymbol{y}\|^2 \leq 0.$$

Rearranging this inequality and dividing through by 4, we get

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle^2 \leq \|\boldsymbol{x}\|^2\|\boldsymbol{y}\|^2$$

and by taking the square root of both sides, we obtain

$$|\langle \boldsymbol{x}, \boldsymbol{y} \rangle| \leq \|\boldsymbol{x}\|\|\boldsymbol{y}\|$$

Now, if $c := \langle \boldsymbol{x}, \boldsymbol{y} \rangle$ is not real, we replace $\boldsymbol{x}$ by $\boldsymbol{x'} = \frac{\bar{c}}{|c|}\boldsymbol{x}$. Then $\|\boldsymbol{x}\| = \|\boldsymbol{x'}\|$ and $|\langle \boldsymbol{x}, \boldsymbol{y} \rangle| = |\langle \boldsymbol{x'}, \boldsymbol{y} \rangle|$. Since $\langle \boldsymbol{x'}, \boldsymbol{y} \rangle = \frac{\bar{c}}{|c|}c = |c|$ is real, the general case follows from the special case we considered above.

■

For $V = \mathbb{R}^n$ the Cauchy/Schwarz' inequality, when squaring both sides, reads

$$(x_1y_1 + \cdots + x_ny_n)^2 \leq \left(x_1^2 + \cdots + x_n^2\right)\left(y_1^2 + \cdots + y_n^2\right),$$

where $\boldsymbol{x} = [x_1, \ldots, x_n]^T$ and $\boldsymbol{y} = [y_1, \ldots, y_n]^T$ are real vectors. It is not obvious that such a relation should hold. In fact, it is rather difficult to show this Cauchy/Schwarz' inequality by means of algebraic methods alone.

The second inequality we want to prove here is the triangle inequality for inner product space. Intuitively, the length of the sum of two vectors $\boldsymbol{x} + \boldsymbol{y}$ should be maximized if the two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ point in the same direction. Moreover, the maximal length of $\boldsymbol{x} + \boldsymbol{y}$ should not be able to be bigger than the sum of the individual lengths of the two vectors. The precise statement behind these ideas can be formulated as follows:

Theorem 2.1.5 Triangle inequality

Let V be an inner product space over $\mathbb{F}$ with inner product $\langle \cdot, \cdot \rangle$ and derived norm $\|\cdot\|$, and let $\boldsymbol{x}, \boldsymbol{y} \in V$. Then

$$\|\boldsymbol{x} + \boldsymbol{y}\| \leq \|\boldsymbol{x}\| + \|\boldsymbol{y}\| \tag{2.15}$$

with equality if and only if $\boldsymbol{x}$ and $\boldsymbol{y}$ points in the same direction, i.e., one

vector is a non-negative scalar multiple of the other.

Proof. By properties of the inner product and Cauchy/Schwarz' inequality, we compute:

$$\begin{aligned}\|\boldsymbol{x}+\boldsymbol{y}\|^2 &= \langle \boldsymbol{x}+\boldsymbol{y}, \boldsymbol{x}+\boldsymbol{y}\rangle \\ &= \langle \boldsymbol{x},\boldsymbol{x}\rangle + \langle \boldsymbol{x},\boldsymbol{y}\rangle + \langle \boldsymbol{y},\boldsymbol{x}\rangle + \langle \boldsymbol{y},\boldsymbol{y}\rangle \\ &= \langle \boldsymbol{x},\boldsymbol{x}\rangle + \langle \boldsymbol{x},\boldsymbol{y}\rangle + \overline{\langle \boldsymbol{x},\boldsymbol{y}\rangle} + \langle \boldsymbol{y},\boldsymbol{y}\rangle \\ &= \|\boldsymbol{x}\|^2 + 2\operatorname{Re}\langle \boldsymbol{x},\boldsymbol{y}\rangle + \|\boldsymbol{y}\|^2 \\ &\le \|\boldsymbol{x}\|^2 + 2|\langle \boldsymbol{x},\boldsymbol{y}\rangle| + \|\boldsymbol{y}\|^2 \\ &\le \|\boldsymbol{x}\|^2 + 2\|\boldsymbol{x}\|\|\boldsymbol{y}\| + \|\boldsymbol{y}\|^2 \\ &= \left(\|\boldsymbol{x}\| + \|\boldsymbol{y}\|\right)^2\end{aligned}$$

where we have used the following well-known properties: $z+\overline{z} = 2\operatorname{Re} z$ and $\operatorname{Re} z \le |z|$ for $z \in \mathbb{C}$ and $a^2+2ab+b^2 = (a+b)^2$ for $a, b \in \mathbb{R}$.

Finally, to prove the "with equality"-statement in the theorem, we need to prove that the two inequalities in the above computations become equalities precisely when $\boldsymbol{x}$ and $\boldsymbol{y}$ points in the same direction. We leave the details to the reader.

■

2.2 Open and closed sets

For functions of one variable the concept of an *open interval* is crucial in order to study continuity and differentiability. The goal of this section is to introduce the corresponding concept in an inner product space V and therefore, in particular, for $\mathbb{R}^n$.

From the norm $\|\cdot\|$ on an inner product space V we will define open and closed sets. This is a so-called *topology* on V, and it is precisely the topology of $\mathbb{R}^n$ that is crucial when we want to consider continuity and differentiability of functions of several variables.

Throughout this section, we let V be an inner product space over $\mathbb{F}$ with inner product $\langle \cdot, \cdot \rangle$ and derived norm $\|\boldsymbol{x}\| = \langle \boldsymbol{x}, \boldsymbol{x}\rangle^{1/2}$. However, you can safely think of $\mathbb{R}^n$ with the euclidean norm whenever you see V. To simplify notation, we will also write vectors $\boldsymbol{x} = [x_1, \ldots, x_n]^T$ in $\mathbb{R}^n$ as $(x_1, \ldots, x_n)$ when this does not lead to confusion.

Definition 2.2.1 Open ball

Fix a vector $\boldsymbol{x}_0 \in V$. The (open) ball at $\boldsymbol{x}_0$ with radius $r > 0$ is the set

$$B(\boldsymbol{x}_0, r) \coloneqq \{\, \boldsymbol{x} \in V \mid \|\boldsymbol{x} - \boldsymbol{x}_0\| < r \,\}. \tag{2.16}$$

In $\mathbb{R}^2$, the ball at a point $\boldsymbol{x}_0$ corresponds precisely to the set of points inside a circle with center at the point.

Example 2.2.1
Let $\boldsymbol{x}_0 = (1, 2) \in \mathbb{R}^2$ and $r = 2$. Then

$$\begin{aligned} B(\boldsymbol{x}_0, r) &= B((1,2), r) \\ &= \{\, \boldsymbol{x} \in \mathbb{R}^2 \mid \|\boldsymbol{x} - (1,2)\| < 2 \,\} \\ &= \{\, (x_1, x_2) \in \mathbb{R}^2 \mid \sqrt{(x_1-1)^2 + (x_2-2)^2} < 2 \,\} \\ &= \{\, (x_1, x_2) \in \mathbb{R}^2 \mid (x_1-1)^2 + (x_2-2)^2 < 2^2 \,\}. \end{aligned}$$

This is precisely the points in $\mathbb{R}^2$ that are located within a circle of radius 2 centered at the point $\boldsymbol{x}_0 = (1, 2)$.

Definition 2.2.2 Open set
Consider a subset U of an inner product space V. Assume that for each $\boldsymbol{x}_0 \in U$, there exists a ball $B(\boldsymbol{x}_0, \epsilon)$ at $\boldsymbol{x}_0$ that is entirely contained in U, that is, there exists $\epsilon > 0$ such that

$$B(\boldsymbol{x}_0, \epsilon) \subseteq U.$$

Then the set U is said to be *open*.

The concept of an open set in $\mathbb{R}^n$, in fact, generalizes the concept of open intervals in $\mathbb{R}$ as the following example shows.

Example 2.2.2 (a) An interval of the form $]a, b[$ in $\mathbb{R}$ is an open set. Indeed, for each $x_0 \in]a, b[$, letting ϵ denote the minimal distance to either a or b, the set $B(x_0, \epsilon/2)$ in $\mathbb{R}$ is completely contained in $]a, b[$.

(b) A ball $B(\boldsymbol{x}_0, r)$ in $\mathbb{R}^n$ is itself an open set.

(c) Letting $a > 0$, a set in $\mathbb{R}^2$ of the form $[-a, a]^2 = [-a, a] \times [-a, a]$ is not open. Indeed, letting, e.g., $\boldsymbol{x}_0 = (a, a)$, every ball $B(\boldsymbol{x}_0, r)$ will contain points outside the set $[-a, a]^2$.

We now define the concept of a *closed set* U in V. Interestingly, this is defined via a condition on the *complement* of the set U, i.e., the set $V \setminus U$.

Definition 2.2.3 Closed set

A set U in V is said to be *closed* if its complement $V \setminus U$ is an open set.

Example 2.2.3 (a) An interval of the form $[a, b]$ in $\mathbb{R}$ is a closed set. Indeed,

$$\mathbb{R} \setminus [a, b] =]-\infty, a[\cup]b, \infty[,$$

which is a union of two open sets and hence open.

(b) Let $\boldsymbol{x}_0 \in \mathbb{R}^2$ and $r > 0$. Then the set

$$\{\boldsymbol{x} \in \mathbb{R}^2 \,\big|\, ||\boldsymbol{x} - \boldsymbol{x}_0|| \leq r\}$$

is closed. Compare the set with the definition of the ball in Definition 2.2.1.

(c) Letting $a > 0$, a set in $\mathbb{R}^2$ of the form $[-a, a]^2 = [-a, a] \times [-a, a]$ is closed.

Note that many sets in V and $\mathbb{R}^n$ are neither open nor closed (e.g., intervals of the form $[a, b[$ in $\mathbb{R}$). On the other hand, there are two special sets with the particular property of being both open and closed, namely, the entire space V and the empty set $\emptyset$.

We need a few more concepts related to sets in $\mathbb{R}^n$.

Definition 2.2.4 Boundary of a set

Consider a subset M of V. The boundary of M, to be denoted by ∂M, consists of precisely the points $\boldsymbol{x}_0 \in V$ for which *each* ball $B(\boldsymbol{x}_0, \epsilon)$ contains points from the set M *and* from the complement $V \setminus M$.

Note that the points in the boundary ∂M do not necessarily belong to the set M.

Example 2.2.4

Consider an interval $]a, b[$ in $\mathbb{R}$. Then the boundary consists of precisely the two points a and b, i.e., $\partial]a, b[= \{a, b\}$.

Similarly, for the set $\{\boldsymbol{x} \in \mathbb{R}^2 \,\big|\, ||\boldsymbol{x} - \boldsymbol{x}_0|| \leq r\}$ considered in Example 2.2.3(b), the boundary is

$$\partial\{\, \boldsymbol{x} \in \mathbb{R}^2 \mid \|\boldsymbol{x} - \boldsymbol{x}_0\| \leq r \,\} = \{\, \boldsymbol{x} \in \mathbb{R}^2 \mid \|\boldsymbol{x} - \boldsymbol{x}_0\| = r \,\}.$$

The reader is encouraged to make a graphical illustration of this result. In

words, it says that the boundary of a disc with radius simply consists of the corresponding circle. If we, e.g., consider the case where $r = 1$ and $\boldsymbol{x}_0 = \mathbf{0} = (0, 0)$, this is precisely the unit circle. It can be described as the set of points

$$\boldsymbol{r}(t) = \begin{bmatrix} \cos(t) \\ \sin(t) \end{bmatrix} \tag{2.17}$$

where $t \in [0, 2\pi]$. We say that the vector function $\boldsymbol{r}$ provides a *parametrization* of the boundary curve; cf. Equation (1.10) on page 21 and the discussion following the equation.

Example 2.2.5

Let $a > 0$ and consider the rectangle

$$[-a, a]^2 = [-a, a] \times [-a, a]$$

in $\mathbb{R}^2$. Then the boundary consists of exactly what our intuitive understanding of the word says, namely, the four line pieces that bound the rectangle. It is a little bit more complicated to give an exact mathematical description than in Example 2.2.4, due to the fact that the boundary consists of four pieces. One way of describing the boundary is as the union

$$(\{a\} \times [-a, a]) \cup (\{-a\} \times [-a, a]) \cup (]-a, a[\times\{a\}) \cup (]-a, a[\times\{-a\}) .$$

Due to this more complicated structure of the boundary, the parametrization is more complicated. Indeed, we have to choose separate parametrizations for each of the four curves. The piece of the boundary consisting of the points $\{a\} \times [-a, a]$ can be parametrized as

$$\boldsymbol{r}(t) = \begin{bmatrix} a \\ t \end{bmatrix}, \tag{2.18}$$

where $t \in [-a, a]$.

Definition 2.2.5 Closure and interior of a set

Consider a subset M of V. The *closure* of the set M, to be denoted by $\overline{M}$, consists of the union of the set M and its boundary ∂M, i.e.,

$$\overline{M} = M \bigcup \partial M.$$

The *interior* of the set M, to be denoted by M°, consists of the points in

M that are *not* boundary points, i.e.,

$$M^\circ = M \setminus \partial M.$$

Using the concept of the norm of a vector in V, we will now define what it means that a set in V is *bounded*.

Definition 2.2.6 Bounded set

A set U in V is said to be *bounded* if there exists $r > 0$ such that $||\boldsymbol{x}|| \leq r$ for all $\boldsymbol{x} \in U$.

Geometrically, a set U is bounded if there exists a sufficiently big ball in V that contains U.

Example 2.2.6 (a) A ball $B(\boldsymbol{x}_0, r)$ in $\mathbb{R}^n$ is itself a bounded set.

(b) Letting $a > 0$, a set in $\mathbb{R}^n$ of the form $[-a, a]^n$ is bounded. Indeed, if $\boldsymbol{x} = (x_1, x_2, \ldots, x_n) \in [-a, a]^n$, then $|x_k| \leq a$ for all $k = 1, \ldots, n$. Hence

$$||\boldsymbol{x}||^2 = x_1^2 + x_2^2 + \cdots + x_n^2 \leq a^2 + a^2 + \cdots a^2 = na^2;$$

this implies that $||\boldsymbol{x}|| \leq \sqrt{n}a$ for all $\boldsymbol{x} \in [-a, a]^n$.

(c) The graph of the function $y = x^2$, i.e., the set

$$\{(x, y) \in \mathbb{R}^2 \,\big|\, y = x^2\}$$

is not a bounded set in $\mathbb{R}^2$.

2.3 Projections onto a line

In $\mathbb{R}^2$ and $\mathbb{R}^3$ we can project a vector (orthogonally) onto another vector as shown in Figure 2.3. The aim of this section is to generalize this construction to arbitrary inner product spaces.

Let $\boldsymbol{y}$ be a non-zero vector in an inner product space V. We consider $Y \coloneqq \text{span}\{\boldsymbol{y}\}$ as the line spanned by $\boldsymbol{y}$; Y is of course a one-dimensional subspace of V. The *orthogonal projection* of a vector $\boldsymbol{x} \in V$ onto the subspace $Y = \text{span}\{\boldsymbol{y}\}$ is defined as:

$$\text{proj}_Y : V \to V, \quad \text{proj}_Y(\boldsymbol{x}) = \frac{\langle \boldsymbol{x}, \boldsymbol{y} \rangle}{\langle \boldsymbol{y}, \boldsymbol{y} \rangle} \boldsymbol{y}.$$

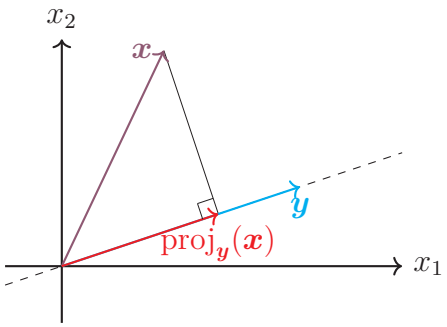


Figure 2.3: Orthogonal projection in $\mathbb{R}^2$.

It is easy to check that proj_Y is a linear mapping (do it!). Note that the image of proj_Y indeed is Y, i.e., $\mathrm{im}(\mathrm{proj}_Y) = Y$.

We can simplify the expression for proj_Y by normalizing the spanning vector $\boldsymbol{y}$. So let $\boldsymbol{u} = \boldsymbol{y}/\|\boldsymbol{y}\|$. Then $Y = \mathrm{span}\{\boldsymbol{y}\} = \mathrm{span}\{\boldsymbol{u}\}$ and

$$\mathrm{proj}_Y(\boldsymbol{x}) = \langle \boldsymbol{x}, \boldsymbol{u} \rangle \boldsymbol{u}. \tag{2.19}$$

With abuse of notation we will also write the projection proj_Y onto Y as $\mathrm{proj}_{\boldsymbol{y}}$, where $\boldsymbol{y}$ is any vector that spans Y.

Exercise 2.3.1

(a) Verify the formula in (2.19). (b) Show that $\mathrm{proj}_Y : V \to V$ is a linear mapping. (c) Show that $\mathrm{proj}_Y \circ \mathrm{proj}_Y = \mathrm{proj}_Y$. Thus, it does not matter if we apply the mapping once or twice (or n times), the output is the same. In other words, nothing happens to the input of the mapping the second time we apply the projections. Why is this a reasonable property of a projection?

Example 2.3.2

We consider the vector space $\mathbb{F}^n$ with the standard inner product $\langle \cdot, \cdot \rangle$. Recall that the standard inner product is given by

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \boldsymbol{y}^* \boldsymbol{x} = \sum_{k=1}^{n} x_k \overline{y_k}$$

for $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{F}^n$. Let $\boldsymbol{u} \in \mathbb{F}^n$ be a unit vector and let $Y = \mathrm{span}\{\boldsymbol{u}\}$ be the one-dimensional subspace spanned by $\boldsymbol{u}$. Then, for any $\boldsymbol{x} \in \mathbb{F}^n$: $\mathrm{proj}_Y(\boldsymbol{x}) = \langle \boldsymbol{x}, \boldsymbol{u} \rangle \boldsymbol{u} = \boldsymbol{u} \langle \boldsymbol{x}, \boldsymbol{u} \rangle = \boldsymbol{u} (\boldsymbol{u}^* \boldsymbol{x}) = (\boldsymbol{u}\boldsymbol{u}^*) \boldsymbol{x} = P\boldsymbol{x}$. The $n \times n$ matrix $P = \boldsymbol{u}\boldsymbol{u}^*$ is called a projection matrix and $\boldsymbol{u}\boldsymbol{u}^*$ is called an outer product (as a matrix product of a column vector and a row vector). The

projection matrix is Hermitian (i.e., $P^* = P$) and idempotent (i.e., $P^2 = P$) as is readily verified:

$$P^* = (\boldsymbol{u}\boldsymbol{u}^*)^* = (\boldsymbol{u}^*)^*\boldsymbol{u}^* = \boldsymbol{u}\boldsymbol{u}^* = P$$

and

$$P^2 = (\boldsymbol{u}\boldsymbol{u}^*)^2 = \boldsymbol{u}\boldsymbol{u}^*\boldsymbol{u}\boldsymbol{u}^* = \boldsymbol{u}(\boldsymbol{u}^*\boldsymbol{u})\boldsymbol{u}^* = \boldsymbol{u}\boldsymbol{u}^* = P$$

since $\boldsymbol{u}^*\boldsymbol{u} = \langle \boldsymbol{u}, \boldsymbol{u}\rangle = \|\boldsymbol{u}\|^2 = 1$. We finally note that $\operatorname{col} P = Y$.

Example 2.3.3

Consider $\boldsymbol{y} = [1, 2, 2]^T \in \mathbb{R}^3$, and let $Y = \operatorname{span}\{\boldsymbol{y}\}$. We compute $\langle \boldsymbol{y}, \boldsymbol{y}\rangle = 1^2 + 2^2 + 2^2 = 9$. The orthogonal projection onto Y is therefore given by

$$\operatorname{proj}_Y(\boldsymbol{x}) = \frac{1}{9}\left\langle \boldsymbol{x}, \begin{bmatrix}1\\2\\2\end{bmatrix}\right\rangle \begin{bmatrix}1\\2\\2\end{bmatrix}$$

for $\boldsymbol{x} \in \mathbb{R}^3$. Alternatively, we normalize the vector $\boldsymbol{y}$ to a unit vector $\boldsymbol{u} = [1/3, 2/3, 2/3]^T \in Y$ and compute

$$P = \boldsymbol{u}\boldsymbol{u}^* = \begin{bmatrix}1/3\\2/3\\2/3\end{bmatrix}[1/32/32/3] = \frac{1}{9}\begin{bmatrix}1 & 2 & 2\\2 & 4 & 4\\2 & 4 & 4\end{bmatrix} \tag{2.20}$$

Hence, the projection $\operatorname{proj}_Y(\boldsymbol{x})$ can also be computed by $P\boldsymbol{x}$ for any $\boldsymbol{x} \in \mathbb{R}^3$.

2.4 Orthonormal basis

Recall from Definition 10.2.4 in Mathematics 1a that a finite ordered basis in a vector space V over $\mathbb{F}$ is a list of linearly independent vectors $\alpha \coloneqq \boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$ with the property that any vector $\boldsymbol{x}$ in V can be written as a linear combination of the vectors from the list α, that is,

$$\boldsymbol{x} = \sum_{k=1}^{n} c_k \boldsymbol{v}_k \qquad \text{for all } \boldsymbol{x} \in V. \tag{2.21}$$

In this text we will simply say that α is a *basis* as all bases will be finite and ordered. Recall also that if V is an n-dimensional vector space, then any list of n linearly independent vectors is a basis.

One can show that for a given vector $\boldsymbol{x} \in V$ the scalars (also called coefficients) $c_k \in \mathbb{F}$, $k = 1, \ldots, n$, are unique. The column vector of

coefficients is the coordinate vector, and it is denoted

$$ {}_{\alpha}\boldsymbol{x} = \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{bmatrix}. $$

That is, there is a one-to-one correspondence between vectors and coordinate vectors. Hence, instead of working with (potentially complicated) vectors $\boldsymbol{x}$ (e.g., a polynomial function), we can always choose to work with column (coordinate) vectors in $\mathbb{F}^n$. However, to actually find ${}_{\alpha}\boldsymbol{x}$ we usually have to solve a linear system of equations which is not always feasible in applications. To address this problem, we introduce a particular nice class of bases called orthonormal bases.

Definition 2.4.1 Orthonormal basis

Let V be an inner product vector space over $\mathbb{F}$. A list β of vectors in V is called an *orthonormal basis* of V if the list is orthonormal and a basis.

Suppose $\beta \coloneqq \boldsymbol{u}_1, \ldots, \boldsymbol{u}_n$ is an orthonormal basis in an inner product vector space V over $\mathbb{F}$. We fix $j = 1, \ldots, n$ and compute

$$ \langle \boldsymbol{x}, \boldsymbol{u}_j \rangle = \langle \sum_{k=1}^{n} c_k \boldsymbol{u}_k, \boldsymbol{u}_j \rangle = \sum_{k=1}^{n} c_k \langle \boldsymbol{u}_j, \boldsymbol{u}_k \rangle $$

However, all the terms in the sum $\sum_{k=1}^{n} c_k \langle \boldsymbol{u}_j, \boldsymbol{u}_k \rangle$ are zero *unless* k is equal to the fixed number j since, by orthonormality of β,

$$ \langle \boldsymbol{u}_j, \boldsymbol{u}_k \rangle = \begin{cases} 1 & \text{if } j = k, \\ 0 & \text{if } j \neq k. \end{cases} $$

Therefore,

$$ \langle \boldsymbol{x}, \boldsymbol{u}_j \rangle = 0 + \cdots + 0 + c_j + 0 + \cdots + 0 = c_j. \tag{2.22} $$

Hence, to find the coefficients c_k, $k = 1, \ldots, n$, we just have to compute the inner products $\langle \boldsymbol{x}, \boldsymbol{u}_k \rangle$.

So, for an orthonormal basis, it is straightforward to compute the coordinate vector of a vector $\boldsymbol{x}$ using the formula:

$$ {}_{\beta}\boldsymbol{x} = \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{bmatrix} = \begin{bmatrix} \langle \boldsymbol{x}, \boldsymbol{u}_1 \rangle \\ \langle \boldsymbol{x}, \boldsymbol{u}_2 \rangle \\ \vdots \\ \langle \boldsymbol{x}, \boldsymbol{u}_n \rangle \end{bmatrix}, $$

and the expansion formula (2.21) becomes:

$$\boldsymbol{x} = \sum_{k=1}^{n} \langle \boldsymbol{x}, \boldsymbol{u}_k \rangle \boldsymbol{u}_k \qquad \text{for all } \boldsymbol{x} \in V.$$

Lemma 2.4.1

Suppose $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n$ is a list of orthonormal vectors in an inner product space. Then $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n$ are linear independent.

Proof. Suppose $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n$ is orthonormal. Consider the expression:

$$\mathbf{0} = \sum_{k=1}^{n} c_k \boldsymbol{u}_k.$$

To show linear independence of $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n$, we have to show that all the coefficients c_k, $k = 1, \ldots, n$, are zero. From the computation (2.22) we know that $c_k = \langle \mathbf{0}, \boldsymbol{u}_k \rangle$ for $k = 1, \ldots, n$. However, for any $k = 1, \ldots, n$, it then follows that $c_k = \langle \mathbf{0}, \boldsymbol{u}_k \rangle = \langle 0\boldsymbol{x}, \boldsymbol{u}_k \rangle = 0\langle \boldsymbol{x}, \boldsymbol{u}_k \rangle = 0$ where $\boldsymbol{x} \in V$.

■

Corollary 2.4.2

Let V be an n-dimensional inner product vector space over $\mathbb{F}$, and let $\beta := \boldsymbol{u}_1, \ldots, \boldsymbol{u}_n$ be a list of n vectors. Then β is an orthonormal basis if and only if β is orthonormal, that is, if β satisfies (2.11).

Proof. Suppose $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n$ is orthonormal. By Lemma 2.4.1 the list is linear independent. Since the list consists of n vectors in an n-dimensional vector space, it must span the vector space. Hence, $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n$ is a basis. It is assumed to be orthonormal so we conclude that $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n$ is an orthonormal basis.

■

Example 2.4.1

We consider the standard basis $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n$ for $\mathbb{F}^n$, where $\boldsymbol{e}_k$ is a size n column vector with all zeros except a one at index k, that is $\boldsymbol{e}_k = [0, \ldots, 0, 1, 0, \ldots, 0]^T$. Since $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n$ is a list of n orthonormal vectors, it follows from Corollary 2.4.2 that the standard basis $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n$ is an orthonormal basis.

The standard basis $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n$ for $\mathbb{F}^n$ is obviously an orthonormal basis for both $\mathbb{R}^n$ and $\mathbb{C}^n$.

Remark 2.4.1

Recall that we consider $\mathbb{R}^n$ as a vector space over the field $\mathbb{R}$ and $\mathbb{C}^n$ as a vector space over the field $\mathbb{C}$. Suppose $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n \in \mathbb{R}^n$. If $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n$ is an orthonormal basis for $\mathbb{R}^n$, then $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n$ is also an orthonormal basis for $\mathbb{C}^n$. (Can you give a proof?)

Exercise 2.4.2

Show that the basis considered in Example 2.1.8 on page 44 is, in fact, an orthonormal basis.

2.5 The Gram-Schmidt process

In the previous section we saw that orthonormal bases are particularly nice bases to work with. Orthonormal bases are also important in numerical linear algebra, where orthogonality is frequently employed to stabilize computations. Given any list of vectors in an inner product space, the Gram-Schmidt process is used to find an orthonormal basis for the subspace spanned by those vectors. The Gram-Schmidt process is closely related to the QR decomposition of matrices used in many different applications in mathematics and engineering.

Suppose we have a list of linearly independent vectors $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_\ell$ in an inner product space V. The goal of the Gram-Schmidt procedure is now to generate a list of orthonormal vectors $\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_\ell$ that span the same subspace. The process is as follows:

(i) (initialization) Set $\boldsymbol{w}_1 = \boldsymbol{v}_1$ and $\boldsymbol{u}_1 = \frac{1}{\|\boldsymbol{w}_1\|}\boldsymbol{w}_1$.

(ii) (orthogonalization & normalization) For $k = 2, \ldots, \ell$ do as follows: construct $\boldsymbol{w}_k$ by removing the component of $\boldsymbol{v}_k$ along each of $\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_{k-1}$:

$$\boldsymbol{w}_k \coloneqq \boldsymbol{v}_k - \sum_{j=1}^{k-1} \operatorname{proj}_{\boldsymbol{u}_j}(\boldsymbol{v}_k),$$

where $\operatorname{proj}_{\boldsymbol{u}_j}(\boldsymbol{v}_k)$ is the orthogonal projection of $\boldsymbol{v}_k$ onto the subspace $\operatorname{span}\{\boldsymbol{u}_j\}$. Then normalize $\boldsymbol{w}_k$ to make sure it is a unit vector:

$$\boldsymbol{u}_k \coloneqq \frac{\boldsymbol{w}_k}{\|\boldsymbol{w}_k\|}. \tag{2.23}$$

Example 2.5.1

Let $V \in \mathsf{M}_{3\times 2}(\mathbb{R})$ be given by

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

Denote the two column vectors of V by $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$. The Gram-Schmidt process goes as follows. In the initialization, we compute the norm of $\boldsymbol{v}_1$

$$\|\boldsymbol{v}_1\| = (1^2 + 2^2 + 2^2)^{1/2} = \sqrt{9} = 3,$$

and normalize $\boldsymbol{v}_1$ to $\boldsymbol{u}_1 = (1/3)\boldsymbol{v}_1 = [1/3, 2/3, 2/3]^T$. The orthogonal projection $\mathrm{proj}_{\boldsymbol{u}_1}$ onto the subspace spanned by $\boldsymbol{u}_1$ was already found in Equation (2.20) on page 55, so here we can directly compute the projection of $\boldsymbol{v}_2$:

$$\mathrm{proj}_{\boldsymbol{u}_1}(\boldsymbol{v}_2) = P\boldsymbol{v}_2 = \frac{1}{9}\begin{bmatrix} 1 & 2 & 2 \\ 2 & 4 & 4 \\ 2 & 4 & 4 \end{bmatrix}\begin{bmatrix} -2 \\ 4 \\ 0 \end{bmatrix} = \begin{bmatrix} 2/3 \\ 4/3 \\ 4/3 \end{bmatrix}.$$

We then construct $\boldsymbol{w}_2$ by

$$\boldsymbol{w}_2 = \boldsymbol{v}_2 - \mathrm{proj}_{\boldsymbol{u}_1}(\boldsymbol{v}_2) = \begin{bmatrix} -2 \\ 4 \\ 0 \end{bmatrix} - \begin{bmatrix} 2/3 \\ 4/3 \\ 4/3 \end{bmatrix} = \begin{bmatrix} -8/3 \\ 8/3 \\ -4/3 \end{bmatrix}.$$

Finally, we need to normalize $\boldsymbol{w}_2$ as

$$\boldsymbol{u}_2 = \frac{1}{\|\boldsymbol{w}_2\|}\boldsymbol{w}_2 = \begin{bmatrix} -2/3 \\ 2/3 \\ -1/3 \end{bmatrix}.$$

Let us summarize. The list of vectors $\boldsymbol{v}_1, \boldsymbol{v}_2$ is a basis for the column space $\mathrm{col}\, V = \mathrm{span}\{\boldsymbol{v}_1, \boldsymbol{v}_2\}$. From the Gram-Schmidt process we found vectors $\boldsymbol{u}_1, \boldsymbol{u}_2$ that constitute an *orthonormal* basis for the *same* subspace $\mathrm{col}\, V = \mathrm{span}\{\boldsymbol{u}_1, \boldsymbol{u}_2\}$.

Exercise 2.5.2

This exercise is a continuation of Example 2.5.1. Suppose

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

Perform the Gram-Schmidt process on the three column vectors of V. Show that the outcome is the three column vectors of U given by

$$U = \begin{bmatrix} \frac{1}{3} & -\frac{2}{3} & -\frac{2}{3} \\ \frac{2}{3} & \frac{2}{3} & -\frac{1}{3} \\ \frac{2}{3} & -\frac{1}{3} & \frac{2}{3} \end{bmatrix}.$$

Note that, if a list $\boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_3$, as in Exercise 2.5.2, spans the entire space $\mathbb{R}^3$, and if we are only asked to find an orthonormal basis for the span of the list (the entire space), there is no need to do the Gram-Schmidt process. We can simple pick the standard basis $\boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{e}_3$. What makes the orthonormal basis found via the Gram-Schmidt process special is the property (2.24) stated below. Let us prove this property and at the same time show that the Gram-Schmidt process always works as claimed:

Theorem 2.5.1 Gram-Schmidt

Suppose $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_\ell$ is a list of linearly independent vectors in an inner product space V. Then the list $\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_\ell$ given by (2.23) is an orthonormal list with the property

$$\operatorname{span}\{\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_k\} = \operatorname{span}\{\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_k\} \tag{2.24}$$

for $k = 1, \ldots, \ell$.

Proof. The proof goes by induction on ℓ. The base case $\ell = 1$ is easy. Since $\boldsymbol{v}_1$ is assumed to be linearly independent, it cannot be the zero vector. Hence, $\boldsymbol{u}_1 = \frac{\boldsymbol{v}_1}{\|\boldsymbol{v}_1\|}$ is well-defined. A list of only one vector is orthonormal precisely when the vector is of norm one. We compute $\|\boldsymbol{u}_1\| = \|\frac{\boldsymbol{v}_1}{\|\boldsymbol{v}_1\|}\| = \frac{1}{\|\boldsymbol{v}_1\|}\|\boldsymbol{v}_1\| = 1$.

For the induction step, let $k \leq \ell$. Suppose that, given $k-1$ linearly independent vectors $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_{k-1}$, the vectors $\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_{k-1}$ are orthonormal and

$$\operatorname{span}\{\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_{k-1}\} = \operatorname{span}\{\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_{k-1}\}.$$

Since $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k$ are linearly independent, it is impossible to write $\boldsymbol{v}_k$ as a linear combination of $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{k-1}$, that is,

$$\boldsymbol{v}_k \notin \operatorname{span}\{\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_{k-1}\} = \operatorname{span}\{\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_{k-1}\}.$$

In particular, $\boldsymbol{v}_k \neq \sum_{j=1}^{k-1} \mathrm{proj}_{\boldsymbol{u}_j}(\boldsymbol{v}_k)$ hence $\boldsymbol{w}_k \neq \boldsymbol{0}$, and the normalization $\boldsymbol{u}_k = \boldsymbol{w}_k/\|\boldsymbol{w}_k\|$ in (2.23) is well-defined.

We need to prove that $\boldsymbol{u}_k$ is orthogonal to $\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_{k-1}$. For $1 \leq m \leq k-1$, we first compute

$$\begin{aligned}\langle \boldsymbol{w}_k, \boldsymbol{u}_m \rangle &= \langle \boldsymbol{v}_k - \sum_{j=1}^{k-1} \mathrm{proj}_{\boldsymbol{u}_j}(\boldsymbol{v}_k), \boldsymbol{u}_m \rangle = \langle \boldsymbol{v}_k, \boldsymbol{u}_m \rangle - \Big\langle \sum_{j=1}^{k-1} \langle \boldsymbol{v}_k, \boldsymbol{u}_j \rangle \boldsymbol{u}_j, \boldsymbol{u}_m \Big\rangle \\ &= \langle \boldsymbol{v}_k, \boldsymbol{u}_m \rangle - \sum_{j=1}^{k-1} \langle \boldsymbol{v}_k, \boldsymbol{u}_j \rangle \langle \boldsymbol{u}_j, \boldsymbol{u}_m \rangle = \langle \boldsymbol{v}_k, \boldsymbol{u}_m \rangle - \sum_{j=1}^{k-1} \langle \boldsymbol{v}_k, \boldsymbol{u}_j \rangle \delta_{j,m} \\ &= \langle \boldsymbol{v}_k, \boldsymbol{u}_m \rangle - \langle \boldsymbol{v}_k, \boldsymbol{u}_m \rangle = 0,\end{aligned}$$

and therefore

$$\langle \boldsymbol{u}_k, \boldsymbol{u}_m \rangle = \Big\langle \frac{\boldsymbol{w}_k}{\|\boldsymbol{w}_k\|}, \boldsymbol{u}_m \Big\rangle = \frac{1}{\|\boldsymbol{w}_k\|} \langle \boldsymbol{w}_k, \boldsymbol{u}_m \rangle = \frac{1}{\|\boldsymbol{w}_k\|} \cdot 0 = 0.$$

Since $\boldsymbol{u}_k$ is defined as a linear combinations of $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_k$, it follows by the induction hypothesis that

$$\mathrm{span}\{\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_k\} \subseteq \mathrm{span}\{\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_k\}.$$

However, the dimension of both these subspaces are k as both lists of vectors are linearly independent by Lemma 2.4.1. We conclude that $\mathrm{span}\{\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_k\}$ is a k-dimensional subspace of a k-dimensional space $\mathrm{span}\{\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_k\}$, and hence these subspaces must agree[2].

■

Exercise 2.5.3

Consider the list $\alpha = 1, x, x^2, x^3$ of polynomials in $P_3([-1,1])$ equipped with the L^2-inner product. Argue that α is a list of linearly independent vectors. Apply the Gram-Schmidt process to α and show that it outputs a *normalized*[a] version of the Legendre polynomials from Example 2.1.9 on page 45.

[a] There is no standard rule of how to "normalize" a polynomial. Considered as a vector in normed vector space, it is natural to normalize polynomials to have norm one, but there are other choices, e.g., one can require that $P(1) = 1$ or that the coefficient of the highest order term is 1.

[2] Strictly speaking, we here assume that the reader is familiar with the fact that the only k-dimensional subspace of a k-dimensional vector space is the vector space itself. We refer the reader to Theorem 2.2.9 in [1] for a proof.

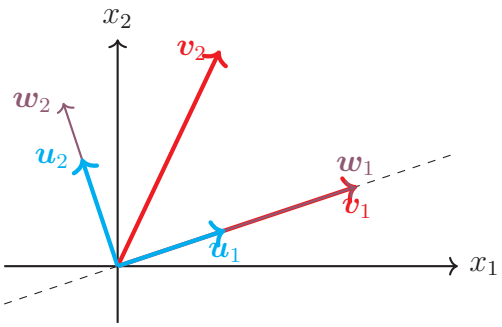


Figure 2.4: Illustration of the Gram-Schmidt process in $\mathbb{R}^2$.

Note that we have formulated the Gram-Schmidt process only for a list of linearly independent vectors. Mathematically, it is not difficult to extend the Gram-Schmidt process to vectors $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_\ell$ that are not linearly independent. The idea is simply to skip a vector $\boldsymbol{v}_k$ in the process whenever the list $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_k$ becomes linearly dependent. Algorithmically, we then update $k \to k+1$ without appending a vector to the output list. Effectively, we remove vectors from the list $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_\ell$ to obtain a linearly independent list of vectors. Numerically, however, the "trick" often leads to instability issues, and it is usually recommended to use an implementation of the QR factorization.

2.6 Unitary and orthogonal matrices

Definition 2.6.1 Unitary and real orthogonal matrix

A square matrix $U \in \mathsf{M}_n(\mathbb{C})$ is unitary if $U^*U = I$. A unitary matrix $Q \in \mathsf{M}_n(\mathbb{R})$ that is real is called *real orthogonal*.

Since $U^*U = I$ if and only if U is invertible with U^* being the inverse, it is easy to invert unitary matrices. If Q is real orthogonal, the adjoint and the transpose are equal, hence it satisfies $Q^TQ = I$. So, for unitary U and real orthogonal Q, we have:

$$U^{-1} = U^* (= \overline{U}^T) \qquad Q^{-1} = Q^T = Q^*.$$

Example 2.6.1

The following 3×3 matrix is real orthogonal and therefore also unitary.

$$U = \begin{bmatrix} \frac{1}{3} & -\frac{2}{3} & -\frac{2}{3} \\ \frac{2}{3} & \frac{2}{3} & -\frac{1}{3} \\ \frac{2}{3} & -\frac{1}{3} & \frac{2}{3} \end{bmatrix} \tag{2.25}$$

This is verified by computing: $U^T U = U^* U = I$.

Consider the vector space $\mathbb{F}^n$ with the standard inner product $\langle \cdot, \cdot \rangle$, i.e., $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \boldsymbol{y}^* \boldsymbol{x}$. Suppose $U \in \mathsf{M}_n(\mathbb{C})$ is unitary and let $\boldsymbol{u}_k$ denote the kth column of U. Then $U^* U = I$ can be written:

$$\begin{bmatrix} \boldsymbol{u}_1^* \\ \boldsymbol{u}_2^* \\ \vdots \\ \boldsymbol{u}_n^* \end{bmatrix} \begin{bmatrix} \boldsymbol{u}_1 & \boldsymbol{u}_2 & \cdots & \boldsymbol{u}_n \end{bmatrix} = \begin{bmatrix} \boldsymbol{u}_1^* \boldsymbol{u}_1 & \boldsymbol{u}_1^* \boldsymbol{u}_2 & \dots & \boldsymbol{u}_1^* \boldsymbol{u}_n \\ \boldsymbol{u}_2^* \boldsymbol{u}_1 & \boldsymbol{u}_2^* \boldsymbol{u}_2 & \dots & \boldsymbol{u}_2^* \boldsymbol{u}_n \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{u}_n^* \boldsymbol{u}_1 & \boldsymbol{u}_n^* \boldsymbol{u}_2 & \cdots & \boldsymbol{u}_n^* \boldsymbol{u}_n \end{bmatrix} = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}$$

which, in turn, is just

$$\langle \boldsymbol{u}_j, \boldsymbol{u}_k \rangle = \begin{cases} 1 & \text{if } j = k, \\ 0 & \text{if } j \neq k. \end{cases}.$$

Hence, a matrix U is unitary, i.e., $U^* U = I$, if and only if the list of its column vectors is an orthonormal list. By Corollary 2.4.2 on page 57 we conclude that the column vectors of a square matrix constitute an orthonormal basis of $\mathbb{C}^n$ if and only if the matrix is unitary. This is one of the key properties of unitary matrices, and therefore also of real orthogonal matrices. We collect a few more characterizing properties in the following result.

Theorem 2.6.1

Let $U \in \mathsf{M}_n(\mathbb{C})$. The following assertions are equivalent:

(i) U is unitary, that is, $U^* U = I$.

(ii) The columns of U are an orthonormal basis of $\mathbb{C}^n$.

(iii) $U U^* = I$.

(iv) U^* is unitary.

(v) The rows of U are an orthonormal basis of $\mathbb{C}^n$.

(vi) For all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{F}^n$: $\langle U\boldsymbol{x}, U\boldsymbol{y} \rangle = \langle \boldsymbol{x}, \boldsymbol{y} \rangle$.

Proof. We have already argued that Item (i) $\Leftrightarrow$ Item (ii). Recall that for two square matrices the identity $AB = I$ is equivalent to $BA = I$ (which, in turn, is equivalent to B being invertible with $B = A^{-1}$). This shows Item (i) $\Leftrightarrow$ Item (iii) with $A = U$ and $B = U^*$. To see Item (iii) $\Leftrightarrow$ Item (iv), we simply write Item (iii) as $(U^*)^*U^* = I$ which by definition is unitarily of U^*.

A list of vectors is an orthonormal basis for $\mathbb{F}^n$ if and only if the list of vectors *complex conjugated* is an orthonormal basis for $\mathbb{F}^n$. The rows of U are the complex conjugates of the columns of U^*. Hence, from Item (i) $\Leftrightarrow$ Item (ii) we see that Item (iv) $\Leftrightarrow$ Item (v).

Assume now that Item (i) holds, that is, assume $U^*U = I$. Then we can compute, for $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{F}^n$,

$$\langle U\boldsymbol{x}, U\boldsymbol{y}\rangle = \langle U^*U\boldsymbol{x}, \boldsymbol{y}\rangle = \langle \boldsymbol{x}, \boldsymbol{y}\rangle,$$

which is Item (vi). Assume then Item (vi) holds. By the same computation as just before, we than have $\langle U^*U\boldsymbol{x}, \boldsymbol{y}\rangle = \langle \boldsymbol{x}, \boldsymbol{y}\rangle$. By Lemma 2.1.2 on page 46 we conclude that $U^*U\boldsymbol{x} = \boldsymbol{x}$ for all $\boldsymbol{x}$ which shows that Item (i) holds. This completes the proof.

■

Using the so-called polarization identities, one can show that the conditions in Theorem 2.6.1 are also equivalent to:

(iii) For all $\boldsymbol{x} \in \mathbb{F}^n$: $\|U\boldsymbol{x}\| = \|\boldsymbol{x}\|$.

From Items (iii) and (vi) we say that linear maps associated with unitary matrices *preserves* the length of vectors and angle between them.

Exercise 2.6.2

Consider the vector space $\mathbb{F}^n$ with the standard inner product $\langle \cdot, \cdot \rangle$, i.e., $\langle \boldsymbol{x}, \boldsymbol{y}\rangle = \boldsymbol{y}^*\boldsymbol{x}$. Suppose $U \in \mathsf{M}_n(\mathbb{C})$ is unitary and let $\boldsymbol{u}_k$ denote the kth column of U. As shown above $U^*U = I$ is just another way of stating that $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n$ is orthonormal. Show that $UU^* = I$ is just another way of writing

$$\boldsymbol{x} = \sum_{k=1}^{n} \langle \boldsymbol{x}, \boldsymbol{u}_k\rangle \boldsymbol{u}_k \qquad \text{for all } \boldsymbol{x} \in \mathbb{F}^n.$$

(Hint: $UU^* = I$ states that $UU^*\boldsymbol{x} = \boldsymbol{x}$ for all $\boldsymbol{x} \in \mathbb{F}^n$. Start by relating $U^*\boldsymbol{x}$ to $\langle \boldsymbol{x}, \boldsymbol{u}_k\rangle$.)

The equivalences in Theorem 2.6.1 can be formulated for real orthogonal matrices. Here, we focus on the most important equivalence.

Corollary 2.6.2

Let $Q \in \mathsf{M}_n(\mathbb{R})$. The following assertions are equivalent:

(i) Q is (real) orthogonal, that is, $Q^T Q = I$.

(ii) The columns of Q are an orthonormal basis of $\mathbb{R}^n$.

Proof. Since $Q \in \mathsf{M}_n(\mathbb{R})$ is real, it is unitary if and only if it is orthogonal. So, by Theorem 2.6.1, Q is orthogonal if and only if the list of the columns of Q is an orthonormal basis of $\mathbb{C}^n$. However, the column vectors are real vectors by assumption, so they also constitute an orthonormal basis of $\mathbb{R}^n$ as the standard inner product in $\mathbb{C}^n$ is identical to the dot product on $\mathbb{R}^n$ for real vectors.

■

Example 2.6.3

The $n \times n$ *Fourier matrix* F_n is a complex matrix $F_n = [f_{i,j}]_{i,j=0}^{n-1}$ whose (i, j) entry is given by

$$f_{i,j} = \frac{1}{\sqrt{n}} \omega_n^{ij} \qquad \text{where } \omega_n = \mathrm{e}^{2\pi i/n}$$

The complex number ω_n is called the nth root of unity (try to sketch it in the complex plane). It can be shown that F_n is unitary for every $n \in \mathbb{N}$. The column vectors of F_n are called the discrete Fourier basis of $\mathbb{C}^n$. The linear mapping

$$\mathrm{DFT} : \mathbb{C}^n \to \mathbb{C}^n, \mathrm{DFT} = \boldsymbol{x} \mapsto F_n^* \boldsymbol{x}$$

is the so-called *Discrete Fourier Transform* (DFT). If n is a power of 2, i.e., $n = 2^m$ for some m, it can be computed very fast[a] using an algorithm called the Fast Fourier Transform (FFT). This transform is extremely useful and has applications in all of engineering.

[a] To be precise: The speed of the FFT depends on the number of factors in the prime factorization of n. For $n = 2^m$ the algorithm uses a constant multiple of $n \log(n)$ floating-point operations, while standard matrix-vector multiplication uses Cn^2 operations.

Note that if $W \in \mathsf{M}_{k \times n}(\mathbb{F})$ has orthonormal columns, it is *only* unitary if it is square $k = n$. However, if $k < n$, that is, we have fewer orthonormal column vectors than needed to span $\mathbb{F}^n$, we can extend the matrix with additional columns so that the extended matrix becomes unitary. This fact is a consequence of the following simple result.

Lemma 2.6.3 extension to orthonormal basis

Suppose $\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_\ell$ is an orthonormal list of vectors in an inner product space V of dimension n. Then there are $n - \ell$ vectors $\boldsymbol{u}_{\ell+1}, \ldots, \boldsymbol{u}_n$ so that the combined list $\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_n$ is an orthonormal basis for V.

Proof. Let $\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_\ell$ is an orthonormal list. Pick $n - r$ vectors[3] $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{n-r}$ so that the combined list $\alpha := \boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_\ell, \boldsymbol{v}_1, \ldots, \boldsymbol{v}_{n-r}$ is linearly independent, that is, so that $\operatorname{span}\{\alpha\} = V$. Now, if we perform the Gram-Schmidt process on the list α, we obtain the desired orthonormal basis.

■

2.7 Diagonalizable matrices

Recall that two square matrices A and B in $\mathsf{M}_n(\mathbb{F})$ are called *similar* if there is an invertible matrix $S \in \mathsf{M}_n(\mathbb{F})$ such that $B = S^{-1}AS$. A complex square matrix is said to be diagonalizable (over $\mathbb{C}$) if and only if it is *similar* with a diagonal matrix. We are interested in the cases where S is not just invertible, but unitary. This leads to the following definitions.

Definition 2.7.1 Diagonalizable matrix

A matrix $A \in \mathsf{M}_n(\mathbb{C})$ is:

(i) *diagonalizable* if there is an invertible matrix $S \in \mathsf{M}_n(\mathbb{C})$ such that $S^{-1}AS$ is diagonal.

(ii) *unitarily diagonalizable* if there is a unitary matrix $U \in \mathsf{M}_n(\mathbb{C})$ such that U^*AU is diagonal.

(iii) *real orthogonally diagonalizable* if there is a real orthogonal matrix $Q \in \mathsf{M}_n(\mathbb{R})$ such that Q^TAQ is diagonal.

In Chapter 12 in Mathematics 1a it is shown that $A \in \mathsf{M}_n(\mathbb{C})$ is diagonalizable if and only if A has n linearly independent eigenvectors which in turn happens if and only if the algebraic and geometric multiplicities agree for all eigenvalues[4].

[3] It is almost obvious that this can be done. However, it does require a proof. The proof is "easy" and can be done by induction of the definition of linear independence. We leave the details to the reader.

[4] See Lemma 12.3.2 in Mathematics 1a and Corollary 12.3.5 in Mathematics 1a.

However, it may not be apparent directly from Definition 2.7.1 that the notion of diagonalibility of a matrix is closely related to eigenvectors and, in fact, the column vectors of S, U and Q in Definition 2.7.1 constitute a basis of eigenvectors. To see this, suppose that A is diagonalizable. Then $S^{-1}AS$ is diagonal for some invertible matrix S. Let us denote this diagonal matrix by $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$. From $S^{-1}AS = \Lambda$ we get $AS = S\Lambda$ by multiplying with S from the left. If we denote the ith column of S by $\boldsymbol{s}_i$, we get

$$AS = A\begin{bmatrix}\boldsymbol{s}_1 & \boldsymbol{s}_2 & \cdots & \boldsymbol{s}_n\end{bmatrix} = \begin{bmatrix}A\boldsymbol{s}_1 & A\boldsymbol{s}_2 & \cdots & A\boldsymbol{s}_n\end{bmatrix} \tag{2.26}$$

and

$$S\Lambda = \begin{bmatrix}\lambda_1\boldsymbol{s}_1 & \lambda_2\boldsymbol{s}_2 & \cdots & \lambda_n\boldsymbol{s}_n\end{bmatrix}. \tag{2.27}$$

Since the matrix-matrix products in (2.26) and (2.27) are equal, each column must be equal, thus:

$$A\boldsymbol{s}_i = \lambda_i\boldsymbol{s}_i, \qquad \text{for } i = 1, \ldots, n. \tag{2.28}$$

Hence, we see that $(\lambda_i, \boldsymbol{s}_i)$ is actually an eigenvalue-eigenvector pair of A. Since S is assumed to be invertible, its column vectors are linearly independent, hence they constitute a basis.

If A is either unitarily diagonalizable or real orthogonal diagonalizable, the same computations carries through. We just have to use that U^* and Q^T are inverse of a unitary matrix U and a real orthogonal matrix Q, respectively. Moreover, in these cases the eigenbasis is an *orthonormal basis* for $\mathbb{R}^n$ and $\mathbb{C}^n$, respectively, by Theorem 2.6.1.

2.8 The Spectral Theorem

The spectral theorem can be formulated for three classes of square matrices:

(i) real, symmetric matrices A, that is, matrices satisfying $A = A^T$ with real entries.

(ii) Hermitian matrices A, that is, matrices satisfying $A = A^*$ with complex entries.

(iii) normal matrices A, that is, matrices satisfying $AA^* = A^*A$ with complex entries.

Real, symmetric matrices are Hermitian matrices, and Hermitian matrices are normal matrices. These statements are easy to prove. E.g., if A is Hermitian, it satisfies $A = A^*$ and therefore also $AA^* = AA = A^*A$ which is the condition for being normal.

Example 2.8.1

Consider the 3×3 matrices A, B, C defined by

$$A = \begin{bmatrix} 2 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 2 \end{bmatrix}, \; B = \begin{bmatrix} 2 & 3+i & 1 \\ 3-i & 2 & -i \\ 1 & i & 2 \end{bmatrix}, \; C = \begin{bmatrix} 2 & 1 & 0 \\ 0 & 2 & 1 \\ 1 & 0 & 2 \end{bmatrix}.$$

The matrix A is real and symmetric; hence, A is also Hermitian and normal. The matrix B is Hermitian (and normal), but not symmetric nor real. The matrix C is neither symmetric nor Hermitian, but it is normal as can be seen by computing CC^* and C^*C and verifying that they agree.

The different versions of the spectral theorem are very similar so we will for brevity focus on the case of real and symmetric matrices. Let $A \in \mathsf{M}_n(\mathbb{R})$ be a symmetric matrix. The spectral theorem, Theorem 2.8.3, we will prove below, implies that A can be decomposed as the matrix product:

$$A = Q\Lambda Q^T, \tag{2.29}$$

where $Q \in \mathsf{M}_n(\mathbb{R})$ is an orthogonal matrix and Λ is a diagonal matrix. The formula (2.29) is called the *spectral decomposition* of A. From Section 2.7 we then know that Λ contains the eigenvalues of A and that the columns of Q contain the corresponding eigenvectors.

Before we prove the spectral decomposition, let us discuss how we can compute (2.29) given that the formula actually holds. So, let A be a real and symmetric matrix. We first compute the eigenvalues of A. Counted with algebraic multiplicity there are exactly n eigenvalues. Hence, we let $\lambda_1, \ldots, \lambda_n$ be the eigenvalues of A. Note that the value of λ_i in this list has to be repeated according to the algebraic multiplicity of the eigenvalue. We define $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$. As we will see, a real and symmetric matrix A has n linearly independent eigenvectors (recall that this is not the case for all $n \times n$ matrices). The list of eigenvectors should be ordered to match the eigenvalues so that the ith eigenvector has eigenvalue λ_i. After applying the Gram-Schmidt procedure to this list of n linearly independent eigenvectors, we end up with an orthonormal basis of eigenvectors $\boldsymbol{q}_1, \ldots, \boldsymbol{q}_n$. Setting $Q = [\boldsymbol{q}_1, \ldots, \boldsymbol{q}_n]$ provides us with the spectral decomposition $A = Q\Lambda Q^T$.

Lemma 2.8.1

If $A \in \mathsf{M}_n(\mathbb{C})$ is Hermitian, i.e., $A = A^*$, then all eigenvalues of A are real.

Proof. Let $\lambda \in \mathbb{C}$ be an eigenvalue of a Hermitian matrix A, and let $\boldsymbol{v} \in \mathbb{C}^n \setminus \{\mathbf{0}\}$ be an associated eigenvector. Hence, $A\boldsymbol{v} = \lambda\boldsymbol{v}$. Then, since $A = A^*$,

$$\lambda\langle \boldsymbol{v}, \boldsymbol{v}\rangle = \langle \lambda\boldsymbol{v}, \boldsymbol{v}\rangle = \langle A\boldsymbol{v}, \boldsymbol{v}\rangle = \langle \boldsymbol{v}, A^*\boldsymbol{v}\rangle = \langle \boldsymbol{v}, A\boldsymbol{v}\rangle = \langle \boldsymbol{v}, \lambda\boldsymbol{v}\rangle = \overline{\lambda}\langle \boldsymbol{v}, \boldsymbol{v}\rangle.$$

Since $\|\boldsymbol{x}\|^2 = \langle \boldsymbol{x}, \boldsymbol{x}\rangle$ for all vectors $\boldsymbol{x}$, we have shown that $\lambda\|\boldsymbol{v}\|^2 = \overline{\lambda}\|\boldsymbol{v}\|^2$. An eigenvector $\boldsymbol{v}$ cannot be the zero vector so $\|\boldsymbol{v}\| \neq 0$. Hence, we can devide through by $\|\boldsymbol{v}\|^2$, and we arrive at $\lambda = \overline{\lambda}$, which shows that λ is real. ■

Eigenvectors of a Hermitian matrix have the following orthogonality property: two eigenvectors associated with two *different* eigenvalues are orthogonal.

Exercise 2.8.2

Suppose $A \in \mathsf{M}_n(\mathbb{C})$ is Hermitian, i.e., $A = A^*$. Let λ_1 and λ_2 by eigenvalues of A with associated eigenvectors $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$, respectively. Show that if $\lambda_1 \neq \lambda_2$, then $\boldsymbol{u}_1 \perp \boldsymbol{u}_2$, i.e., $\langle \boldsymbol{u}_1, \boldsymbol{u}_2\rangle = 0$

A real and symmetric matrix is automatically Hermitian since $A = A^T = A^*$. Hence, real, symmetric matrices inherit the same properties as Hermitian matrix, in particular, they have real eigenvalues by Lemma 2.8.1. However, not only will a real and symmetric matrix have real eigenvalues, it will also have a real eigenvector:

Lemma 2.8.2

A real, symmetric matrix A has a real eigenvector.

Proof. The characteristic polynomial of A is a degree n polynomial and by the fundamental theorem of algebra, we know that it has at least one root λ_1. By Lemma 2.8.1 we know that λ_1 is real. The fact that λ_1 is an eigenvalue means that $A - \lambda_1 I$ is singular, hence there is a non-zero vector $\boldsymbol{q}_1$ such that $A\boldsymbol{q}_1 = \lambda_1\boldsymbol{q}_1$. We need to prove that we can find a *real* eigenvector of A associated with λ_1. For the (possibly complex) non-zero eigenvector $\boldsymbol{q}_1$ we write $\boldsymbol{q}_1 = \operatorname{Re}\boldsymbol{q}_1 + i\operatorname{Im}\boldsymbol{q}_1$, where we take the real and imaginary parts coordinate wise. The equation $A\boldsymbol{q}_1 = \lambda_1\boldsymbol{q}_1$ holds if and only if the real and

imaginary parts of the left and right hand side agrees. Since both A and λ_1 are real, we arrive at the two equations:

$$A \operatorname{Re} \boldsymbol{q}_1 = \lambda_1 \operatorname{Re} \boldsymbol{q}_1 \quad \text{and} \quad A \operatorname{Im} \boldsymbol{q}_1 = \lambda_1 \operatorname{Im} \boldsymbol{q}_1$$

Note that both $\operatorname{Re} \boldsymbol{q}_1$ and $\operatorname{Im} \boldsymbol{q}_1$ are real vectors and that they cannot both be the zero vector (as then $\boldsymbol{q}_1$ would be the zero vector). Hence, we have shown the existence of at least one real eigenvector.

■

Exercise 2.8.3

Find the eigenvalues and eigenvectors of the Hermitian matrix $A = \begin{bmatrix} 0 & i \\ -i & 0 \end{bmatrix}$. The eigenvalues of A are real by Lemma 2.8.1, but are there any real eigenvectors of A?

We are now finally ready to state and prove the spectral theorem for real, symmetric matrices.

Theorem 2.8.3 Spectral Theorem (the real case)

Let $A \in \mathsf{M}_n(\mathbb{R})$. The following assertions are equivalent:

(i) A is a symmetric matrix.

(ii) A is real orthogonally diagonalizable, that is, there is a real orthogonal matrix $Q \in \mathsf{M}_n(\mathbb{R})$ such that $Q^T A Q$ is diagonal.

(iii) $\mathbb{R}^n$ has an orthonormal basis consisting of eigenvectors of A.

Proof. (i) ⇒ (ii): Suppose A is real and symmetric. We need to show that A is real orthogonally diagonalizable. The proof is by induction on the size n of the matrix A.

The result is trivial for $n = 1$: The matrix $A \in \mathsf{M}_1(\mathbb{R})$ is just a scalar, i.e., $A = [a_{1,1}]$ for $a_{1,1} \in \mathbb{R}$. Hence, $\lambda_1 = a_{1,1}$ is the eigenvalue of A and with $Q = [1]$ or $Q = [-1]$, we have $A = Q \Lambda Q^T$. Note that both $Q = [1]$ or $Q = [-1]$ are real orthogonal 1×1 matrices.

Now, we turn to the induction step. So, let $n > 1$ and assume the implication (i) ⇒ (ii) is true for any matrix of size $n - 1$.

Let λ_1 be an eigenvalue of A. By Lemma 2.8.1 it is real, and by Lemma 2.8.2, we can find a real eigenvector $\boldsymbol{u}_1$ associated with λ_1, i.e., $A\boldsymbol{u}_1 = \lambda_1 \boldsymbol{u}_1$. If $\boldsymbol{u}_1$ is not normalized, i.e., of length one, we can replace it by $\boldsymbol{u}_1 / \|\boldsymbol{u}_1\|$ which is an eigenvector of norm one. Hence, we can assume that

$\|\boldsymbol{u}_1\| = 1$. Note that there is no risk of division by zero since eigenvectors are always non-zero.

By a result in Lemma 2.6.3 on page 66 we can extend $\boldsymbol{u}_1$ to an orthonormal basis for $\mathbb{R}^n$. The first element in this list is still $\boldsymbol{u}_1$ since it was a normalized vector, but the remaining $n-1$ vectors are not necessarily eigenvectors of A. We let the last $n-1$ vectors in the orthonormal basis be columns in an $n \times (n-1)$ matrix V_1. Then $[\boldsymbol{u}_1, V_1]$ is an orthogonal matrix since the columns are an orthonormal basis of $\mathbb{R}^n$.

The matrix $V_1^T A V_1$ is symmetric since it is equal to its transpose:

$$(V_1^T A V_1)^T = V_1^T A^T (V_1^T)^T = V_1^T A V_1 \tag{2.30}$$

So, by our induction hypothesis, the $(n-1) \times (n-1)$ symmetric matrix $V_1^T A V_1$ is real orthogonally diagonalizable, that is, there is a real ortogonal $(n-1) \times (n-1)$-matrix Q_1 such that $Q_1^T V_1^T A V_1 Q_1$ is diagonal. We denote this diagonal matrix by Λ_1, i.e.,

$$\Lambda_1 = Q_1^T V_1^T A V_1 Q_1. \tag{2.31}$$

We define the $n \times (n-1)$ matrix $U_1 := V_1 Q_1$. By construction the columns of U_1 are linear combinations of the columns of V_1; we can express this as $\operatorname{col} U_1 \subseteq \operatorname{col} V_1$ [5]. Since $\boldsymbol{u}_1$ is orthogonal to all columns in V_1, it is therefore also orthogonal to all columns of U_1. The conclusion that $\boldsymbol{u}_1$ is orthogonal to all columns in U_1 can be summarized as

$$U_1^T \boldsymbol{u}_1 = \mathbf{0}_{n-1}, \tag{2.32}$$

where $\mathbf{0}_{n-1}$ is the column vector of $n-1$ zeros. We claim that the matrix $U := [\boldsymbol{u}_1, U_1]$ is real orthogonal. To see this claim, we have to show that $U^T U = I$, but this follows from the following computation:

$$\begin{aligned} U^T U &= \begin{bmatrix} \boldsymbol{u}_1^T \\ U_1^T \end{bmatrix} \begin{bmatrix} \boldsymbol{u}_1 & U_1 \end{bmatrix} \\ &= \begin{bmatrix} \boldsymbol{u}_1^T \boldsymbol{u}_1 & \boldsymbol{u}_1^T U_1 \\ U_1^T \boldsymbol{u}_1 & U_1^T U_1 \end{bmatrix} = \begin{bmatrix} 1 & \mathbf{0}_{n-1}^T \\ \mathbf{0}_{n-1} & I_{n-1} \end{bmatrix} = I_n, \end{aligned}$$

since $U_1^T U_1 = Q_1^T V_1^T V_1 Q_1 = Q_1^T I_{n-1} Q_1 = I_{n-1}$.

From equation (2.32), we see that $U_1^T A \boldsymbol{u}_1 = \lambda_1 U_1^T \boldsymbol{u}_1 = \lambda_1 \mathbf{0}_{n-1} = \mathbf{0}_{n-1}$. Note also that Equation (2.31) reads $\Lambda_1 = U_1^T A U_1$. Using these relations, we can compute:

$$U^T A U = U^T A \begin{bmatrix} \boldsymbol{u}_1 & U_1 \end{bmatrix} = \begin{bmatrix} \boldsymbol{u}_1^T \\ U_1^T \end{bmatrix} \begin{bmatrix} A\boldsymbol{u}_1 & A U_1 \end{bmatrix}$$

[5] In fact, $\operatorname{col} U_1 = \operatorname{col} V_1$, but we will not need this property.

$$= \begin{bmatrix} \boldsymbol{u}_1^T A\boldsymbol{u}_1 & \boldsymbol{u}_1^T AU_1 \\ U_1^T A\boldsymbol{u}_1 & U_1^T AU_1 \end{bmatrix} = \begin{bmatrix} \lambda_1 & \mathbf{0}_{n-1}^T \\ \mathbf{0}_{n-1} & \Lambda_1 \end{bmatrix},$$

where in the last step we have used that $\boldsymbol{u}_1^T A\boldsymbol{u}_1 = \langle A\boldsymbol{u}_1, \boldsymbol{u}_1\rangle = \langle \lambda_1\boldsymbol{u}_1, \boldsymbol{u}_1\rangle = \lambda_1\|\boldsymbol{u}_1\|^2 = \lambda_1$, and

$$\boldsymbol{u}_1^T AU_1 = (U_1^T A\boldsymbol{u}_1)^T = \mathbf{0}_{n-1}^T$$

Hence, we have exhibited an orthogonal matrix $U \in \mathsf{M}_n(\mathbb{R})$ such that $U^T AU$ is diagonal. This completes the induction proof.

(ii) $\Rightarrow$ (iii): We assume that there is a real orthogonal matrix $Q \in \mathsf{M}_n(\mathbb{R})$ such that $Q^T AQ$ is diagonal $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$. Write $Q = [\boldsymbol{q}_1, \ldots, \boldsymbol{q}_n]$. As in (2.28), we then have:

$$A\boldsymbol{q}_i = \lambda_i \boldsymbol{q}_i, \qquad \text{for } i = 1, \ldots, n.$$

So $\boldsymbol{q}_1, \ldots, \boldsymbol{q}_n$ are eigenvectors of A. Since Q is real orthogonal, the list $\boldsymbol{q}_1, \ldots, \boldsymbol{q}_n$ is an orthonormal basis (of eigenvectors).
(iii) $\Rightarrow$ (i): Suppose that $\boldsymbol{q}_1, \ldots, \boldsymbol{q}_n$ are an orthonormal basis of $\mathbb{R}^n$ and let $\lambda_1, \ldots, \lambda_n$ be the associated eigenvalues, i.e.,

$$A\boldsymbol{q}_i = \lambda_i \boldsymbol{q}_i \text{ for } i = 1, \ldots, n. \tag{2.33}$$

Then $Q = [\boldsymbol{q}_1, \ldots, \boldsymbol{q}_n]$ is unitary, and we can write Equation (2.33) as matrix-matrix products as in Section 2.7:

$$AQ = Q\Lambda.$$

Since $Q^{-1} = Q^T$, we can express A as $A = Q\Lambda Q^T$. From this expression of A, it is easy to see that it is symmetric:

$$A^T = (Q\Lambda Q^T)^T = (Q^T)^T \Lambda^T Q^T = Q\Lambda Q^T = A.$$

■

Corollary 2.8.4

Let $A \in \mathsf{M}_n(\mathbb{R})$ be a symmetric matrix. Then *the spectral decomposition*

$$A = Q\Lambda Q^T$$

holds, where Λ is a diagonal matrix whose diagonal consists of the eigenvalues of A (repeated according to their algebraic multiplicity) and where $Q \in \mathsf{M}_n(\mathbb{R})$ is a real orthogonal matrix, whose columns are eigenvectors of A.

Proof. From 2.8.3 (ii) we have that Q^TAQ is diagonal. Denote this diagonal matrix by $\Lambda \coloneqq Q^TAQ$. Then

$$Q\Lambda Q^T = QQ^TAQQ^T = I_nAI_n = A$$

since Q is real orthogonal, i.e., $QQ^T = I$.

■

The spectral decomposition shows how we pick the simplest possible matrix representations of linear mappings defined through real, symmetric matrices by changing basis to the orthonormal eigenvector basis.

Example 2.8.4

Let $A \in \mathsf{M}_n(\mathbb{R})$ be a symmetric matrix and define the linear map $L_A : \mathbb{R}^n \to \mathbb{R}^n$ by $L_A = \boldsymbol{x} \mapsto A\boldsymbol{x}$, i.e., $L_A\boldsymbol{x} = A\boldsymbol{x}$.

If we let e denote the standard basis for $\mathbb{R}^n$, i.e., $e = \boldsymbol{e}_1, \ldots, \boldsymbol{e}_n$, then the matrix representation of L_A is simply ${}_e[L_A]_e = A$. However, the spectral decomposition provide us with a much simpler matrix representation of L_A as follows:

Let $A = Q\Lambda Q^T$ be the spectral decomposition of A, where $Q = [\boldsymbol{q}_1, \ldots, \boldsymbol{q}_n]$ contains the orthonormal basis of eigenvectors. Let $q \coloneqq \boldsymbol{q}_1, \ldots, \boldsymbol{q}_n$. Recall from Lemma 12.1.3 in Mathematics 1a:

$${}_e[L_A]_e = {}_eM_q\, {}_q[L_A]_q\, {}_qM_e,$$

where ${}_eM_q$ is the change-of-basis matrix from q-basis to e-basis. This is in fact just the spectral decomposition of A is disguise. We first note that $Q = {}_eM_q$ and $Q^T = Q^{-1} = {}_qM_e$. Hence, the spectral decomposition of A can be written

$${}_e[L_A]_e = A = Q\Lambda Q^T = {}_eM_q\, \Lambda\, {}_qM_e.$$

We conclude that ${}_q[L_A]_q = \Lambda$.

The spectral theorem can also be used to simplify the study of functions, in particular, quadratic forms. Consider a quadratic form $q : \mathbb{R}^n \to \mathbb{R}$. The following simple result says that we may always assume that the matrix $A \in \mathsf{M}_n(\mathbb{R})$ in Equation (1.3) on page 16 is *symmetric*.

Lemma 2.8.5

Let $q : \mathbb{R}^n \to \mathbb{R}$ be a quadratic form:

$$q(\boldsymbol{x}) = \boldsymbol{x}^TA\boldsymbol{x} + \boldsymbol{x}^T\boldsymbol{b} + c,$$

where $A \in \mathsf{M}_n(\mathbb{R})$, $\boldsymbol{b} \in \mathbb{R}^n$, and $c \in \mathbb{R}$. Let $B \coloneqq (A + A^T)/2$. Then B is symmetric and

$$q(\boldsymbol{x}) = \boldsymbol{x}^T B \boldsymbol{x} + \boldsymbol{x}^T \boldsymbol{b} + c$$

Exercise 2.8.5

Prove Lemma 2.8.5.

Example 2.8.6 Reduction of quadratic forms

Let $q : \mathbb{R}^n \to \mathbb{R}$ be a quadratic form. By Lemma 2.8.5 we can write q as

$$q(\boldsymbol{x}) = \boldsymbol{x}^T A \boldsymbol{x} + \boldsymbol{x}^T \boldsymbol{b} + c$$

where $A \in \mathsf{M}_n(\mathbb{R})$ is symmetric, $\boldsymbol{b} \in \mathbb{R}^n$, and $c \in \mathbb{R}$.

By Corollary 2.8.4, we have the spectral decomposition

$$A = Q \Lambda Q^T,$$

where Λ is a diagonal matrix whose diagonal consists of the eigenvalues of A (repeated according to their algebraic multiplicity) and where $Q \in \mathsf{M}_n(\mathbb{R})$ is a real orthogonal matrix. The column vectors $\boldsymbol{q}_i$ of Q constitute an orthonormal basis of $\mathbb{R}^n$, i.e., $\beta = \boldsymbol{q}_1, \boldsymbol{q}_2, \ldots, \boldsymbol{q}_n$ is an orthonormal basis and Q^T is the change-of-basis matrix from e-basis to standard β-basis.

We consider the column vector $\boldsymbol{x}$ as the coordinates of the vector $\boldsymbol{x}$ with respect to the e-basis. However, there is of course no difference on $\boldsymbol{x}$ and ${}_e\boldsymbol{x}$, but if we now define $\boldsymbol{y} = Q^T \boldsymbol{x}$ we see that $\boldsymbol{y} = {}_q\boldsymbol{x}$ represents the coordinates of $\boldsymbol{x}$ with respect to the eigenvector basis β. That is, given $\boldsymbol{x} \in \mathbb{R}^n$, we write as $\boldsymbol{x} = y_1 \boldsymbol{q}_1 + y_2 \boldsymbol{q}_2 + \ldots y_n \boldsymbol{q}_n$ instead of $\boldsymbol{x} = x_1 \boldsymbol{e}_1 + x_2 \boldsymbol{e}_2 + \ldots x_n \boldsymbol{e}_n$

Let us see why this change-of-basis is useful. Consider the second order terms:

$$\boldsymbol{x}^T A \boldsymbol{x} = \boldsymbol{x}^T Q \Lambda Q^T \boldsymbol{x} = \boldsymbol{y}^T \Lambda \boldsymbol{y} = \lambda_1 y_1^2 + \lambda_2 y_2^2 + \cdots + \lambda_n y_n^2,$$

hence there are no cross-terms of the form $y_i y_j$ $(i \neq j)$ in the β-basis.

We can therefore write q as

$$q(\boldsymbol{x}) = \lambda_1 y_1^2 + \lambda_2 y_2^2 + \cdots + \lambda_n y_n^2 + \boldsymbol{y}^T Q^T \boldsymbol{b} + c$$

where $\boldsymbol{y} = {}_q\boldsymbol{x}$. Hence, we can analyze quadratic forms in a simplified form:

$$\lambda_1 y_1^2 + \lambda_2 y_2^2 + \cdots + \lambda_n y_n^2 + \boldsymbol{y}^T \boldsymbol{b}' + c$$

where $\boldsymbol{b}' \coloneqq Q^T \boldsymbol{b}$. This reduced expression is exactly the quadratic forms considered in (1.18), for which we described its levelsets in Example 1.4.8 on page 32.

Finally, with a little more work, we can prove the complex version of the spectral theorem. We leave the proof to the reader.

Theorem 2.8.6 Spectral Theorem (the complex case)
Let $A \in \mathsf{M}_n(\mathbb{C})$. The following assertions are equivalent:

(i) A is a normal matrix.

(ii) A is unitarily diagonalizable, that is, there is a unitary matrix $U \in \mathsf{M}_n(\mathbb{C})$ such that U^*AU is diagonal.

(iii) $\mathbb{C}^n$ has an orthonormal basis consisting of eigenvectors of A.

Proof. Most of the proof of Theorem 2.8.3 generalize to the complex case. The only real difficulty is proving that $V_1^*AV_1$, see (2.30), is normal which is needed in order to apply our induction hypothesis. It requires a bit of work, and we leave the details to the reader.

■

From Theorem 2.8.6 we obtain a spectral decomposition of normal matrices A. The proof of the following result is essentially identical to the proof of Corollary 2.8.4 on page 72 so we do not repeat it.

Corollary 2.8.7
Let $A \in \mathsf{M}_n(\mathbb{C})$ be a normal matrix. Then *the spectral decomposition*

$$A = U\Lambda U^*$$

holds, where Λ is a diagonal matrix whose diagonal consists of the eigenvalues of A (repeated according to their algebraic multiplicity) and where $U \in \mathsf{M}_n(\mathbb{C})$ is a unitary matrix, whose columns are eigenvectors of A.

2.9 Positive definite and semi-definite matrices

Hermitian matrices have, as we have seen, real eigenvalues. Hermitian matrices with only *nonnegative* or *positive* eigenvalues are of particular interest, and they are the subject of this chapter. They can be seen as the matrix counterparts to nonnegative or positive real numbers. These matrices frequently appear in various fields: in statistics, they emerge as correlation matrices and in the normal equations for least squares fitting problems;

in Lagrangian mechanics, they represent the kinetic energy functional; in quantum mechanics, they are crucial as density matrices; and in mathematics, they characterize inner products.

Definition 2.9.1

Let $A \in \mathsf{M}_n(\mathbb{F})$.

(i) A is positive definite if A is Hermitian and $\langle A\boldsymbol{x}, \boldsymbol{x}\rangle > 0$ for all nonzero $\boldsymbol{x} \in \mathbb{F}^n$.

(ii) A is positive semi-definite if A is Hermitian and $\langle A\boldsymbol{x}, \boldsymbol{x}\rangle \geq 0$ for all $\boldsymbol{x} \in \mathbb{F}^n$.

(iii) A is negative definite if $-A$ is positive definite.

(iv) A is negative semi-definite if $-A$ is positive semi-definite.

Since $\langle A\mathbf{0}, \mathbf{0}\rangle = 0$, we see that a positive definite matrix A is also a positive semi-definite matrix. Note also that the inequalities in Definition 2.9.1 implicitly assume that $\langle A\boldsymbol{x}, \boldsymbol{x}\rangle$ is a real scalar. This potential issue is solved by assuming A is Hermitian since $\langle A\boldsymbol{x}, \boldsymbol{x}\rangle \in \mathbb{R}$ whenever A is Hermitian.

We already know several examples of positive semi-definite matrices. The zero $n \times n$ matrix and the projection matrix (from Example 2.3.2 on page 54) are positive semi-definite matrices, and the identity matrix is positive definite since $\langle \boldsymbol{x}, \boldsymbol{x}\rangle > 0$ for all nonzero $\boldsymbol{x} \in \mathbb{F}^n$. It is easy to construct other (and less trivial) positive semi-definite matrices as we will see in the next example.

Example 2.9.1

Let $B \in \mathsf{M}_{m\times n}(\mathbb{F})$. We claim that both $BB^* \in \mathsf{M}_m(\mathbb{F})$ and $B^*B \in \mathsf{M}_n(\mathbb{F})$ are positive semi-definite matrices. Let us prove this claim for $A = B^*B$. The matrix A is Hermitian since

$$A^* = (B^*B)^* = B^*(B^*)^* = B^*B = A.$$

From the computation, using that the norm is always non-negative,

$$\langle A\boldsymbol{x}, \boldsymbol{x}\rangle = \langle B^*B\boldsymbol{x}, \boldsymbol{x}\rangle = \langle B\boldsymbol{x}, (B^*)^*\boldsymbol{x}\rangle = \langle B\boldsymbol{x}, B\boldsymbol{x}\rangle = \|B\boldsymbol{x}\|^2 \geq 0,$$

we see that A is a positive semi-definite matrix.

The following characterization of positive definite (positive semi-definite) is a consequence of the Spectral Theorem. There are actually two theorems stated in Theorem 2.9.1: the equivalence holds either for all the statements in the parentheses or for all the statements without the parentheses.

Theorem 2.9.1

Let $A \in \mathsf{M}_n(\mathbb{F})$ be Hermitian. Then the following are equivalent:

(i) A is positive definite (positive semi-definite)

(ii) The eigenvalues of A are positive (non-negative)

(iii) There exists a positive (non-negative) constant c such that $\langle A\boldsymbol{x}, \boldsymbol{x}\rangle \geq c\langle \boldsymbol{x}, \boldsymbol{x}\rangle$ for all $\boldsymbol{x} \in \mathbb{F}^n$.

Proof. We will only prove that A is positive semi-definite if and only if A is Hermitian and all its eigenvalues are non-negative and leave the rest of the proof to the reader. If A is positive semi-definite, then A is Hermitian, and for an eigenvalue-eigenvector pair $(\lambda, \boldsymbol{u})$, where we have normalizes $\boldsymbol{u}$, we have

$$\lambda = \lambda\langle \boldsymbol{u}, \boldsymbol{u}\rangle = \langle \lambda\boldsymbol{u}, \boldsymbol{u}\rangle = \langle A\boldsymbol{u}, \boldsymbol{u}\rangle \geq 0.$$

Suppose A is Hermitian with non-negative eigenvalues $\lambda_1, \ldots, \lambda_n$. Let $A = U\Lambda U^*$ be the spectral decomposition of A, where U is a unitary matrix and $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$. Define $B = U\Gamma U^*$, where $\Gamma = \operatorname{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_n})$. Then

$$B^*B = (U\Gamma U^*)^* U\Gamma U^* = U\Gamma^* U^* U\Gamma U^* = U\Gamma^2 U^* = U\Lambda U^* = A$$

From Example 2.9.1 it follows that A is positive semi-definite.

■

If $A \in \mathsf{M}_n(\mathbb{F})$ is Hermitian, then it is normal. From the Spectral Theorem it follows that $A = U\Lambda U^*$ for some unitary matrix U and diagonal matrix Λ. Theorem 2.9.1 tells us that the matrix A is positive definite precisely when all the eigenvalues in the diagonal of Λ are positive.

2.10 Projections onto a subspace

This section is optional and can be skipped as it will not be used in the rest of the text. However, it is included since projections onto subspaces are important in many applications, e.g., in statistics, principal component analysis, and dimensionality reduction. In dimensionality reduction, one

typically has data living in a high dimensional space, e.g., $\mathbb{F}^n$ where n is large, but one believes that the data fits (or approximately fits) within a much lower dimensional subspace. To extract the essential information from the data, we therefore need to find a way to project the data onto this lower dimensional subspace. This is done to reduce the complexity of the data, e.g., to minimize the number of features or parameters needed in a model.

We have already studied projections onto a line in Section 2.3 on page 53. In this section, we will generalize this concept to projections onto subspaces of V. So, let Y be a non-trivial subspace of an inner product space V. Let ℓ denote the dimension of Y, and let $\alpha = \boldsymbol{u}_1, \ldots, \boldsymbol{u}_\ell$ be an orthonormal basis of Y, found, e.g., by the Gram-Schmidt process. The *orthogonal projection* $\mathrm{proj}_Y : V \to V$ of a vector $\boldsymbol{x} \in V$ onto the subspace Y is defined as:

$$\begin{aligned} \mathrm{proj}_Y(\boldsymbol{x}) &= \sum_{k=1}^{\ell} \mathrm{proj}_{\boldsymbol{u}_k}(\boldsymbol{x}) \\ &= \sum_{k=1}^{\ell} \langle \boldsymbol{x}, \boldsymbol{u}_k \rangle \boldsymbol{u}_k. \end{aligned} \tag{2.34}$$

The projection is a linear map on V since it is a sum of linear maps, and the image of proj_Y indeed is Y, i.e., $\mathrm{im}(\mathrm{proj}_Y) = Y$. It is important to use an orthonormal basis for Y when computing the projection in (2.34) as, otherwise, the expression will not be valid.

For $V = \mathbb{F}^n$ the orthogonal projection onto a subspace Y can be computed by using the projection matrix P onto Y as follows:

$$\mathrm{proj}_Y(\boldsymbol{x}) = P\boldsymbol{x} \quad \text{with } P \coloneqq UU^*,$$

where $U = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_\ell]$ is the $n \times \ell$ matrix whose columns are the orthonormal basis vectors of Y. To see why this holds, we can follow the reasoning in Example 2.3.2 on page 54, but we will not repeat it here. We simply mention that the projection matrix P is Hermitian (i.e., $P^* = P$) and idempotent (i.e., $P^2 = P$), and $\mathrm{col}\, P = Y$.

Let us show by an example how such a projection matrix can be computed.

Exercise 2.10.1

Let Y be the subspace of $\mathbb{R}^3$ spanned by the column vectors of $V \in \mathsf{M}_{3\times 2}(\mathbb{R})$ given by

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

In Example 2.5.1 on page 59 we found an orthonormal basis $\boldsymbol{u}_1, \boldsymbol{u}_2$ for Y by applying the Gram-Schmidt process to the columns of V:

$$U := \begin{bmatrix} \boldsymbol{u}_1 & \boldsymbol{u}_2 \end{bmatrix} = \begin{bmatrix} \frac{1}{3} & -\frac{2}{3} \\ \frac{2}{3} & \frac{2}{3} \\ \frac{2}{3} & -\frac{1}{3} \end{bmatrix}.$$

The projection matrix P onto Y is then given by

$$P = UU^* = UU^T = \begin{bmatrix} \frac{1}{3} & -\frac{2}{3} \\ \frac{2}{3} & \frac{2}{3} \\ \frac{2}{3} & -\frac{1}{3} \end{bmatrix} \begin{bmatrix} \frac{1}{3} & \frac{2}{3} & \frac{2}{3} \\ -\frac{2}{3} & \frac{2}{3} & -\frac{1}{3} \end{bmatrix} = \begin{bmatrix} \frac{5}{9} & -\frac{4}{9} & -\frac{2}{9} \\ -\frac{4}{9} & \frac{8}{9} & \frac{4}{9} \\ -\frac{2}{9} & \frac{4}{9} & \frac{5}{9} \end{bmatrix}.$$

Hence, to compute the projection $\mathrm{proj}_Y(\boldsymbol{x})$ of a vector in $\mathbb{R}^3$, say $\boldsymbol{x} = [1, 2, 3]^T$, onto Y, we simply have to compute the matrix-vector product $P\boldsymbol{x}$.

We finally remark that P can also be computed as $P = \boldsymbol{u}_1\boldsymbol{u}_1^T + \boldsymbol{u}_2\boldsymbol{u}_2^T$ (check it!).

CHAPTER 3

Continuity and Differentiability

The current chapter forms the core of the book, and we will see that the introduced topics play crucial roles in the later chapters.

We expect the reader is familiar with calculus of scalar functions of one variable, but we will briefly review some of the theory in Section 3.1 for the reader's convenience. As for functions of a single real variable, the topics of continuity and differentiability are central in the analysis of functions of several variables. We start by considering continuity in Section 3.2; the good news here is that having defined continuity of functions of one variable in a precise way, we can easily extend the definition to functions of several variables. Section 3.3 introduces the concept of partial derivatives for functions of several variables, in much the same way as differentiability is defined for functions of one variable. This is generalized in Section 3.5, where we consider partial derivatives of higher order and the so-called Hessian matrix.

The material in Sections 3.3 to 3.5 unfortunately shows that some of the standard results for functions of one variable do not immediately generalize to functions of several variables. For this reason we introduce the topic of differentiability in Section 3.6. Section 3.7 discusses the so-called chain-rule, an important result that generalizes the well-known results for differentiation of a composite function to the case of functions of several variables. Finally, in Section 3.8, we generalize the previous discussions of differentiability to vector functions of several variables.

3.1 Analysis of functions of one variable

The mathematical analysis of functions is based on the concepts *continuity* and *differentiability*. We first consider functions of the form $f : I \to \mathbb{R}$,

where either $I = \mathbb{R}$ or I is an open interval of the form $]-\infty, a[$, $]a, \infty[$ or $]a, b[$ for some $a, b \in \mathbb{R}$.

Definition 3.1.1 Continuity
A function $f : I \to \mathbb{R}$ is continuous at $x_0 \in I$ if

$$f(x) \to f(x_0) \text{ whenever } x \to x_0. \tag{3.1}$$

For short, the function f is said to be *continuous* if it is continuous at all $x_0 \in I$.

The condition (3.1) can alternatively be stated as

$$|f(x) - f(x_0)| \to 0 \text{ whenever } |x - x_0| \to 0. \tag{3.2}$$

It is phrased as "$f(x)$ tends to $f(x_0)$ if x tends to x_0". Intuitively, this means that the function values $f(x)$ are "close" to the function value $f(x_0)$ for points x that are "close" to x_0. The mathematically exact way of formulating the definition is:

For any given $\epsilon > 0$ there exists a $\delta > 0$ such that if $x \in I$ satisfies $|x - x_0| < \delta$ then $|f(x) - f(x_0)| < \epsilon$ holds.

We can write this much shorter using the so-called universal quantifier $\forall$ (which is read "for all") and the existential quantifier $\exists$ (which is read "there exists").

$$\forall \epsilon > 0 \exists \delta > 0 \forall x \in I : |x - x_0| < \delta \Rightarrow |f(x) - f(x_0)| < \epsilon. \tag{3.3}$$

A function is, on the other hand, said to be *discontinuous* at a point x_0 if it is not continuous at this point. Negating the condition (3.3) gives the following equivalent statement for discontinuity of f at x_0:

$$\exists \epsilon > 0 \forall \delta > 0 \exists x \in I : |x - x_0| < \delta \text{ and } |f(x) - f(x_0)| \geq \epsilon. \tag{3.4}$$

We should read this as "there exists an $\epsilon > 0$ such that for all $\delta > 0$ there exists an x such that $|x - x_0| < \delta$ and $|f(x) - f(x_0)| \geq \epsilon$".

It is well-known that, e.g., polynomials, exponential functions, and trigonometric functions are continuous within their domains. Let us here show that x^k, in fact, is continuous for any $k \in \mathbb{N}$.

Example 3.1.1

The polynomial function $f(x) = x^k$ is continuous on $\mathbb{R}$ for all $k \in \mathbb{N}$. We have to show that for every $x_0 \in \mathbb{R}$ and every $\epsilon > 0$, there exists $\delta > 0$ such that

$$|x - x_0| < \delta \quad \Rightarrow \quad |x^k - x_0^k| < \epsilon.$$

We start by rewriting the difference $x^k - x_0^k$ using the factorization:

$$x^k - x_0^k = (x - x_0)\Big(x^{k-1} + x^{k-2}x_0 + \cdots + x_0^{k-1}\Big).$$

Taking absolute values yields

$$|x^k - x_0^k| = |x - x_0| \cdot \Big| x^{k-1} + x^{k-2}x_0 + \cdots + x_0^{k-1}\Big|. \tag{3.5}$$

We will assume that x is close to x_0, i.e., $|x - x_0| < 1$. We set $M = |x_0| + 1$ Then, for $|x - x_0| < 1$, we have

$$|x| \le |x_0| + |x - x_0| < |x_0| + 1 = M.$$

Each term in the sum (3.5) can be estimated as follows:

$$|x^{k-1}| \le |x|^{k-1} \le M^{k-1},$$

$$|x^{k-2}x_0| \le |x|^{k-2}\,|x_0| \le M^{k-2}\,|x_0| \le M^{k-1},$$

and so on. In total, there are k terms, each bounded above by M^{k-1}. Consequently,

$$\Big| x^{k-1} + x^{k-2}x_0 + \cdots + x_0^{k-1}\Big| \le k\,M^{k-1}.$$

Combining the factorization and the bound on the sum, for any x with $|x - x_0| < 1$, we get

$$|x^k - x_0^k| = |x - x_0| \cdot \Big| x^{k-1} + \cdots + x_0^{k-1}\Big| \le |x - x_0|\,k\,M^{k-1}.$$

Let $\epsilon > 0$ be given. We want

$$|x^k - x_0^k| < \epsilon.$$

From the above estimate, it suffices to show that

$$|x - x_0|\,k\,M^{k-1} < \epsilon.$$

Therefore, it is enough to require

$$|x - x_0| < \frac{\epsilon}{k\, M^{k-1}},$$

and also $|x - x_0| < 1$. Thus we define

$$\delta = \min\left\{ 1,\ \frac{\epsilon}{k\, M^{k-1}} \right\} = \min\left\{ 1,\ \frac{\epsilon}{k(|x_0| + 1)^{k-1}} \right\}.$$

Then, whenever $|x - x_0| < \delta$, both conditions $|x - x_0| < 1$ and $|x - x_0| < \epsilon/(k\, M^{k-1})$ hold, leading to

$$|x^k - x_0^k| < \epsilon.$$

which is what we wanted to show.

From Example 3.1.1 above, we see that proving a function's continuity can require quite lengthy arguments. Fortunately, standard algebraic operations on continuous functions yield continuous functions. Specifically,

- If f and g are continuous functions defined on the same domain, the functions $f + g$ and fg are continuous.
- If f and g are continuous functions and the range of f is contained in the domain of g, the composite function $h(x) \coloneqq g(f(x))$ is continuous. Recall that we denote the composition by $h = g \circ f$.

Thus, in many cases, continuity of a given function can be checked without referring explicitly to Definition 3.1.1.

Example 3.1.2

Consider the function $f : \mathbb{R} \to \mathbb{R}$ given by

$$f(x) = x^2 + \cos(x^4),\ x \in \mathbb{R}.$$

We showed above in Example 3.1.1 that the function x^k is continuous on $\mathbb{R}$. Assuming that we also know that $\cos(x)$ is continuous[a] on $\mathbb{R}$, we can argue that the composition $\cos(x^4)$ is continuous. The sum of two continuous functions is also continuous so $x^2 + \cos(x^4)$ is continuous as a function of x. We conclude that f is continuous on $\mathbb{R}$.

[a] Unless stated otherwise, we will always assume that continuity and differentiability properties of standard functions, such as polynomial functions, exponential functions, and trigonometric functions, are known.

Definition 3.1.2 Differentiability

A function $f : I \to \mathbb{R}$ is differentiable at $x_0 \in I$ if the expression

$$\frac{f(x_0 + h) - f(x_0)}{h} \tag{3.6}$$

has a limit as $h \to 0$; if this is the case, the limit is denoted by $f'(x_0)$, i.e.,

$$f'(x_0) = \lim_{h \to 0} \frac{f(x_0 + h) - f(x_0)}{h}. \tag{3.7}$$

For short, the function f is said to be differentiable if it is differentiable at all $x_0 \in I$. If f is differentiable and $f'(x_0)$ depends continuously on x_0, the function f is said to be continuously differentiable. The value $f'(x_0)$ is called the derivative of the function f at the point x_0.

The derivative $f'(x_0)$ is the slope of the tangent line to the graph of f at the point $(x_0, f(x_0))$. Differentiability of a function at a point x_0 implies continuity at x_0; we will prove later in Theorem 3.6.4 (see also Theorem 3.8.2) that this also holds for functions of several variables.

As for continuity, the concept of differentiability behaves in a convenient way regarding the standard mathematical operations: a sum, a product, or a composition of two differentiable functions $f, g : \mathbb{R} \to \mathbb{R}$ is again differentiable, and the derivatives are given by:

$$(f + g)'(x) = f'(x) + g'(x), \tag{3.8}$$

$$(fg)'(x) = f(x)g'(x) + f'(x)g(x), \tag{3.9}$$

and

$$(g \circ f)'(x) = g'(f(x))f'(x). \tag{3.10}$$

Example 3.1.3

Let $c \in \mathbb{R}$ denote any constant, and consider the function given as

$$h : \mathbb{R} \to \mathbb{R}, \ h(x) = \sin(x^2 + c).$$

Then $h(x) = (g \circ f)(x)$ with $g(x) = \sin(x)$ and $f(x) = x^2 + c$. Thus h is differentiable, and

$$h'(x) = 2x \cos(x^2 + c).$$

Note that in order to check that a function $f : I \to \mathbb{R}$ is differentiable, we need to examine differentiability at each single point $x_0 \in I$. There exist functions that are differentiable at certain points, but not in all points:

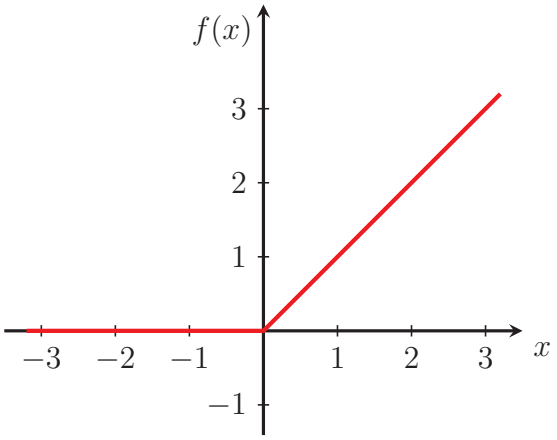


Figure 3.1: The graph of the function $f(x) = \mathrm{ReLU}(x)$.

Example 3.1.4

Let us consider the differentiability of the ReLU introduced in Equation (1.7). The function ReLU is differentiable for all $x \neq 0$, but not for $x = 0$; see Figure 3.1. To prove that ReLU is not differentiable at $x_0 = 0$, we compute:

$$\frac{f(x_0 + h) - f(x_0)}{h} = \frac{\mathrm{ReLU}(h)}{h} = \begin{cases} 1 & \text{for } h > 0, \\ 0 & \text{for } h < 0. \end{cases}$$

The limit of this expression as $h \to 0$ does not exist, and hence ReLU is not differentiable at $x_0 = 0$.

Vector-valued functions of one variable. We now turn to *vector* functions of one variable. These are functions of the form $\boldsymbol{f} : I \to \mathbb{R}^k$, where I is an interval of $\mathbb{R}$. Such functions are used to parameterize curves in $\mathbb{R}^k$ as we discussed in Item (e) on page 21. We have indeed already met curves, e.g., in our discussion of the boundary of a set in $\mathbb{R}^2$, see Examples 2.2.4 and 2.2.5. Here are a few more illustrative examples to keep in mind. As is typical when considering curves, we use the symbols $\boldsymbol{r}$ and $t \in I$ instead of $\boldsymbol{f}$ and $x \in I$.

Example 3.1.5 (a) Consider the function

$$\boldsymbol{r} : [0, 2] \to \mathbb{R}^2, \ \boldsymbol{r}(t) := \begin{bmatrix} t \\ t \end{bmatrix} = t \begin{bmatrix} 1 \\ 1 \end{bmatrix}. \tag{3.11}$$

Note that by letting the parameter t traverse through the interval $[0, 2]$, the vectors $\boldsymbol{r}(t)$ traverse through the straight line starting at the point $\boldsymbol{r}(0) = (0, 0)$ and ending at the point $\boldsymbol{r}(2) = (2, 2)$.

(b) Similarly, we can consider the function

$$\boldsymbol{r} : [0, \pi] \to \mathbb{R}^2,\ \boldsymbol{r}(t) \coloneqq \begin{bmatrix} \cos t \\ \sin t \end{bmatrix}. \tag{3.12}$$

When we let the parameter t traverse through the interval $[0, \pi]$, the vectors $\boldsymbol{r}(t)$ traverse through "half the unit circle", starting at the point $\boldsymbol{r}(0) = (1, 0)$ and ending at the point $\boldsymbol{r}(\pi) = (-1, 0)$.

More generally, the concept of a function $\boldsymbol{r} : I \to \mathbb{R}^k$ in a natural fashion leads to the concept of a *curve*:

Definition 3.1.3 Curve in $\mathbb{R}^k$

Let I denote an interval in $\mathbb{R}$, and consider a function $\boldsymbol{r} : I \to \mathbb{R}^k$. The image of $\boldsymbol{r}$

$$\operatorname{im}(\boldsymbol{r}) = \{\, \boldsymbol{r}(t) \mid t \in I \,\} \tag{3.13}$$

is called a curve in $\mathbb{R}^k$. The function $\boldsymbol{r}$ is said to provide a parametrization of the curve, and the variable $t \in I$ is called the parameter of $\boldsymbol{r}$. Curves will often be denoted by $\mathcal{C}$.

Note that a curve $\mathcal{C}$ can be parameterized in different ways; that is, a set of the form (3.13) will occur by different choices of the function $\boldsymbol{r}$. For example, the function

$$\boldsymbol{r} : [0, 1] \to \mathbb{R}^2,\ \boldsymbol{r}(t) \coloneqq \begin{bmatrix} 2t \\ 2t \end{bmatrix} \tag{3.14}$$

yields exactly the same set of points in $\mathbb{R}^2$ as the function in (3.11) — we just "travel through the curve with double speed". When we speak about a curve, we will always consider a *fixed* choice of a parametrization.

As in Equation (1.5) on page 18 we write $\boldsymbol{r}(t) = (r_1(t), r_2(t), \ldots, r_k(t))$ using the coordinate functions $r_i : I \to \mathbb{R}$. The coordinate functions are scalar-valued functions of one variable, which we can analyze using the result presented in the current section. E.g., if each coordinate function is differentiable, we can differentiate $\boldsymbol{r}$ by $\boldsymbol{r}'(t) \coloneqq (r_1'(t), r_2'(t), \ldots, r_k'(t))$ for $t \in I$.

Definition 3.1.4 Tangent vector

Let $\boldsymbol{r} : I \to \mathbb{R}^k$. We say that $\boldsymbol{r}$ is (continuously) differentiable if all the functions $r_1, r_2, \ldots, r_k$ are (continuously) differentiable[a]. The vector $\boldsymbol{r}'(t) = (r_1'(t), r_2'(t), \ldots, r_k'(t))$ is the *tangent vector* at $\boldsymbol{r}(t)$. If $\boldsymbol{r}'(t) \neq \boldsymbol{0}$ for all $t \in I$, the parametrization $\boldsymbol{r}$ is *regular*.

[a] If I is not an open interval, we assume that $r_1, r_2, \ldots, r_k$ are continuously differentiable on an open interval containing I

Example 3.1.6

We consider the parametrization of the half circle in (3.12). Since $\boldsymbol{r}'(t) = (-\sin t, \cos t) \neq (0,0)$ for all $t \in [0, \pi]$, the parametrization is continuously differentiable and regular. Fix $t_0 \in I$. The *tangent line* or just tangent at $\boldsymbol{r}(t_0)$ can then be written as

$$\boldsymbol{r}(t_0) + s\boldsymbol{r}'(t_0) = (\cos t_0, \sin t_0) + s(-\sin t_0, \cos t_0), \quad s \in \mathbb{R}.$$

E.g., with $t_0 = \pi/2$, we find the tangent at the mid-point $\boldsymbol{r}(\pi/2) = (0,1)$ to be $(0,1) + s(-1,0) = (-s, 1)$ for $s \in \mathbb{R}$.

In Definition 3.1.3, the starting point is the function $\boldsymbol{r}$ from which, we obtain a curve. In applications we often face the opposite scenario, where a curve is given, and we subsequently want to find a parametrization. In the next example we describe a parametrization of the straight line connecting two given points in $\mathbb{R}^k$.

Example 3.1.7

Given arbitrary points $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^k$, the straight line starting at $\boldsymbol{a}$ and ending at $\boldsymbol{b}$ can be parameterized as

$$\boldsymbol{r} : [0,1] \to \mathbb{R}^k, \quad \boldsymbol{r}(t) = \boldsymbol{a} + t(\boldsymbol{b} - \boldsymbol{a}) = (1-t)\boldsymbol{a} + t\boldsymbol{b}.$$

3.2 Continuity of vector functions of several variables

We are now ready to start the approach of generalizing the key concepts in mathematical analysis from functions of one variable to functions of n variables. We begin with continuity.

Recall the way we described continuity for a function of one variable, $f : I \to \mathbb{R}$, in (3.2). Here, the term $|x - x_0|$ measures the distance from the

point $x \in I$ to the fixed point $x_0 \in I$. In $\mathbb{R}^n$ we measure the distance between two points $\boldsymbol{x}$ and $\boldsymbol{x}_0$ by the norm $\|\boldsymbol{x} - \boldsymbol{x}_0\|$ introduced in Equation (2.1) on page 36. In fact, the norm $\|\cdot\|$ on $\mathbb{R}^n$ for $n = 1$ is just the absolute value. Therefore, continuity for a function of n variables can be defined precisely as we did for functions of one variable, simply by replacing $|x - x_0|$ by $\|\boldsymbol{x} - \boldsymbol{x}_0\|$, where $\boldsymbol{x}$ and $\boldsymbol{x}_0$ now are vectors. Similarly, we replace the $|f(x) - f(x_0)|$ by $\|\boldsymbol{f}(x) - \boldsymbol{f}(x_0)\|$. The precise definition reads as follows:

Definition 3.2.1 Continuity of vector functions

Let A denote a set in $\mathbb{R}^n$. A vector function $\boldsymbol{f} : A \to \mathbb{R}^k$ is continuous at $\boldsymbol{x}_0 \in A$ if

$$\|\boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{f}(\boldsymbol{x}_0)\| \to 0 \quad \text{whenever} \quad \|\boldsymbol{x} - \boldsymbol{x}_0\| \to 0. \tag{3.15}$$

We will also write the condition (3.15) as

$$\boldsymbol{f}(\boldsymbol{x}) \to \boldsymbol{f}(\boldsymbol{x}_0) \quad \text{whenever} \quad \boldsymbol{x} \to \boldsymbol{x}_0. \tag{3.16}$$

The function $\boldsymbol{f}$ is continuous if it is continuous at all $\boldsymbol{x}_0 \in A$.

Definition 3.2.1 is a proper generalization of the definition for a function of one variable. Indeed, if we write $\boldsymbol{x} = (x_1, \ldots, x_n)$ and $\boldsymbol{x}_0 = (x_{1,0}, \ldots, x_{n.0})$, then

$$\|\boldsymbol{x} - \boldsymbol{x}_0\| = \sqrt{(x_1 - x_{1,0})^2 + \cdots + (x_n - x_{n,0})^2};$$

in the case $n = 1$ this corresponds to $||\boldsymbol{x} - \boldsymbol{x}_0|| = \sqrt{(x_1 - x_{1,0})^2} = |x_1 - x_{1,0}|$, which is the expression appearing in (3.2).

As for a function of one variable, the condition (3.16) means that the function values $\boldsymbol{f}(\boldsymbol{x})$ are "close" to the function value $\boldsymbol{f}(\boldsymbol{x}_0)$ for vectors $\boldsymbol{x}$ that are "close" to $\boldsymbol{x}_0$. The mathematically exact way of phrasing this is:

For any given $\epsilon > 0$ there exists a $\delta > 0$ such that

$$\forall \boldsymbol{x} \in A : ||\boldsymbol{x} - \boldsymbol{x}_0|| < \delta \Rightarrow \|\boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{f}(\boldsymbol{x}_0)\| < \epsilon. \tag{3.17}$$

Going from continuity of scalar functions to vector functions is not a real complication. In fact, we can check continuity of a vector function by looking at each coordinate separately:

Theorem 3.2.1 Continuity of coordinate functions

Let A denote a set in $\mathbb{R}^n$. A vector function $\boldsymbol{f} = (f_1, \ldots, f_k) : A \to \mathbb{R}^k$ is continuous at $\boldsymbol{x}_0$ if and only if all the functions $f_i : A \to \mathbb{R}$, $i = 1, \ldots, k$ are

continuous at $\boldsymbol{x}_0$.

Proof. For $i = 1, \ldots, k$ we compute:

$$\begin{aligned}|f_i(\boldsymbol{x}) - f_i(\boldsymbol{x}_0)| &= \sqrt{|f_i(\boldsymbol{x}) - f_i(\boldsymbol{x}_0)|^2} \\ &\leq \sqrt{|f_1(\boldsymbol{x}) - f_1(\boldsymbol{x}_0)|^2 + \cdots + |f_k(\boldsymbol{x}) - f_k(\boldsymbol{x}_0)|^2} = \|\boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{f}(\boldsymbol{x}_0)\|.\end{aligned}$$

Thus, if $\|\boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{f}(\boldsymbol{x})\| < \epsilon$, then $|f_i(\boldsymbol{x}) - f_i(\boldsymbol{x}_0)| < \epsilon$. It follows from Equation (3.17) on the previous page that if $\boldsymbol{f} = (f_1, \ldots, f_k) : A \to \mathbb{R}^k$ is continuous at $\boldsymbol{x}_0$, then so is f_i for each i.

Conversely, assume all the coordinate functions f_i are continuous at $\boldsymbol{x}_0$ and we are given an $\epsilon > 0$. For each functions f_i we can find a $\delta_i > 0$ such that $\|\boldsymbol{x} - \boldsymbol{x}_0\| < \delta_i \Rightarrow |f_i(\boldsymbol{x}) - f_i(\boldsymbol{x}_0)| < \epsilon/k$. We now put $\delta = \min\{\delta_1, \ldots, \delta_k\}$. If $\|\boldsymbol{x} - \boldsymbol{x}_0\| < \delta$ then $|f_i(\boldsymbol{x}) - f_i(\boldsymbol{x}_0)| < \epsilon/m$ for all $i = 1, \ldots, k$ and hence

$$\begin{aligned}\|\boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{f}(\boldsymbol{x}_0)\| &= \sqrt{|f_1(\boldsymbol{x}) - f_1(\boldsymbol{x}_0)|^2 + \cdots + |f_k(\boldsymbol{x}) - f_k(\boldsymbol{x}_0)|^2} \\ &< \sqrt{k\left(\frac{\epsilon}{k}\right)^2} = \frac{\epsilon}{\sqrt{k}} < \epsilon.\end{aligned}$$

■

Example 3.2.1

Recall that, by Example 3.1.1 on page 82, the polynomial function $x \mapsto x^k$, $\mathbb{R} \to \mathbb{R}$, is continuous for each $k \in \mathbb{N}$.

(a) The functions

$$\boldsymbol{f} : \mathbb{R} \to \mathbb{R}^2,\ \boldsymbol{f}(x) = [x^k, 0]^T \ \text{ and } \ \boldsymbol{g} : \mathbb{R} \to \mathbb{R}^2,\ \boldsymbol{g}(x) = [0, x^k]^T$$

are also continuous. This follows from Theorem 3.2.1 since the two coordinate functions $x \mapsto x^k$ and $x \mapsto 0$ are continuous.

(b) Consider the functions

$$f : \mathbb{R}^2 \to \mathbb{R},\ f(x, y) = x^k \ \text{ and } \ g : \mathbb{R}^2 \to \mathbb{R},\ g(x, y) = y^k,$$

where $k \in \mathbb{N}$; for $k = 1$, the functions are called coordinate projections. We want to argue that f is continuous; the argument for g is analogous. Let $\boldsymbol{x}_0 = (x_0, y_0) \in \mathbb{R}^2$ be given, and let $\epsilon > 0$ be arbitrary. We need to find a $\delta > 0$ such that

$$\|(x, y) - \boldsymbol{x}_0\| < \delta \quad \Rightarrow \quad |f(x, y) - f(x_0, y_0)| < \epsilon.$$

Since $f(x,y) = x^k$, we almost have the situation as in Example 3.1.1, where we proved that the function $x \mapsto x^k$ is continuous on $\mathbb{R}$. Indeed, we can use the same δ as in that example. Note that

$$\|(x,y) - \boldsymbol{x}_0\| = \sqrt{(x-x_0)^2 + (y-y_0)^2} \geq |x - x_0|$$

Hence, $\|(x,y) - \boldsymbol{x}_0\| < \delta$ implies $|x - x_0| < \delta$ and therefore that

$$|f(x,y) - f(x_0,y_0)| = |x^k - x_0^k| < \epsilon.$$

As for functions of one variable, continuity of a vector function is preserved under standard algebraic operations. In particular, the sum, the scalar product, and the composition of continuous vector functions are again continuous vector functions. We state this in the following two theorems which we present without proof for brevity.

Theorem 3.2.2

Let A denote a set in $\mathbb{R}^n$. Suppose $\boldsymbol{f}, \boldsymbol{g} : A \to \mathbb{R}^k, h : A \to \mathbb{R}$ are vector functions that are continuous at $\boldsymbol{x}_0 \in A$. Then the following vector functions are continuous at $\boldsymbol{x}_0 \in A$:

(i) The sum $\boldsymbol{f} + \boldsymbol{g}$ and difference $\boldsymbol{f} - \boldsymbol{g}$,

(ii) The product $h\boldsymbol{f}$, where h is a scalar function and $h\boldsymbol{f}(\boldsymbol{x}) \coloneqq h(\boldsymbol{x})\boldsymbol{f}(\boldsymbol{x})$,

(iii) The norm $\|\boldsymbol{f}\|$ and the inner product $\langle \boldsymbol{f}, \boldsymbol{g} \rangle$.

If $h(\boldsymbol{x}_0) \neq 0$, the quotient $\boldsymbol{f}/h$ is also continuous at $\boldsymbol{x}_0$.

Note that, in Theorem 3.2.2, the product $h\boldsymbol{f}$ (and the quotient $\boldsymbol{f}/h$) is only defined when h is a scalar function which is why we require h to have codomain co-dom$(h) = \mathbb{R}$.

Example 3.2.2 softmax is continuous

The softmax vector function introduced in Example 1.3.3 on page 20 is continuous. Let us argue why this is the case. By Theorem 3.2.1, it suffices to show that each coordinate function $\mathrm{e}^{x_i}/s(\boldsymbol{x})$, $i = 1, 2, \ldots, n$, of softmax is continuous, where the sum function $s : \mathbb{R}^n \to \mathbb{R}$ is defined by

$$s(\boldsymbol{x}) = \mathrm{e}^{x_1} + \mathrm{e}^{x_2} + \cdots + \mathrm{e}^{x_n}, \quad \boldsymbol{x} = [x_1, x_2, \ldots, x_n]^T \in \mathbb{R}^n,$$

We first note that the exponential function e^x is continuous on $\mathbb{R}$ and

that e^{x_i}, $i = 1, 2, \ldots, n$, are continuous functions on $\mathbb{R}^n$ as they depend only on a single coordinate x_i of $\boldsymbol{x} \in \mathbb{R}^n$. From Theorem 3.2.2(i) it follows that the sum function s is continuous since it is a sum of continuous functions. Finally, since $s(\boldsymbol{x}) \neq 0$ for all $\boldsymbol{x} \in \mathbb{R}^n$, we can use the final statement of Theorem 3.2.2 to conclude that each coordinate function $e^{x_i}/s(\boldsymbol{x})$, $i = 1, 2, \ldots, n$, of softmax is continuous.

Theorem 3.2.3 Composition of continuous vector functions

Let $A \subseteq \mathbb{R}^n$ and $B \subseteq \mathbb{R}^k$ be sets. Suppose that $\boldsymbol{f} : A \to B$ and $\boldsymbol{g} : B \to \mathbb{R}^\ell$ are continuous at $\boldsymbol{x}_0 \in A$ and at $\boldsymbol{f}(\boldsymbol{x}_0) \in B$, respectively. Then the composition $\boldsymbol{h} = \boldsymbol{g} \circ \boldsymbol{f}, A \to \mathbb{R}^\ell$ is continuous at $\boldsymbol{x}_0$.

We are usually interested in continuity in more than in a single point. This leads to the following definition.

Definition 3.2.2 C^0 vector function

A vector function $\boldsymbol{f} : A \to \mathbb{R}^k$ is called *continuous* if it is continuous at all points $\boldsymbol{x} \in A$. Continuous vector functions are called C^0 vector functions.

Let us show how we can argue for the continuity of a function without explicitly to refer to Definition 3.2.1.

Example 3.2.3

The function

$$\boldsymbol{h} : \mathbb{R}^2 \to \mathbb{R}^2, \ \boldsymbol{h}(x, y) = [x^2 + xy + \sin(x + y), e^{xy}]^T$$

is a sum and composition of continuous functions, and thus, by Theorems 3.2.2 and 3.2.3, it is continuous.

Example 3.2.4 Neural networks are continuous

Consider a (feedforward) neural network as introduced in Definition 1.3.2 on page 22. Such a neural network $\boldsymbol{\Phi} : \mathbb{R}^{n_0} \to \mathbb{R}^{n_L}$ can be viewed as a composition of functions

$$\boldsymbol{\Phi}(\boldsymbol{x}) = \boldsymbol{f}^{(L)} \circ \boldsymbol{f}^{(L-1)} \circ \cdots \circ \boldsymbol{f}^{(1)}(\boldsymbol{x}), \tag{3.18}$$

where each function $\boldsymbol{f}^{(\ell)} = \sigma^{(\ell)} \circ T_{A_\ell, \boldsymbol{b}_\ell}$, $\ell = 1, 2, \ldots, L$, represents a layer in the neural network. If the activation functions $\sigma^{(\ell)}$ are continuous (as is typically the case), then each layer map $\boldsymbol{f}^{(\ell)} = \sigma^{(\ell)} \circ T_{A_\ell, \boldsymbol{b}_\ell}$ is

continuous because affine maps $T_{A_\ell,\boldsymbol{b}_\ell}(\boldsymbol{x}) = A_\ell\boldsymbol{x} + \boldsymbol{b}_\ell$ are continuous and compositions of continuous maps are continuous by Theorem 3.2.3. By applying Theorem 3.2.3 to (3.18) repeatedly, we conclude that the neural networks $\boldsymbol{\Phi}$ are continuous vector functions on $\mathbb{R}^{n_0}$.

Special attention is needed for functions that are defined by multiple expressions as illustrated in the next example.

Example 3.2.5

Consider the function

$$f : \mathbb{R}^2 \to \mathbb{R},\ f(x,y) = \begin{cases} \frac{x^2 y}{x^4+y^2} & \text{whenever } (x,y) \neq (0,0), \\ 0 & \text{whenever } (x,y) = (0,0). \end{cases}$$

Then f is continuous for all $(x,y) \neq (0,0)$ since it is a composition of continuous functions. The point $(x,y) = (0,0)$ needs special attention — and the function is indeed not continuous at this point! In order to see this, consider points in $\mathbb{R}^2$ on the curve $y = x^2$, $x > 0$, i.e., points of the form $(x,y) = (x,x^2)$. For any such points we see that

$$f(x,y) = f(x,x^2) = \frac{x^2 x^2}{x^4 + (x^2)^2} = \frac{x^4}{2x^4} = \frac{1}{2};$$

that is, along the curve $y = x^2$, the function f takes the constant value $\frac{1}{2}$. Since there are points on the curve $y = x^2$, $x > 0$, arbitrarily close to the point $(0,0)$ and $f(0,0) = 0$, this implies that the function f is not continuous at $(x,y) = (0,0)$.

One may wonder if the function in Example 3.2.5 can be made continuous at $(0,0)$ if we define $f(0,0)$ to be $1/2$ instead of 0. In other words, we ask if we can "patch" the function f at $(0,0)$ by assigning it a value different from 0 at this point. The answer is no. No matter what value we assign to $f(0,0)$, the function cannot be continuous at $(0,0)$. In order to see this, consider points in $\mathbb{R}^2$ on the line $y = 0$, i.e., points of the form $(x,y) = (x,0)$. Hence, along the line $y = 0$, the function f takes the value $f(x,0) = 0$ for all $x \neq 0$. This argument shows that $f(0,0)$ needs to be 0 in order for f to be continuous at $(0,0)$, but in Example 3.2.5 we saw that $f(0,0)$ also had to be $1/2$. As it is impossible for a function to take two different values at the same point, we conclude that f cannot be continuous at $(0,0)$ regardless of the value we assign to $f(0,0)$.

3.3 Partial derivatives of first order and the gradient vector

In Section 3.2 we saw that continuity for functions of n variables can be defined in a similar fashion as for functions of one variable. In contrast, the theory for differentiability is significantly more involved for functions of several variables than for functions of one variable.

In the next three sections we will develop the theory of differentiability of *scalar* functions of several variables. In Section 3.8 we will return to vector functions.

Let us begin with an introductory example.

Example 3.3.1

Consider the function

$$f : \mathbb{R}^2 \to \mathbb{R}, \ f(x, y) \coloneqq xy^2 + x^2 + y^3.$$

The function f is obviously continuous as it is composed of well-known continuous functions, but we have not yet defined how we should interpret differentiability for a function of two variables.

However, note that if we *fix* an arbitrary value of the variable y and consider the function f only as a function of the remaining variable x, we obtain the function h_y (the subscript is chosen to indicate that the function depends on the chosen y) given by

$$h_y : \mathbb{R} \to \mathbb{R}, \ h_y(x) = f(x, y) = xy^2 + x^2 + y^3.$$

For all (fixed) values of y, the function h_y is a differentiable function of the variable x, and

$$h_y'(x) = y^2 + 2x. \tag{3.19}$$

Note that the expression (3.19) is obtained simply by considering y as a constant and applying standard rules for differentiation of a function of one variable.

We will now introduce a new terminology to describe the situation. Being a little ahead of the formal definition (to be given in Definition 3.3.1) we will say that *the partial derivative of the function f with respect to the variable x, to be denoted by $\frac{\partial f}{\partial x}(x, y)$, exists, and is given by*

$$\frac{\partial f}{\partial x}(x, y) = h_y'(x) = y^2 + 2x. \tag{3.20}$$

Similarly, if we *fix* an arbitrary value of the variable x and consider the function f only as a function of the remaining variable y, we obtain the function

$$k_x : \mathbb{R} \to \mathbb{R},\ k_x(y) = f(x,y) = xy^2 + x^2 + y^3.$$

Regardless of the chosen x, the function k_x is a differentiable function of the variable y, and

$$k_x'(y) = 2xy + 3y^2. \tag{3.21}$$

We say that *the partial derivative of the function f with respect to the variable y, to be denoted by $\frac{\partial f}{\partial y}(x,y)$, exists, and is given by*

$$\frac{\partial f}{\partial y}(x,y) = k_x'(y) = 2xy + 3y^2. \tag{3.22}$$

Note that in order to obtain the expressions (3.20) and (3.22), it is not necessary to introduce the functions h_y and k_x. Since, in order to calculate the partial derivative $\frac{\partial f}{\partial x}(x,y)$, we simply consider y as a *constant* and differentiate the function f with respect to the variable x. Similarly, we calculate the partial derivative $\frac{\partial f}{\partial y}(x,y)$ by differentiating the expression $f(x,y)$ with respect to the variable y when considering x as a *constant*. Let us illustrate this by one more example.

Example 3.3.2

Consider the function

$$f : \mathbb{R}^2 \to \mathbb{R},\ \ f(x,y) = \sin(x^2 + y).$$

Then (compare with the calculations in Example 3.1.3)

$$\frac{\partial f}{\partial x}(x,y) = 2x\cos(x^2 + y)$$

and

$$\frac{\partial f}{\partial y}(x,y) = \cos(x^2 + y).$$

Already in Example 3.1.4, we saw that a function of one variable can be differentiable at some points and non-differentiable at other points. The same complication occurs for functions of several variables. Therefore we would like to stress the fact that the existence of the partial derivatives for a function $f : \mathbb{R}^2 \to \mathbb{R}$ has to be checked *individually* for each point $(x,y) \in \mathbb{R}^2$.

Example 3.3.3

Consider the function

$$f : \mathbb{R}^2 \to \mathbb{R}, \;\; f(x, y) = \mathrm{ReLU}(x)\, y^2.$$

For each fixed $x \in \mathbb{R}$, the function $k_x(y) = \mathrm{ReLU}(x)\, y^2$ is differentiable for all choices of $y \in \mathbb{R}$; thus, the partial derivative $\frac{\partial f}{\partial y}(x, y)$ exists for all $(x, y) \in \mathbb{R}^2$, and

$$\frac{\partial f}{\partial y}(x, y) = 2\ \mathrm{ReLU}(x)\, y.$$

The situation is slightly more complicated when we fix $y \in \mathbb{R}$ and want to consider f as a function of x. The reason is that as we saw in Example 3.1.4, the function ReLU is not differentiable at $x = 0$. In order to handle this, let us write the function f as

$$f(x, y) = \begin{cases} xy^2 & \text{whenever } x \geq 0, \\ 0 & \text{whenever } x < 0. \end{cases}$$

From here we see that for any fixed $y \in \mathbb{R}$, the function $h_y(x) = \mathrm{ReLU}(x)\, y^2$ is differentiable whenever $x \neq 0$; thus, for $(x, y) \in \mathbb{R}^2$ with $x \neq 0$ the partial derivative $\frac{\partial f}{\partial x}(x, y)$ exists and

$$\frac{\partial f}{\partial x}(x, y) = \begin{cases} y^2 & \text{whenever } x > 0, \\ 0 & \text{whenever } x < 0. \end{cases}$$

However, when we fix any $y \neq 0$, the function h_y is not differentiable at $x = 0$, as we saw in Example 3.1.4; that is, the partial derivatives $\frac{\partial f}{\partial x}(0, y)$ do not exist.

Complications like in Example 3.3.3 makes it evident that we need a formal definition of the partial derivatives. We will now state such a definition for a function of n variables $(x_1, x_2, \ldots, x_n)$.

Definition 3.3.1 Partial derivative

Let U be an open subset of $\mathbb{R}^n$, and let $f : U \to \mathbb{R}$ be a scalar function. Consider a point $(x_1, x_2, \ldots, x_n) \in U$. Fix some $j = 1, \ldots, n$, and assume that the expression

$$\frac{f(x_1, x_2, \ldots, x_j + h, \ldots x_n) - f(x_1, x_2, \ldots, x_j, \ldots, x_n)}{h}$$

has a limit as $h \to 0$. Then we say that the *partial derivative* $\frac{\partial f}{\partial x_j}$ exists at the point $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$, and we define

$$\frac{\partial f}{\partial x_j}(\boldsymbol{x}) \;=\; \lim_{h\to 0} \frac{f(x_1, x_2, \ldots, x_j + h, \ldots, x_n) - f(x_1, x_2, \ldots, x_j, \ldots, x_n)}{h}.$$

Let us explicitly formulate the definition of the partial derivative with respect to the first variable x_1. If the expression

$$\frac{f(x_1 + h, x_2, \ldots, x_n) - f(x_1, x_2, \ldots, x_n)}{h}$$

has a limit as $h \to 0$, the partial derivative $\frac{\partial f}{\partial x_1}$ exists at the point $(x_1, x_2, \ldots, x_n)$ and it is given by

$$\frac{\partial f}{\partial x_1}(x_1, x_2, \ldots, x_n) = \lim_{h\to 0} \frac{f(x_1 + h, x_2, \ldots, x_n) - f(x_1, x_2, \ldots, x_n)}{h}.$$

If all the partial derivatives $\frac{\partial f}{\partial x_k}, k = 1, \ldots, n$ exist at a given point $(x_1, \ldots, x_n)$, we will collect the information in a single vector:

Definition 3.3.2 Gradient vector

If all the partial derivatives $\frac{\partial f}{\partial x_k}, k = 1, \ldots, n$, exist at a given point $(x_1, \ldots, x_n)$, the *gradient vector* of the function f at the point $\boldsymbol{x} = (x_1, \ldots, x_n)$ is defined as the vector

$$\nabla f(\boldsymbol{x}) \coloneqq \left(\frac{\partial f}{\partial x_1}(\boldsymbol{x}), \frac{\partial f}{\partial x_2}(\boldsymbol{x}), \cdots, \frac{\partial f}{\partial x_n}(\boldsymbol{x}) \right). \tag{3.23}$$

The symbol ∇ is pronounced "nabla". The gradient vector will be considered as a *column* vector unless stated otherwise.

Example 3.3.4

For the function

$$f : \mathbb{R}^2 \to \mathbb{R},\; f(x, y) = xy^2 + x^2 + y^3,$$

we calculated the partial derivatives in Example 3.3.1. Formulated in terms of the gradient vector, the result says that

$$\nabla f(x, y) = (y^2 + 2x, 2xy + 3y^2). \tag{3.24}$$

Example 3.3.5

Consider the function

$$f : \mathbb{R}^3 \to \mathbb{R},\ \ f(x_1, x_2, x_3) = x_1 x_2 + x_2^2 x_3^3.$$

Then all the partial derivatives exist at all points $(x_1, x_2, x_3) \in \mathbb{R}$, and

$$\begin{aligned}\nabla f(x_1, x_2, x_3) &= \left(\frac{\partial f}{\partial x_1}(x_1, x_2, x_3), \frac{\partial f}{\partial x_2}(x_1, x_2, x_3), \frac{\partial f}{\partial x_3}(x_1, x_2, x_3)\right) \\ &= (x_2, x_1 + 2x_2 x_3^3, 3x_2^2 x_3^2).\end{aligned}$$

Recall that if a function $f : \mathbb{R} \to \mathbb{R}$ is differentiable, then f is also continuous. For a function $f : \mathbb{R}^n \to \mathbb{R}$ it is therefore natural to ask whether existence of all the partial derivatives implies that the function is continuous. The answer turns out to be no:

Example 3.3.6

Consider again the function in Example 3.2.5,

$$f : \mathbb{R}^2 \to \mathbb{R},\ \ f(x, y) = \begin{cases} \frac{x^2 y}{x^4 + y^2} & \text{whenever } (x, y) \neq (0, 0), \\ 0 & \text{whenever } (x, y) = (0, 0). \end{cases}$$

For $(x, y) \neq (0, 0)$, the function f is composed of functions that are differentiable with respect to as well x as y, i.e., the partial derivatives $\frac{\partial f}{\partial x}(x, y)$ and $\frac{\partial f}{\partial y}(x, y)$ exist. In order to check whether the partial derivative $\frac{\partial f}{\partial x}(x, y)$ exists at the point $(x, y) = (0, 0)$, we apply Definition 3.3.1 and calculate

$$\frac{f(h, 0) - f(0, 0)}{h} = \frac{0 - 0}{h} = 0;$$

since this expression indeed has a limit (namely, 0) as $h \to 0$, we conclude that also the partial derivative $\frac{\partial f}{\partial x}(0, 0)$ exists. By a similar argument, also the partial derivative $\frac{\partial f}{\partial x}(0, 0)$ exists, i.e., both partial derivatives exist in all points. However, as we saw in Example 3.2.5, the function f is not continuous.

In order to avoid issues like in Example 3.3.6 we need to consider a stronger differentiability condition than just existence of the partial derivatives. We will do so in Section 3.6.

3.4 Directional derivatives

When computing the partial derivative $\frac{\partial f}{\partial x_i}$ in $\boldsymbol{x}_0 \in \mathbb{R}^n$, one moves along the ith coordinate axes, in the sense that $\frac{\partial f}{\partial x_i}(\boldsymbol{x}_0)$ is equal to the usual derivative as in Definition 3.1.2 of the function $h \mapsto f(\boldsymbol{x}_0 + h\boldsymbol{e}_i), \mathbb{R} \to \mathbb{R}$, evaluated at $h = 0$, where $\boldsymbol{e}_i = [0, \dots, 0, 1, 0, \dots, 0]^T$ is the ith standard basis vector of $\mathbb{R}^n$. However, there is nothing special about the coordinate axes, and it is just as natural to move along other directions. This is the idea of directional derivatives.

Definition 3.4.1 Directional derivative

Let $\boldsymbol{v}$ be a non-zero vector in $\mathbb{R}^n$, let U be an open subset of $\mathbb{R}^n$, and let $f : U \to \mathbb{R}$ be a scalar function. Define $\boldsymbol{u} = \boldsymbol{v}/\|\boldsymbol{v}\|$. Consider a point $(x_1, x_2, \dots, x_n) \in U$. Assume that the expression

$$\frac{f(\boldsymbol{x} + h\boldsymbol{v}) - f(\boldsymbol{x})}{h}$$

has a limit as $h \to 0$. Then we say that the *directional derivative* $\nabla_{\boldsymbol{v}} f$ exists at $\boldsymbol{x} = (x_1, x_2, \dots, x_n)$, and we define

$$\nabla_{\boldsymbol{v}} f(\boldsymbol{x}) \coloneqq \lim_{h \to 0} \frac{f(\boldsymbol{x} + h\boldsymbol{u}) - f(\boldsymbol{x})}{h}.$$

Note that the directional derivative $\nabla_{\boldsymbol{v}} f$ evaluated at a point $\boldsymbol{x} \in U$ is just a scalar in $\mathbb{R}$. Moreover, $\nabla_{\boldsymbol{e}_i} f(\boldsymbol{x}) = \frac{\partial f}{\partial x_i}(\boldsymbol{x})$.

The length of the vector $\boldsymbol{v}$ in the definition of $\nabla_{\boldsymbol{v}} f$ is irrelevant, and we see that only the direction matters. In particular, if $\boldsymbol{w} \in \operatorname{span}\{\boldsymbol{v}\}$ is a non-zero vector, then $\nabla_{\boldsymbol{w}} f = \pm \nabla_{\boldsymbol{v}} f$. (The minus sign is needed if $\boldsymbol{w}$ and $\boldsymbol{v}$ points in opposite directions, i.e., $\boldsymbol{w} = c\boldsymbol{v}$ for $c < 0$.)

The directional derivative can easily be computed from the gradient vector using the standard inner product in $\mathbb{R}^n$.

Lemma 3.4.1

If $\boldsymbol{u}$ is a unit vector, then

$$\nabla_{\boldsymbol{u}} f(\boldsymbol{x}) = \langle \boldsymbol{u}, \nabla f(\boldsymbol{x}) \rangle \tag{3.25}$$

for $\boldsymbol{x} \in \mathbb{R}^n$.

We leave the proof to the reader. In case $\boldsymbol{v}$ is not necessarily of norm one, the formula in Equation (3.25) should be changed to $\nabla_{\boldsymbol{v}} f(\boldsymbol{x}) = \frac{1}{\|\boldsymbol{v}\|} \langle \boldsymbol{v}, \nabla f(\boldsymbol{x}) \rangle$.

The directional derivative $|\nabla_{\boldsymbol{v}} f(\boldsymbol{x})|$ is maximized if $\boldsymbol{v}$ is a unit vector in the span of the gradient vector as shown in Exercise 3.4.1. Loosely speaking, we say that the function $f : \mathbb{R}^n \to \mathbb{R}$ increases and decreases the most along the gradient vector.

Exercise 3.4.1

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a function for which all directional derivatives exist at $\boldsymbol{x} \in \mathbb{R}^n$. Show that $\boldsymbol{u} = \nabla f(\boldsymbol{x})/\|\nabla f(\boldsymbol{x})\|$ is a unit vector. Show that the scalar $|\nabla_{\boldsymbol{v}} f(\boldsymbol{x})|$ is maximized for $\boldsymbol{v} = \pm\boldsymbol{u}$. *Hint: Equation (3.25) and Theorem 2.1.4 on page 47.* Determine now the sign of $\nabla_{\boldsymbol{u}} f(\boldsymbol{x})$ and $\nabla_{-\boldsymbol{u}} f(\boldsymbol{x}) = -\nabla_{\boldsymbol{u}} f(\boldsymbol{x})$, and explain why $\boldsymbol{u}$ is called a direction of steepest ascent, while $-\boldsymbol{u}$ is called a direction of steepest descent.

From Exercise 3.4.1 we see that the gradient vector at a point $\boldsymbol{x}$ contains information about the rate of change of f in *all* directions from the point $\boldsymbol{x}$. This observation leads to the gradient ascent (descent) method for finding local maxima (minima) of functions $f : \mathbb{R}^n \to \mathbb{R}$. Local minima and maxima will be thoroughly studied in Chapter 5, but here in the next example we just want to illustrate how the gradient vector can be used in optimization.

Example 3.4.2 Gradient ascent method

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a function for which the directional derivatives exist for all $\boldsymbol{x} \in \mathbb{R}^n$. We visualize the graph of f as a landscape where high values of f correspond to high altitudes and low values correspond to low altitudes. Our goal is to find the maximum of f — analogous to finding the mountain peak on a hiking trip — using only "local" information about the landscape.

To make this concrete, consider the example $f(x, y) = 300 - 10(x + 1)^2 - 4(y - 5)^2$ for $n = 2$. This function has a unique global maximum at $(x, y) = (-1, 5)$ where the height of the mountain peak is $f(-1, 5) = 300$.

Let $\boldsymbol{x}_0 \in \mathbb{R}^n$ be a point such that $\nabla f(\boldsymbol{x}_0) \neq \boldsymbol{0}$. In our concrete example, we could choose $\boldsymbol{x}_0 = (0, 0)$ which corresponds to starting our hike at the point $(0, 0)$ in the xy-plane where the altitude is $f(0, 0) = 300 - 10 - 100 = 190$. Of course, if we could see the mountain peak from $\boldsymbol{x}_0$, we would just walk straight to the peak, but in many practical situations this is not possible. In our analogy this could be the case if there was a thick fog that only allows us to see a few meters around us. Instead, we have to rely on local information about the landscape, i.e., the values of f and its gradient vector at $\boldsymbol{x}_0$.

From Exercise 3.4.1 we know that the gradient vector $\nabla f(\boldsymbol{x}_0)$ points in the direction of steepest ascent. This means that if we move a small

step from $\boldsymbol{x}_0$ in the direction of $\nabla f(\boldsymbol{x}_0)$, the value of the function f will increase the most compared to moving in any other direction. In our concrete example, we have

$$\nabla f(x,y) = \begin{bmatrix} \frac{\partial f}{\partial x}(x,y) \\ \frac{\partial f}{\partial y}(x,y) \end{bmatrix} = \begin{bmatrix} -20(x+1) \\ -8(y-5) \end{bmatrix},$$

and thus $\nabla f(\boldsymbol{x}_0) = \nabla f(0,0) = (-20, 40)$. This means that if we start our hike at the point $(0,0)$, we should walk in the direction $(-1,2)$ in order to increase our altitude as much as possible.

This observation is the basis for the *gradient ascent method* for finding local maxima of functions $f : \mathbb{R}^n \to \mathbb{R}$. The method starts with an initial guess $\boldsymbol{x}_0$ for a local maximum, and then iteratively updates the guess according to

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k \nabla f(\boldsymbol{x}_k), \tag{3.26}$$

where $\alpha_k > 0$ is a step size (also called learning rate) that may depend on the iteration k.

In applications, in particular in machine learning, the learning rate is often chosen to be rather small constant (e.g., $\alpha_k = 0.01$ or 0.001) in order to avoid overshooting the maximum. In our analogy of hiking in a foggy landscape, this corresponds to taking small steps to avoid stepping past the mountain peak without noticing. Conversely, if the step size is too small, convergence becomes very slow. There is no universal rule for choosing the learning rate; in practice, it is determined by trial and error, depending on the complexity (or rather curvature) of the graph of the function f.

Exercise 3.4.3 Gradient descent method

The gradient ascent method from Example 3.4.2 can be used to find a local *maximum* of a differentiable function $f : \mathbb{R}^n \to \mathbb{R}$. How should the iteration formula in (3.26) be changed in order to find a local minimum instead of a local maximum? This method is called the *gradient descent method* (or *steepest descent method*).

3.5 Partial derivatives of second order and the Hessian matrix

We will now give a short introduction to higher-order partial derivatives. In Chapters 4 and 5 we will see that they play a central role in connection with

Taylor polynomials and optimization problems for functions of n variables (like $f''(x)$ does for a function of a single variable!).

Consider again a function $f : U \to \mathbb{R}$, where U is an open set in $\mathbb{R}^n$. If the partial derivative $\frac{\partial f}{\partial x_k}$ exists at all points $(x_1, \ldots, x_n) \in U$ for some index $k = 1, \ldots, n$, we can consider the function $\frac{\partial f}{\partial x_k}$ as a new function of n variables,

$$\frac{\partial f}{\partial x_k} : U \to \mathbb{R}.$$

Thus, we can ask whether the function $\frac{\partial f}{\partial x_k}$ has partial derivatives, say, with respect to x_j for some $j = 1, \ldots, n$, at a point $(x_1, \ldots, x_n) \in U$; if this is the case, the partial derivative will be denoted by

$$\frac{\partial^2 f}{\partial x_j \partial x_k}(x_1, \ldots, x_n).$$

In the special case where $j = k$ we will use the shorter notation

$$\frac{\partial^2 f}{\partial x_k^2}(x_1, \ldots, x_n) \coloneqq \frac{\partial^2 f}{\partial x_k \partial x_k}(x_1, \ldots, x_n).$$

Example 3.5.1

For the function

$$f : \mathbb{R}^3 \to \mathbb{R}, \ f(x_1, x_2, x_3) = x_1 x_2 + x_2^2 x_3^3,$$

we saw in Example 3.3.5 on page 97 that, for $\boldsymbol{x} = (x_1, x_2, x_3)$,

$$\frac{\partial f}{\partial x_1}(\boldsymbol{x}) = x_2, \ \ \frac{\partial f}{\partial x_2}(\boldsymbol{x}) = x_1 + 2x_2 x_3^3, \ \ \frac{\partial f}{\partial x_3}(\boldsymbol{x}) = 3x_2^2 x_3^2.$$

All these partial derivatives have partial derivatives with respect to x_1, x_2, and x_3; by direct calculation we get that

$$\begin{aligned}
&\frac{\partial^2 f}{\partial x_1^2}(\boldsymbol{x}) = 0, \quad \frac{\partial^2 f}{\partial x_2 \partial x_1}(\boldsymbol{x}) = 1, \quad \frac{\partial^2 f}{\partial x_3 \partial x_1}(\boldsymbol{x}) = 0,\\
&\frac{\partial^2 f}{\partial x_1 \partial x_2}(\boldsymbol{x}) = 1, \quad \frac{\partial^2 f}{\partial x_2^2}(\boldsymbol{x}) = 2x_3^3, \quad \frac{\partial^2 f}{\partial x_3 \partial x_2}(\boldsymbol{x}) = 6x_2 x_3^2,\\
&\frac{\partial^2 f}{\partial x_1 \partial x_3}(\boldsymbol{x}) = 0, \quad \frac{\partial^2 f}{\partial x_2 \partial x_3}(\boldsymbol{x}) = 6x_2 x_3^2, \quad \frac{\partial^2 f}{\partial x_3^2}(\boldsymbol{x}) = 6x_2^2 x_3,
\end{aligned}$$

where $\boldsymbol{x} = (x_1, x_2, x_3)$.

By inspection of the second-order partial derivatives in Example 3.5.1 we notice that for all values of $j, k = 1, 2, 3$ we have that

$$\frac{\partial^2 f}{\partial x_j \partial x_k}(x_1, x_2, x_3) = \frac{\partial^2 f}{\partial x_k \partial x_j}(x_1, x_2, x_3);$$

that is, the order in which the partial derivatives is taken does not matter. This turns out to hold in general, provided that an extra assumption introduced in Section 3.6 is satisfied — see Theorem 3.6.5.

All information about the second-order partial derivatives are collected in the so-called *Hessian matrix* defined as follows.

Definition 3.5.1 Hessian matrix

Assuming that all the second-order partial derivatives exist for a scalar function $f : \mathbb{R}^n \to \mathbb{R}$, the associated *Hessian matrix* is the $n \times n$ matrix given by:

$$\boldsymbol{H}_f(\boldsymbol{x}) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2}(\boldsymbol{x}) & \frac{\partial^2 f}{\partial x_1 \partial x_2}(\boldsymbol{x}) & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n}(\boldsymbol{x}) \\ \frac{\partial^2 f}{\partial x_2 \partial x_1}(\boldsymbol{x}) & \frac{\partial^2 f}{\partial x_2^2}(\boldsymbol{x}) & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n}(\boldsymbol{x}) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1}(\boldsymbol{x}) & \frac{\partial^2 f}{\partial x_n \partial x_2}(\boldsymbol{x}) & \cdots & \frac{\partial^2 f}{\partial x_n^2}(\boldsymbol{x}) \end{bmatrix}. \tag{3.27}$$

For a function $f : \mathbb{R}^2 \to \mathbb{R}$ the Hessian matrix takes the form

$$\boldsymbol{H}_f(x, y) = \begin{bmatrix} \frac{\partial^2 f}{\partial x}(x, y) & \frac{\partial^2 f}{\partial x \partial y}(x, y) \\ \frac{\partial^2 f}{\partial y \partial x}(x, y) & \frac{\partial^2 f}{\partial y^2}(x, y) \end{bmatrix}. \tag{3.28}$$

Example 3.5.2

For the function

$$f : \mathbb{R}^2 \to \mathbb{R}, \;\; f(x, y) \coloneqq \sin(x^2 + y),$$

we saw in Example 3.3.2 that

$$\frac{\partial f}{\partial x}(x, y) = 2x \cos(x^2 + y), \;\text{ and }\; \frac{\partial f}{\partial y}(x, y) = \cos(x^2 + y).$$

Thus

$$\frac{\partial^2 f}{\partial x^2}(x, y) = 2\cos(x^2 + y) - 4x^2 \sin(x^2 + y)$$

$$\frac{\partial^2 f}{\partial x \partial y}(x,y) = -2x\sin(x^2+y)$$
$$\frac{\partial^2 f}{\partial y \partial x}(x,y) = -2x\sin(x^2+y)$$
$$\frac{\partial^2 f}{\partial y^2}(x,y) = -\sin(x^2+y).$$

The Hessian matrix is therefore

$$\begin{aligned}\boldsymbol{H}_f(x,y) &= \begin{bmatrix} \frac{\partial^2 f}{\partial x}(x,y) & \frac{\partial^2 f}{\partial x \partial y}(x,y) \\ \frac{\partial^2 f}{\partial y \partial x}(x,y) & \frac{\partial^2 f}{\partial y^2}(x,y) \end{bmatrix} \\ &= \begin{bmatrix} 2\cos(x^2+y) - 4x^2\sin(x^2+y) & -2x\sin(x^2+y) \\ -2x\sin(x^2+y) & -\sin(x^2+y) \end{bmatrix}.\end{aligned}$$

Example 3.5.3

For the function f in Example 3.5.1, the Hessian matrix is

$$\begin{aligned}\boldsymbol{H}_f(\boldsymbol{x}) &= \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2}(x_1,x_2,x_3) & \frac{\partial^2 f}{\partial x_1 \partial x_2}(x_1,x_2,x_3) & \frac{\partial^2 f}{\partial x_1 \partial x_3}(x_1,x_2,x_3) \\ \frac{\partial^2 f}{\partial x_2 \partial x_1}(x_1,x_2,x_3) & \frac{\partial^2 f}{\partial x_2^2}(x_1,x_2,x_3) & \frac{\partial^2 f}{\partial x_2 \partial x_3}(x_1,x_2,x_3) \\ \frac{\partial^2 f}{\partial x_3 \partial x_1}(x_1,x_2,x_3) & \frac{\partial^2 f}{\partial x_3 \partial x_2}(x_1,x_2,x_3) & \frac{\partial^2 f}{\partial x_3^2}(x_1,x_2,x_3) \end{bmatrix} \\ &= \begin{bmatrix} 0 & 1 & 0 \\ 1 & 2x_3^2 & 6x_2x_3^2 \\ 0 & 6x_2x_3^2 & 6x_2^2x_3 \end{bmatrix},\end{aligned}$$

where $\boldsymbol{x} = x_1, x_2, x_3 \in \mathbb{R}^3$.

Example 3.5.4 Quadratic forms

Let $q : \mathbb{R}^n \to \mathbb{R}$ be a quadratic form. Recall from Lemma 2.8.5 on page 73 that we may assume that

$$q(\boldsymbol{x}) = \boldsymbol{x}^T A \boldsymbol{x} + \boldsymbol{x}^T \boldsymbol{b} + c,$$

where $A \in \mathsf{M}_n(\mathbb{R})$ is symmetric, $\boldsymbol{b} \in \mathbb{R}^n$, and $c \in \mathbb{R}$. It can be shown using Example 2.8.6 on page 74 that the Hessian $\boldsymbol{H}_q$ is constant and equals

$$\boldsymbol{H}_q = 2A.$$

Exercise 3.5.5

Prove the claims made in Example 3.5.4.

3.6 Differentiability of scalar functions of several variables

We have already mentioned that differentiability is a much more complicated issue for functions of several variables than for functions of one variable. One reason is that Definition 3.1.2 is restricted to the one-dimensional case: in contrast to our approach to continuity, the condition can not be changed to a condition on $\mathbb{R}^n$ simply by exchanging the point x_0 by a vector $\boldsymbol{x}_0$. The argument for this is that if we want to consider the numerator in (3.6) with x_0 replaced by a vector $\boldsymbol{x}_0$, i.e., an expression of the form

$$f(\boldsymbol{x}_0 + h) - f(\boldsymbol{x}_0),$$

then h is forced to be a vector in $\mathbb{R}^n$, and therefore the division with h in (3.6) becomes meaningless as we have no concept of division by a vector.

In order to generalize the concept of differentiability to higher dimensions, we therefore need to reformulate the condition in Definition 3.1.2 in such a way that we obtain a condition that can be transferred into $\mathbb{R}^n$. This is done next.

Lemma 3.6.1

Consider a function $f : I \to \mathbb{R}$, where I is an open interval in $\mathbb{R}$, and let $x_0 \in I$. Fix $c \in \mathbb{R}$. Then the following are equivalent:

(i) The function f is differentiable at x_0 and $f'(x_0) = c$.

(ii) There exists a function $\varepsilon : \mathbb{R} \to \mathbb{R}$ such that $\varepsilon(h) \to 0$ as $h \to 0$ and

$$f(x_0 + h) - f(x_0) - ch - \varepsilon(h)|h| = 0 \quad \text{for all } h \in \mathbb{R}. \tag{3.29}$$

Proof. We first note that we in (ii) can replace (3.29) with the equation

$$f(x_0 + h) - f(x_0) - ch - \varepsilon(h)h = 0 \quad \text{for all } h \in \mathbb{R}.$$

Indeed, the two ε functions are related by a changing the sign for $h < 0$. We will in the rest of the prove use this version of (3.29).

Now assume that (i) holds. Then, by (3.7)

$$\frac{f(x_0+h)-f(x_0)}{h} \to f'(x_0) = c \text{ as } h \to 0.$$

Now define the function ε by

$$\varepsilon(h) := \begin{cases} \frac{f(x_0+h)-f(x_0)}{h} - c & \text{whenever } h \neq 0, \\ 0 & \text{whenever } h = 0. \end{cases}$$

Then $\varepsilon(h) \to 0$ as $h \to 0$, and by multiplying the expression for $\varepsilon(h)$ with h and collecting all terms on one side of the equality sign, we see that (3.29) holds.

On the other hand, if (ii) holds, then

$$\varepsilon(h)h = f(x_0+h) - f(x_0) - ch.$$

Dividing by h now yields that

$$\varepsilon(h) = \frac{f(x_0+h)-f(x_0)}{h} - c;$$

applying now that $\varepsilon(h) \to 0$ as $h \to 0$, we conclude that

$$\frac{f(x_0+h)-f(x_0)}{h} \to c \text{ as } h \to 0.$$

This means precisely that the function f is differentiable at x_0 and that $f'(x_0) = c$, which proves (i).

■

Let us discuss condition (ii) in Lemma 3.6.1. First, we remark that the constant c appearing in (ii) is not arbitrary, indeed, it is equal to $f'(x_0)$. Second, we note that ch for a given constant c defines a linear map $h \mapsto ch$, $\mathbb{R} \to \mathbb{R}$. Hence, the derivative of a function is closely related to linear maps in h. This linear map also "appears" in the tangent line of the graph at $(x_0, f(x_0))$[1].

In contrast to our original definition of differentiability, the condition (ii) in Lemma 3.6.1 has a natural generalization to functions of several variables. Let us explain this in detail.

Consider a function $f : U \to \mathbb{R}$, where U is an open set in $\mathbb{R}^n$, and let $\boldsymbol{x}_0 \in U$. In order to generalize the condition (3.29) we will do the following:

[1]The graph of this linear map is a line in $\mathbb{R}^2$ through $(0,0)$. If we add $(x_0, f(x_0))$ to each point in this set, we obtain the tangent line.

1) We will replace the scalar h by a (column) vector $\boldsymbol{h} \in \mathbb{R}^n$.

2) We will replace the scalar c by a (column) vector $\boldsymbol{c} \in \mathbb{R}^n$ and the quantity ch by the scalar $\boldsymbol{c}^T\boldsymbol{h}$.

3) We will replace the function $\varepsilon : \mathbb{R} \to \mathbb{R}$ by a function $\varepsilon : \mathbb{R}^n \to \mathbb{R}$ of n variables, and we replace the expression $\varepsilon(h)|h|$ by $\varepsilon(\boldsymbol{h})\|\boldsymbol{h}\|$.

Note that in 2) $\boldsymbol{h} \mapsto \boldsymbol{c}^T\boldsymbol{h}$ is a linear map from $\mathbb{R}^n$ to $\mathbb{R}$ defined by the $1 \times n$ "matrix" $\boldsymbol{c}^T$. Note also that the $\boldsymbol{c}^T\boldsymbol{h}$ is just the standard inner product on $\mathbb{R}^n$, i.e., $\boldsymbol{c}^T\boldsymbol{h} = \boldsymbol{h} \cdot \boldsymbol{c} = \langle \boldsymbol{h}, \boldsymbol{c} \rangle$. In 3), the norm $\|\boldsymbol{h}\|$ reduces to $|h_1|$ if $n = 1$ and $\boldsymbol{h} = [h_1]$.

These modifications lead to the following definition of differentiability:

Definition 3.6.1 Differentiability

Consider a function $f : U \to \mathbb{R}$, where U is an open set in $\mathbb{R}^n$, and let $\boldsymbol{x}_0 \in U$. We say that the function f is differentiable at the point $\boldsymbol{x}_0$ if there exist a vector $\boldsymbol{c} \in \mathbb{R}^n$ and a function $\varepsilon : \mathbb{R}^n \to \mathbb{R}$ such that $\varepsilon(\boldsymbol{h}) \to 0$ as $\boldsymbol{h} \to 0$ and

$$f(\boldsymbol{x}_0 + \boldsymbol{h}) - f(\boldsymbol{x}_0) - \boldsymbol{c}^T\boldsymbol{h} - \varepsilon(\boldsymbol{h})\|\boldsymbol{h}\| = 0, \quad \forall \boldsymbol{h} \in \mathbb{R}^n. \tag{3.30}$$

We will now show that there is a strong link between Definition 3.6.1 and the concept of partial derivatives in Section 3.3. In particular, if the function f is differentiable at the point $\boldsymbol{x}_0$, all the partial derivatives $\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \ldots, \frac{\partial f}{\partial x_n}$ exist at the point $\boldsymbol{x}_0$ and the vector $\boldsymbol{c}$ in (3.30) equals the gradient vector $\nabla f(\boldsymbol{x}_0)$. Note that this is similar to the situation in Lemma 3.6.1 on page 104, where the constant c is equal to $f'(x_0)$. The linear map $\boldsymbol{h} \mapsto \nabla f(\boldsymbol{x}_0)^T\boldsymbol{h}$ is called the *differential* of f at $\boldsymbol{x}_0$.

Theorem 3.6.2

Assume that the function $f : U \to \mathbb{R}$ is differentiable at the point $\boldsymbol{x}_0$. Then all the partial derivatives $\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \ldots, \frac{\partial f}{\partial x_n}$ exist at the point $\boldsymbol{x}_0$ and the vector $\boldsymbol{c}$ in Definition 3.6.1 equals the gradient vector $\nabla f(\boldsymbol{x}_0)$.

Proof. Assume that f is differentiable at the point $\boldsymbol{x}_0$ and choose the vector $\boldsymbol{c}$ and the function $\varepsilon(\boldsymbol{h})$ such that (3.30) is satisfied. We will now consider certain special choices of the vector $\boldsymbol{h}$, namely $\boldsymbol{h} = (h, 0, \ldots, 0)$, where

$h \in \mathbb{R}$. The matrix product $\boldsymbol{c}^T\boldsymbol{h}$ becomes:

$$\boldsymbol{c}^T\boldsymbol{h} = \begin{bmatrix} c_1 & c_2 & \cdots & c_n \end{bmatrix} \begin{bmatrix} h \\ 0 \\ \vdots \\ 0 \end{bmatrix} = c_1 h.$$

Since $\|\boldsymbol{h}\| = \sqrt{h^2 + 0 + \cdots + 0} = |h|$, we also see that

$$\varepsilon(\boldsymbol{h})\|\boldsymbol{h}\| = \varepsilon(\boldsymbol{h})|h|$$

Inserting into (3.30), we obtain

$$f\left(\boldsymbol{x}_0 + (h, 0, \ldots, 0)\right) - f(\boldsymbol{x}_0) - c_1 h - \varepsilon(h, 0, \ldots, 0)|h| = 0,$$

where

$$\varepsilon(h, 0, \ldots, 0) \to 0 \quad \text{whenever } h \to 0.$$

Applying now Lemma 3.6.1, we see that the partial derivative $\frac{\partial f}{\partial x_1}$ exists at the point $\boldsymbol{x}_0$ and that $\frac{\partial f}{\partial x_1}(\boldsymbol{x}_0) = c_1$. A similar argument applies to the other partial derivatives, and the conclusion is that

$$\boldsymbol{c} = (c_1, c_2, \ldots, c_n) = \left(\frac{\partial f}{\partial x_1}(\boldsymbol{x}_0), \frac{\partial f}{\partial x_2}(\boldsymbol{x}_0), \cdots, \frac{\partial f}{\partial x_n}(\boldsymbol{x}_0)\right) = \nabla f(\boldsymbol{x}_0)$$

as claimed.

■

Theorem 3.6.2 shows that differentiability of a function $f : U \to \mathbb{R}$ implies existence of the partial derivatives. On the other hand, existence of the partial derivatives is not enough to guarantee differentiability! Indeed, in Example 3.3.6 we saw an example of a function, for which all the partial derivatives exist; but the function is not continuous, and hence, by Theorem 3.6.4 on the following page, not differentiable. However, a slightly weaker result hold: existence *and* continuity of all partial derivatives implies differentiability.

Theorem 3.6.3

Let $U \subseteq \mathbb{R}^n$ be an open set and let $f : U \to \mathbb{R}$ be a scalar function. If the partial derivatives $\frac{\partial f}{\partial x_k}$ exist at all points and are continuous, then f is differentiable for all $\boldsymbol{x} \in U$.

Proof. For a proof we refer the reader to Theorem 3.22 in the additional notes.

■

Example 3.6.1

Let $A \in \mathbb{R}^{n\times n}$, $\boldsymbol{b} \in \mathbb{R}^n$, and $c \in \mathbb{R}$. A quadratic form $q(\boldsymbol{x}) = \boldsymbol{x}^T A\boldsymbol{x} + \boldsymbol{b}^T\boldsymbol{x} + c$ is differentiable at all points $\boldsymbol{x} \in \mathbb{R}^n$, and the gradient vector at $\boldsymbol{x}_0$ is given by $\nabla q(\boldsymbol{x}) = (A + A^T)\boldsymbol{x} + \boldsymbol{b}$. Now, if A is symmetric, then $A + A^T = 2A$, and so the gradient vector becomes $\nabla q(\boldsymbol{x}) = 2A\boldsymbol{x} + \boldsymbol{b}$. Recall that the derivative of the second degree polynomial function $p : \mathbb{R} \to \mathbb{R}, p(x) = ax^2 + bx + c$ is $p'(x) = 2ax + b$. Hence, $\nabla q(\boldsymbol{x}) = (A + A^T)\boldsymbol{x} + \boldsymbol{b}$ is a proper generalization of this well-known fact.

Similarly to what we know for functions of one variable, differentiable functions of n variables are continuous[2]:

Theorem 3.6.4

Assume that the function $f : U \to \mathbb{R}$ is differentiable at the point $\boldsymbol{x}_0 \in U$. Then f is also continuous at $\boldsymbol{x}_0$.

Proof. Assume that f is differentiable at the point $\boldsymbol{x}_0 \in U$. Then (3.30) holds for some function $\varepsilon : \mathbb{R}^n \to \mathbb{R}$ such that $\varepsilon(\boldsymbol{h}) \to 0$ as $\boldsymbol{h} \to 0$. It then follows that

$$f(\boldsymbol{x}_0 + \boldsymbol{h}) - f(\boldsymbol{x}_0) = \boldsymbol{c}^T\boldsymbol{h} + \varepsilon(\boldsymbol{h})\|\boldsymbol{h}\| \to 0 \text{ as } \boldsymbol{h} \to \boldsymbol{0},$$

i.e., the function f is indeed continuous at the point $\boldsymbol{x}_0$.

■

The concept of differentiability also yields the condition that implies that the order in which we take partial derivatives of second order is irrelevant:

Theorem 3.6.5

Let U be an open set in $\mathbb{R}^n$ and consider a differentiable function $f : U \to \mathbb{R}$ for which all the second-order partial derivatives $\frac{\partial^2 f}{\partial x_j \partial x_k}$, $j, k = 1, \ldots, n$, exist on U and are continuous. Then

$$\frac{\partial^2 f}{\partial x_j \partial x_k}(x_1, x_2, \ldots, x_n) = \frac{\partial^2 f}{\partial x_k \partial x_j}(x_1, x_2, \ldots, x_n)$$

for all $j, k = 1, \ldots, n$ and all $(x_1, x_2, \ldots, x_n) \in U$.

The proof of Theorem 3.6.5 is technical and will not be given here. Again, we refer to Theorem 3.34 in the additional notes for a proof.

[2]You should compare this fact to the issues we encountered in Example 3.3.6.

3.7 The chain rule for scalar functions

The rule (3.10) concerning differentiation of a composed function $g \circ f$ has an important generalization to the case where g is a function of several variables. For the sake of convenient notation we will only state it for a function g that is defined and differentiable on all of $\mathbb{R}^n$. We will state the general case in Theorem 3.8.4 on page 115.

Theorem 3.7.1

Consider a differentiable function $g : \mathbb{R}^n \to \mathbb{R}$, and let $f_1, f_2, \ldots f_n : \mathbb{R} \to \mathbb{R}$ denote differentiable functions. Then the composite function

$$h : \mathbb{R} \to \mathbb{R}, \quad h(x) \coloneqq g(f_1(x), f_2(x), \cdots, f_n(x)) \tag{3.31}$$

is differentiable, and

$$\begin{aligned} h'(x) = &\frac{\partial g}{\partial x_1}(f_1(x), f_2(x), \ldots, f_n(x)) f_1'(x) \\ &+ \frac{\partial g}{\partial x_2}(f_1(x), f_2(x), \ldots, f_n(x)) f_2'(x) \\ &+ \cdots + \frac{\partial g}{\partial x_n}(f_1(x), f_2(x), \ldots, f_n(x)) f_n'(x). \end{aligned} \tag{3.32}$$

The proof of Theorem 3.7.1 follows closely the classical proof of (3.10). The expression (3.32) can be written in a more compressed form using the inner product between the gradient of g at the point $(f_1(x), f_2(x), \ldots, f_n(x))$ and the vector $(f_1'(x), f_2'(x), \ldots, f_n'(x))$. Combining $\boldsymbol{f}(x) = (f_1(x), f_2(x), \ldots, f_n(x))$ into a vector function of one variable, we can express the derivative of $h(x) \coloneqq g(\boldsymbol{f}(x))$ as follows.

Corollary 3.7.2

The expression (3.32) means precisely that

$$\begin{aligned} h'(x) &= \langle (f_1'(x), f_2'(x), \ldots, f_n'(x)), \nabla g(f_1(x), f_2(x), \ldots, f_n(x)) \rangle \\ &= \langle \boldsymbol{f}'(x), \nabla g(\boldsymbol{f}(x)) \rangle = \nabla g(\boldsymbol{f}(x))^T \, \boldsymbol{f}'(x). \end{aligned} \tag{3.33}$$

The result in Theorem 3.7.1 and Corollary 3.7.2 is known as the *chain rule*. The generalized chain rule for vector functions can be found in Theorem 3.8.4 on page 115.

If the functions g and $f_1, f_2, \ldots, f_n$ in Corollary 3.7.2 are *explicitly* given, we can directly use (3.31) to obtain a formula for the function h, which then

can be differentiated "as usual" without reference to the chain rule. Indeed, let us illustrate this with an example.

Example 3.7.1

Consider the function

$$g : \mathbb{R}^2 \to \mathbb{R}, \quad g(x_1, x_2) = x_1^2 + x_2$$

and let

$$f_1(x) = \sin(x), \quad f_2(x) = \cos(x).$$

Then

$$h(x) := g(f_1(x), f_2(x)) = \sin^2(x) + \cos(x),$$

so we directly see that h is differentiable, with

$$h'(x) = 2\sin(x)\,\cos(x) - \sin(x). \tag{3.34}$$

Let us derive the same result by applying the chain rule as formulated in Corollary 3.7.2. We see that the gradient of the function g is

$$\nabla g(x_1, x_2) = (2x_1, 1),$$

so

$$\nabla g(f_1(x), f_2(x)) = (2\sin(x), 1).$$

Also,

$$(f_1'(x), f_2'(x)) = (\cos(x), -\sin(x)).$$

Thus, via (3.34)

$$\begin{aligned} h'(x) &= \langle \nabla g(f_1(x), f_2(x)), (f_1'(x), f_2'(x)) \rangle \\ &= \langle (2\sin(x), 1), (\cos(x), -\sin(x)) \rangle \\ &= 2\sin(x)\cos(x) - \sin(x), \end{aligned}$$

as we also saw in (3.34).

While the chain rule did not really make the calculations easier for us in Example 3.7.1, it is important to become familiar with the use of the gradient notation and the inner product in Equation (3.33). Indeed, as we will see in Section 4.6 on page 134, this type of notation is very convenient when considering Taylor polynomials in higher dimensions.

The chain rule has many applications, e.g., it is used in modeling physical phenomena where quantities depend on other quantities that in turn depend

on yet other quantities, and it is fundamental in the training of neural networks, as we will discuss in the following example.

Example 3.7.2 Training a neural network

Let $\boldsymbol{\Phi}_\theta : \mathbb{R}^{n_0} \to \mathbb{R}^{n_L}$ denote a neural network with L layers, where the input layer has n_0 neurons and the output layer has n_L neurons, and $\theta = \{A_\ell, \boldsymbol{b}_\ell\}_{\ell=1}^{L}$ denotes the network parameters (weights and biases). Suppose we wish to approximate an unknown target map $\boldsymbol{y} : \mathbb{R}^{n_0} \to \mathbb{R}^{n_L}$ from a finite training set $\{(\boldsymbol{x}^{(i)}, \boldsymbol{y}(\boldsymbol{x}^{(i)}))\}_{i=1}^{m}$ by choosing θ so that $\boldsymbol{\Phi}_\theta(\boldsymbol{x}^{(i)}) \approx \boldsymbol{y}(\boldsymbol{x}^{(i)})$ for all i.

To measure approximation quality, introduce a so-called loss function $L : \mathbb{R}^{n_L} \times \mathbb{R}^{n_L} \to \mathbb{R}$, e.g., the mean squared error

$$L(\boldsymbol{\Phi}_\theta(\boldsymbol{x}), \boldsymbol{y}(\boldsymbol{x})) = \tfrac{1}{2}\|\boldsymbol{\Phi}_\theta(\boldsymbol{x}) - \boldsymbol{y}(\boldsymbol{x})\|^2.$$

The factor $\frac{1}{2}$ is included only for convenience when differentiating as it cancels the 2 arising from the square. The overall loss function (called the empirical risk) $R : \mathbb{R}^N \to \mathbb{R}$ is

$$R(\theta) = \frac{1}{m}\sum_{i=1}^{m} L\Big(\boldsymbol{\Phi}_\theta(\boldsymbol{x}^{(i)}), \boldsymbol{y}(\boldsymbol{x}^{(i)})\Big),$$

where N is the total number of parameters in θ (see Equation (1.12) on page 23).

Training the neural network then reduces to minimizing the overall loss $R(\theta)$. This is usually done by iterative gradient-based method: compute the gradient $\nabla_\theta R(\theta)$ and update parameters by gradient descent, e.g.

$$\theta_{k+1} = \theta_k - \eta_k \nabla_\theta R(\theta_k),$$

with learning rate (i.e., step size) $\eta_k > 0$. This is completely analogous to the iterative updates discussed in Example 3.4.2, but here we minimize the loss rather than maximize.

Since there can be millions of parameters in θ, computing $\nabla_\theta R(\theta)$ can be very expensive. Here, the chain rule comes to the rescue as it enables efficient computation of $\nabla_\theta R(\theta)$ by re-using intermediate layer outputs; applying it backward through the network yields the backpropagation algorithm, which computes the gradients of the loss with respect to *each* entry in A_ℓ and $\boldsymbol{b}_\ell$. These gradients are then used by the gradient descent method (and its variants[a]) to update the parameters and improve the network's performance. We will not go into details here on how this is done as it is beyond the scope of this book.

[a] In applications, one typically uses the *stochastic gradient descent* (SGD), which updates a *subset* of parameters using a randomly selected *subset* of training data at each iteration, rather than the entire training set.

3.8 Differentiability of vector functions of several variables

We will now turn to differentiability of a *vector* function $\boldsymbol{f} = (f_1, \ldots, f_k) : U \to \mathbb{R}^k$, where U is an open set in $\mathbb{R}^n$. We again need to generalize Equation (3.30) on page 106, this time to an equality in $\mathbb{R}^k$ since $\boldsymbol{f}(\boldsymbol{x}_0) \in \mathbb{R}^k$. The modifications are as follows.

1) We will replace the linear map $\boldsymbol{h} \mapsto \boldsymbol{c}^T\boldsymbol{h}$, $\mathbb{R}^n \to \mathbb{R}$ with a linear map $L : \mathbb{R}^n \to \mathbb{R}^k$, i.e., $\boldsymbol{h} \mapsto L(\boldsymbol{h})$.

2) We will replace the function $\varepsilon : \mathbb{R}^n \to \mathbb{R}$ by a vector function $\boldsymbol{\varepsilon} : \mathbb{R}^n \to \mathbb{R}^k$.

Definition 3.8.1 Differentiability of vector functions

Consider a vector function $\boldsymbol{f} : U \to \mathbb{R}^k$, where U is an open set in $\mathbb{R}^n$, and let $\boldsymbol{x}_0 \in U$. We say that the function $\boldsymbol{f}$ is differentiable at the point $\boldsymbol{x}_0$ if there exist a linear map $L : \mathbb{R}^n \to \mathbb{R}^k$ and a function $\boldsymbol{\varepsilon} : \mathbb{R}^n \to \mathbb{R}^k$ such that $\boldsymbol{\varepsilon}(\boldsymbol{h}) \to \boldsymbol{0}$ as $\boldsymbol{h} \to \boldsymbol{0}$ and

$$\boldsymbol{f}(\boldsymbol{x}_0 + \boldsymbol{h}) - \boldsymbol{f}(\boldsymbol{x}_0) - L(\boldsymbol{h}) - \boldsymbol{\varepsilon}(\boldsymbol{h})\|\boldsymbol{h}\| = 0, \quad \text{for all } \boldsymbol{h} \in \mathbb{R}^n. \tag{3.35}$$

The linear map L is called the *differential* of $\boldsymbol{f}$ at $\boldsymbol{x}_0$ and is denoted $\mathrm{d}\boldsymbol{f}_{\boldsymbol{x}_0}$. From Lemma 11.3.2 in Mathematics 1a we know that L is of the form $L(\boldsymbol{h}) = \boldsymbol{J}_{\boldsymbol{f}}\boldsymbol{h}$ where $\boldsymbol{J}_{\boldsymbol{f}} \in \mathsf{M}_{k\times n}(\mathbb{R})$ is a unique matrix called the *Jacobian matrix*.

We already know examples of vector functions that can be differentiated. Indeed, a linear map, or more generally an affine map, is differentiable:

Lemma 3.8.1 Linear and affine maps are differentiable

If $L : \mathbb{R}^n \to \mathbb{R}^k$ is linear, then it is differentiable at all points $\boldsymbol{x}_0 \in \mathbb{R}^n$ and the differential is $\mathrm{d}L_{\boldsymbol{x}_0} = L$. That is, if $A \in \mathsf{M}_{k\times n}(\mathbb{R})$, then the linear map $\boldsymbol{x} \mapsto A\boldsymbol{x}, \mathbb{R}^n \to \mathbb{R}^k$ is a differentiable vector function at all points $\boldsymbol{x}_0 \in \mathbb{R}^n$, and its Jacobian matrix is (the constant) matrix A. More generally, the affine map $T_{A,\boldsymbol{b}} : \mathbb{R}^n \to \mathbb{R}^k$, defined by $T_{A,\boldsymbol{b}}(\boldsymbol{x}) = A\boldsymbol{x} + \boldsymbol{b}$, where $\boldsymbol{b} \in \mathbb{R}^k$, is also a differentiable vector at all points $\boldsymbol{x}_0 \in \mathbb{R}^n$, and its Jacobian matrix is matrix A.

Proof. Since $L : \mathbb{R}^n \to \mathbb{R}^k$ is linear it satisfies

$$L(\boldsymbol{x}_0 + \boldsymbol{h}) - L(\boldsymbol{x}_0) - L(\boldsymbol{h}) = \boldsymbol{0}$$

for all $\boldsymbol{x}_0, \boldsymbol{h} \in \mathbb{R}^n$. Thus, if we take $\boldsymbol{\varepsilon}$ to be the zero vector function $\boldsymbol{\varepsilon}(\boldsymbol{h}) = \mathbf{0}$ for all $\boldsymbol{h} \in \mathbb{R}^n$, then (3.35) is satisfied. From $L(\boldsymbol{h}) = A\boldsymbol{h}$, we see that the Jacobian matrix is A. The proof for affine maps is similar.

■

As for scalar functions, differentiability of vector functions implies continuity.

Theorem 3.8.2

Assume that the function $\boldsymbol{f} : U \to \mathbb{R}^k$ is differentiable at the point $\boldsymbol{x}_0 \in U$. Then $\boldsymbol{f}$ is also continuous at $\boldsymbol{x}_0$.

Proof. The proof of Theorem 3.6.4 on page 108 carries over verbatim to the setting of Theorem 3.8.2 so we will not repeat it here.

■

In terms of the Jacobian matrix Equation (3.35) reads

$$\boldsymbol{f}(\boldsymbol{x} + \boldsymbol{h}) = \boldsymbol{f}(\boldsymbol{x}) + \boldsymbol{J}_{\boldsymbol{f}}\boldsymbol{h} + \boldsymbol{\varepsilon}(\boldsymbol{h})\|\boldsymbol{h}\|. \tag{3.36}$$

We will return to this expressions in Section 4.7 on page 136.

As for scalar functions, where $\boldsymbol{c}$ in (3.30) was equal to the gradient vector, the Jacobian matrix also has a very special form. If $\boldsymbol{f} : U \to \mathbb{R}^k$ is differentiable at $\boldsymbol{x}$, the Jacobian matrix $\boldsymbol{J}_{\boldsymbol{f}}$ is a $k \times n$ matrix given by:

$$\boldsymbol{J}_{\boldsymbol{f}} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1}(\boldsymbol{x}) & \frac{\partial f_1}{\partial x_2}(\boldsymbol{x}) & \cdots & \frac{\partial f_1}{\partial x_n}(\boldsymbol{x}) \\ \frac{\partial f_2}{\partial x_1}(\boldsymbol{x}) & \frac{\partial f_2}{\partial x_2}(\boldsymbol{x}) & \cdots & \frac{\partial f_2}{\partial x_n}(\boldsymbol{x}) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_k}{\partial x_1}(\boldsymbol{x}) & \frac{\partial f_k}{\partial x_2}(\boldsymbol{x}) & \cdots & \frac{\partial f_k}{\partial x_n}(\boldsymbol{x}) \end{bmatrix}. \tag{3.37}$$

The proof is similar to the proof of Theorem 3.6.2 on page 106 and will not be repeated here. Note that $\boldsymbol{J}_{\boldsymbol{f}}$ consists of gradient vectors of f_i as row vectors. This shows that if we take $k = 1$ in Definition 3.8.1, we recover Definition 3.6.1 on page 106.

As for continuity of vector functions, we can also check differentiability of a vector function by looking at each coordinate separately:

Theorem 3.8.3 Differentiability of coordinate functions

A vector function $\boldsymbol{f} = (f_1, \ldots, f_k) : U \to \mathbb{R}^k$ is differentiable at $\boldsymbol{x}_0$ if and only if all the functions $f_i : U \to \mathbb{R}$, $i = 1, \ldots, k$ are differentiable at $\boldsymbol{x}_0$.

The proof is similar to the proof of Theorem 3.2.1 on page 88 and will not be given here.

When speaking of the Jacobian matrix, we should really say "the Jacobian matrix evaluated at $\boldsymbol{x}$" as the matrix clearly depends on $\boldsymbol{x}$. If we want to stress this fact, we will write $\boldsymbol{J_f}(\boldsymbol{x})$. Consider a vector function $\boldsymbol{f} : U \to \mathbb{R}^k$, where U is an open set in $\mathbb{R}^n$. The map $\boldsymbol{x} \mapsto \boldsymbol{J_f}(\boldsymbol{x})$, $U \mapsto \mathbb{R}^{k\times n}$ is then called the *differential* of $\boldsymbol{f}$, and it is denoted $\mathrm{d}\boldsymbol{f}_{\boldsymbol{x}} : U \mapsto \mathbb{R}^{k\times n}$. This map is clearly only well-defined for vector functions differentiable at *all* points $\boldsymbol{x} \in U$.

Definition 3.8.2 C^1 **vector function**

Let U be an open set in $\mathbb{R}^n$. A vector function $\boldsymbol{f} : U \to \mathbb{R}^k$ is called *differentiable* if it is differentiable at all points $\boldsymbol{x} \in U$. If all partial derivatives $\frac{\partial f_i}{\partial x_j} : U \to \mathbb{R}$ appearing in the Jacobian matrix are continuous functions, then the vector function $\boldsymbol{f}$ is said to be *continuously differentiable* or, alternatively, $\boldsymbol{f}$ is said to be a C^1 vector function.

A note on terminology: the class of C^0 vector functions consists of all continuous vector functions and the class C^1 consists of all differentiable functions whose derivative is continuous (such functions are also called continuously differentiable). More generally, the class of C^m vector functions consists of functions for which all partial derivatives up to order m of *each* coordinate function, that is, partial derivatives of the form

$$\frac{\partial^\ell f_i}{\partial x_{j_1}\partial x_{j_2}\cdots\partial x_{j_\ell}}, \quad \ell = 1, 2, \ldots, m, \quad i = 1, 2, \ldots, k,$$

where $j_1, j_2, \ldots, j_\ell$ are indices from the set $\{1, 2, \ldots, n\}$, *exist* and are *continuous*.

Definition 3.8.3 C^m **vector function**

Let $m \in \mathbb{N}$, and let U be an open set in $\mathbb{R}^n$. A vector function $\boldsymbol{f} : U \to \mathbb{R}^k$ is said to be a C^m vector function if all partial derivatives up to order m exist and are continuous functions on U . The vector function $\boldsymbol{f}$ is said to be of class C^∞ (i.e., smooth or infinitely differentiable) if it is of class C^m for all values of m.

Just like for scalar functions, differentiability is preserved by the usual arithmetic operations, e.g., if $\boldsymbol{f} : U \to \mathbb{R}^k$ and $\boldsymbol{g} : U \to \mathbb{R}^k$ are differentiable at $\boldsymbol{x}_0 \in U$, then so is $\boldsymbol{f} + \boldsymbol{g}$.

Exercise 3.8.1 softmax is continuously differentiable

Consider the softmax function defined in Example 1.3.3 on page 20. Show that the softmax function is a C^1 vector function on $\mathbb{R}^n$ and compute its Jacobian matrix. Can you argue that the softmax function is, in fact, a C^∞ vector function?

Moreover, also the composition of differentiable vector functions is differentiable. This result is the *generalized chain rule*.

Theorem 3.8.4 Generalized chain rule

Let $U \subseteq \mathbb{R}^n$ and $V \subseteq \mathbb{R}^k$ be open sets. Suppose that $\boldsymbol{f} : U \to \mathbb{R}^k$ and $\boldsymbol{g} : V \to \mathbb{R}^\ell$ are vector functions such that $\operatorname{im}(f) \subseteq V$, $\boldsymbol{f}$ is differentiable at $\boldsymbol{x}_0 \in U$, and $\boldsymbol{g}$ is differentiable at $\boldsymbol{y}_0 \coloneqq \boldsymbol{f}(\boldsymbol{x}_0)$. Then $\boldsymbol{g} \circ \boldsymbol{f} : U \to \mathbb{R}^\ell$ is differentiable at $\boldsymbol{x}_0$ with differential $\mathrm{d}(\boldsymbol{g} \circ \boldsymbol{f})_{x_0} = \mathrm{d}\boldsymbol{g}_{y_0} \circ \mathrm{d}\boldsymbol{f}_{x_0}$ and Jacobian matrix

$$\boldsymbol{J}_{\boldsymbol{g}\circ\boldsymbol{f}}(\boldsymbol{x}_0) = \boldsymbol{J}_{\boldsymbol{g}}(\boldsymbol{y}_0)\boldsymbol{J}_{\boldsymbol{f}}(\boldsymbol{x}_0), \tag{3.38}$$

where $\boldsymbol{J}_{\boldsymbol{g}\circ\boldsymbol{f}} \in \mathsf{M}_{\ell\times n}(\mathbb{R})$, $\boldsymbol{J}_{\boldsymbol{g}} \in \mathsf{M}_{\ell\times k}(\mathbb{R})$, and $\boldsymbol{J}_{\boldsymbol{f}} \in \mathsf{M}_{k\times n}(\mathbb{R})$.

The proof is not particularly difficult, however, for brevity, we will not give it here, but refer to Theorem 3.28 in the additional notes. The generalized chain rule shows that the Jacobian matrix of composed vector functions is just a matrix-matrix product of other Jacobian matrices (evaluated at two different points). Note that, compared to the chain rule for scalar functions of one variable in Equation (3.10) on page 84, the order of the Jacobian matrices in (3.38) is important. Indeed, $\boldsymbol{J}_{\boldsymbol{g}}(\boldsymbol{y}_0)$ is a matrix of size $\ell \times k$, while $\boldsymbol{J}_{\boldsymbol{f}}(\boldsymbol{x}_0)$ is of size $k \times n$. Hence, if $\ell \neq n$, only the matrix product in (3.38) is well-defined.

By inductive application of Theorem 3.8.4, we can also show that the composition of C^m vector functions is again a C^m vector function.

Theorem 3.8.5 Composition of C^m vector functions

Let $U \subseteq \mathbb{R}^n$ and $V \subseteq \mathbb{R}^k$ be open sets. Suppose that $\boldsymbol{f} : U \to V$ and $\boldsymbol{g} : V \to \mathbb{R}^\ell$ are C^m vector functions. Then the composition $\boldsymbol{h} = \boldsymbol{g} \circ \boldsymbol{f}, U \to \mathbb{R}^\ell$ is a C^m vector function.

Example 3.8.2 Differentiability of neural networks

Consider a neural network $\boldsymbol{\Phi} : \mathbb{R}^{n_0} \to \mathbb{R}^{n_L}$ with L layers as defined in Definition 1.3.2 on page 22. Assume that the activation functions $\boldsymbol{\sigma}_\ell : \mathbb{R}^{n_\ell} \to \mathbb{R}^{n_\ell}$ used in each layer are C^∞ functions, i.e., infinitely often

differentiable. That is, $\boldsymbol{\sigma}_\ell$ could be the sigmoid function or the hyperbolic tangent function. Then the neural network $\boldsymbol{\Phi}$ is a differentiable vector function, in fact, it is a C^∞ vector function. To see this, note that $\boldsymbol{\Phi}$ is a composition of affine linear maps and smooth activation functions. Each affine linear map is differentiable (Lemma 3.8.1), and the activation functions are differentiable by assumption. By the generalized chain rule the function $\boldsymbol{\Phi}$ is differentiable. Affine maps are actually C^m vector functions for any $m \in \mathbb{N}$, and since the activation functions are also C^m by assumption, it follows from Theorem 3.8.5 that $\boldsymbol{\Phi}$ is a C^m vector function for any $m \in \mathbb{N}$, i.e., a C^∞ vector function.

On the other hand, if the ReLU activation function ReLU is used in any layer, the neural network $\boldsymbol{\Phi}$ will be C^0 but not C^1, since ReLU is continuous but not differentiable at 0.

CHAPTER 4

Taylor Approximation

In mathematics, physics, and engineering we often face functions that are complicated to work with. The reason can be, e.g., that the function is given by a complicated expression, or that we only know the function in terms of its graph. In such cases it is helpful if we can approximate the function with a simpler function, for example, a polynomial. The precise meaning of this will be clear soon, but the idea is to search for a polynomial P such that the deviation between the function values $f(x)$ of the given "complicated function" f and $P(x)$ is small for all x in the considered interval. In many cases it is then possible to do the further calculations with the polynomial P instead of the function f, and still obtain useful and realistic results.

The starting point of Taylor approximations is the notion of differentiability from Definition 3.8.1 on page 112. Suppose $\boldsymbol{f}$ is differentiable at $\boldsymbol{x}_0 \in U$ as in Definition 3.8.1. By a simple rearrangement of the terms in (3.35), we get, by setting $\boldsymbol{x} := \boldsymbol{x}_0 + \boldsymbol{h}$,

$$\begin{aligned} \boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{f}(\boldsymbol{x}_0 + \boldsymbol{h}) &= \boldsymbol{f}(\boldsymbol{x}_0) + \boldsymbol{J}_{\boldsymbol{f}}(\boldsymbol{x}_0)(\boldsymbol{h}) + \boldsymbol{\varepsilon}(\boldsymbol{h})\|\boldsymbol{h}\| \\ &= \boldsymbol{f}(\boldsymbol{x}_0) + \boldsymbol{J}_{\boldsymbol{f}}(\boldsymbol{x}_0)(\boldsymbol{x} - \boldsymbol{x}_0) + \boldsymbol{\varepsilon}(\boldsymbol{x} - \boldsymbol{x}_0)\|\boldsymbol{x} - \boldsymbol{x}_0\| \end{aligned} \tag{4.1}$$

for $\boldsymbol{h} \in \mathbb{R}^n$. The goal is to generalize the formula (4.1) to higher order derivatives and, in particular, to quantify the behavior of the $\boldsymbol{\varepsilon}$-function. For now, we just know that $\boldsymbol{\varepsilon}(\boldsymbol{x} - \boldsymbol{x}_0) \to \boldsymbol{0}$ as $\boldsymbol{x} - \boldsymbol{x}_0 \to \boldsymbol{0}$, i.e., as $\boldsymbol{x} \to \boldsymbol{x}_0$. The function $P_{1,\boldsymbol{f},\boldsymbol{x}_0} : \mathbb{R}^n \to \mathbb{R}^k$ defined by $P_{1,\boldsymbol{f},\boldsymbol{x}_0}(\boldsymbol{x}) = \boldsymbol{f}(\boldsymbol{x}_0) + \boldsymbol{J}_{\boldsymbol{f}}(\boldsymbol{x}_0)(\boldsymbol{x} - \boldsymbol{x}_0)$ for $\boldsymbol{x} \in \mathbb{R}^n$ is the so-called first-degree Taylor polynomial for $\boldsymbol{f}$ at the point $\boldsymbol{x}_0$. Hence, from Chapter 3, we already know that, for small $\boldsymbol{h} = \boldsymbol{x} - \boldsymbol{x}_0$, i.e., for $\boldsymbol{x}$ close to $\boldsymbol{x}_0$, we can approximate $\boldsymbol{f}(\boldsymbol{x})$ by a first degree polynomial $P_{1,\boldsymbol{f},\boldsymbol{x}_0}(\boldsymbol{x})$ (in n variables) and the error $\boldsymbol{f}(\boldsymbol{x}) - P_{1,\boldsymbol{f},\boldsymbol{x}_0}(\boldsymbol{x})$ goes to zero faster than $\|\boldsymbol{h}\|$. As we will see, we can obtain better approximations using higher degree polynomials.

However, let us take a step back and first consider the simple setup of approximating a scalar function f of *one* variable by a first-degree polynomial. This is just the well-known tangent line for a function of one variable and the topic of Section 4.1. This approximation is generalized and improved in Section 4.2, where we approximate the function f by higher-degree polynomials, the so-called Taylor polynomials. The associated error estimates, i.e., the estimates of the deviation between the function values $f(x)$ and the values $P(x)$ of the Taylor polynomials, are considered in Section 4.3. In the Sections 4.4 to 4.6 we run precisely the same program for functions of several variables. Finally, in Section 4.7 we return to Taylor polynomials for vector functions.

The Taylor polynomials we will encounter in this chapter depend on the given function $\boldsymbol{f}$, the desired degree of the Taylor polynomial K, and a fixed "expansion" point $\boldsymbol{x}_0$ in the domain of $\boldsymbol{f}$. The polynomial will be denoted by $P_{K,\boldsymbol{f},\boldsymbol{x}_0}(\boldsymbol{x})$, $P_{K,\boldsymbol{f}}(\boldsymbol{x})$, or simply $P_K(\boldsymbol{x})$ if $\boldsymbol{f}$ and $\boldsymbol{x}_0$ are clear from the context.

4.1 The tangent for a function of one variable

Definition 4.1.1

Let $I \subset \mathbb{R}$ denote an interval and consider a differentiable function $f : I \to \mathbb{R}$. Fix a point $x_0 \in I$. Then the first-degree approximating polynomial — or the *Taylor polynomial of first degree* — at the point x_0 is defined by

$$P_{1,f,x_0}(x) = P_1(x) = f(x_0) + f'(x_0)(x - x_0), \ x \in \mathbb{R}. \tag{4.2}$$

The graph of the function P_1, i.e., the straight line given by

$$y = f(x_0) + f'(x_0)(x - x_0), \ x \in \mathbb{R}, \tag{4.3}$$

is called the *tangent* of the function f at the point x_0.

Remark 4.1.1

The form of the polynomial $P_1 = P_{1,f,x_0}$ in (4.2) implies that two very particular properties hold, namely,

$$P_1(x_0) = f(x_0) \ \text{ and } \ P_1'(x_0) = f'(x_0).$$

That is, the function value of P_1 at the point x_0 coincides with the function value $f(x_0)$, i.e., the graph of the approximating polynomial P_1 goes through

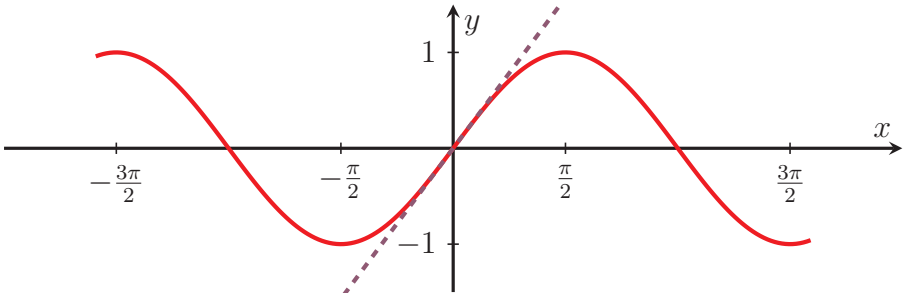


Figure 4.1: The function $f(x) = \sin(x)$ and its first-degree Taylor polynomial at $x_0 = 0$ (dashed).

the point $(x_0, f(x_0))$. The fact that also the first derivative of P_1 and f coincide at x_0, implies that the polynomial P_1 provides an approximation of the function f in a (small) neighborhood of x_0. The graph of P_1 it is a straight line with slope $f'(x_0)$; see the subsequent Example 4.1.1 for an illustration of this. In particular, the slope $f'(x_0)$ tells how the function f evolves around the point x_0 : for example, a positive value of $f'(x_0)$ means that the function f is increasing at the point x_0.

Example 4.1.1

Consider the function $f : \mathbb{R} \to \mathbb{R}$ given by $f(x) = \sin(x)$, $x \in \mathbb{R}$. Then $f'(x) = \cos(x)$. If we choose $x_0 = 0$, the Taylor polynomial of first degree at x_0 is given by

$$P_1(x) = \sin(0) + \cos(0)(x - 0) = x,\ x \in \mathbb{R}.$$

We see from Figure 4.1 that P_1 indeed provides a quite good approximation of the function f for x close to $x_0 = 0$.

In practice it depends on a given application whether we can consider the Taylor polynomial of first degree as a satisfying approximation of a given function. Depending on the function f, the interval where P_1 is approximating the function f might be very small:

Example 4.1.2

Consider the function $f : \mathbb{R} \setminus \{0\} \to \mathbb{R}$ given by

$$f(x) = \frac{1}{x} \sin\left(\frac{1}{x}\right),\ x \neq 0. \tag{4.4}$$

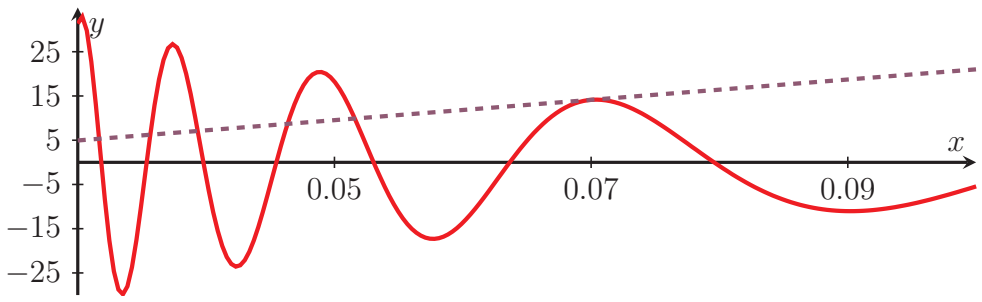


Figure 4.2: The function f in (4.4) and its first-degree Taylor polynomial at $x = 0.07$, shown on the interval $[0.03, 0.1]$ (dashed).

The graph of f is sketched on Figure 4.2 for $x \in [0.03]0.1$; note that the function has "infinitely" many oscillations around $x = 0$. Due to the strong oscillations of the function f it is clear that for values of x_0 close to 0, the Taylor polynomial of first degree at x_0 can only provide us with a reasonable approximation in a very small interval containing x_0. In other words: overall, we can not claim that the first-degree Taylor polynomial provides us with a good approximation of the function f. See Figure 4.2 for an illustration of this.

We also note that the information provided by the first-degree Taylor polynomial is too limited to distinguish even very different functions:

Example 4.1.3

Figure 4.3 shows the graphs of the functions

$$f_1(x) = x^2, \quad f_2(x) = -x^2, \quad f_3(x) = x^3. \tag{4.5}$$

It is clear that these three functions behave very differently; nevertheless they all have the same first-degree Taylor polynomial at the point $x_0 = 0$, namely, $P_1(x) = 0$.

Motivated by our observations in Examples 4.1.2 and 4.1.3 we will now turn our attention to a more general theory for approximation using higher-degree polynomials.

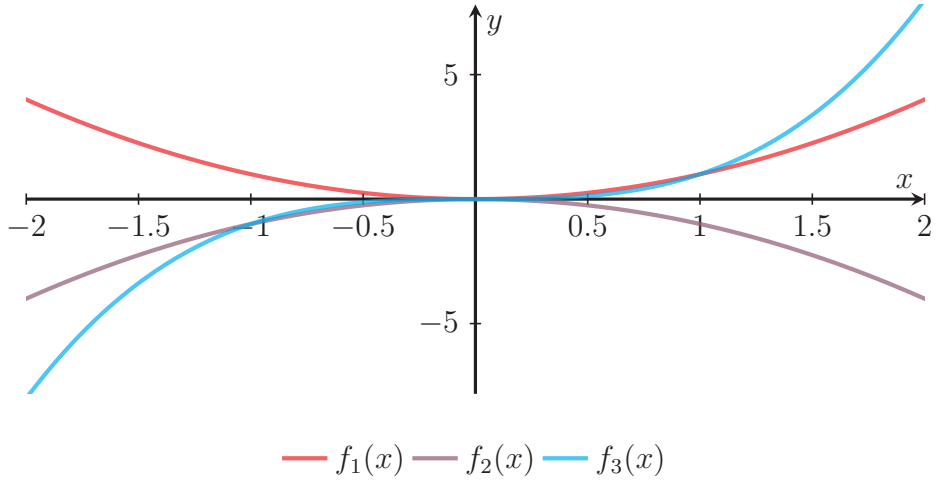


Figure 4.3: The functions in (4.5).

4.2 Taylor polynomials for functions of one variable

Examples 4.1.2 and 4.1.3 demonstrate clearly that we need to develop a finer approximation theory than the one provided by first-degree Taylor polynomials. In order to do so, it is natural to use the expression (4.2) as the starting point. Indeed, as we saw in Remark 4.1.1 the graph of P_1 goes through the point $(x_0, f(x_0))$ and has the "correct slope", so maybe we can obtain better approximation properties by adding "small correction terms" of the form $a(x-x_0)^2, b(x-x_0)^3, c(x-x_0)^4$, etc? This turns out indeed to be the case whenever the constants a, b, c etc are chosen correctly. We will now provide a general formula for an approximating polynomial of degree K for any positive integer K. In order to do so, we need to assume that the function f is K times differentiable; we will denote its derivatives at the point x_0 by $f'(x_0), f^{(2)}(x_0), f^{(3)}(x_0)$, etc.

Definition 4.2.1 Taylor polynomial of Kth degree

Fix a positive integer K. Let $I \subset \mathbb{R}$ denote an interval and consider a K times differentiable function $f : I \to \mathbb{R}$. Fix a point $x_0 \in I$. Then the *Taylor polynomial of Kth degree* $P_{K,f,x_0} = P_K$ at the point x_0 is defined by

$$P_K(x) = f(x_0) + f'(x_0)(x-x_0) + \frac{f^{(2)}(x_0)}{2!}(x-x_0)^2 + \cdots + \frac{f^{(K)}(x_0)}{K!}(x-x_0)^K. \tag{4.6}$$

Recall that for a positive integer K, the expression $K!$ (pronounced "K factorial") is defined by

$$K! = K(K-1)(K-2)\cdots 1.$$

For convenience we also define $0! = 1$. In particular, we thus have

$$0! = 1,\ \ 1! = 1,\ \ 2! = 2,\ \ 3! = 6,\ \ 4! = 24,\ \ 5! = 120,\ \ 6! = 720.$$

Note that for $K = 1$ the expression (4.6) indeed corresponds to our definition of the first degree Taylor polynomial. The Kth Taylor polynomial can be written in a more compressed form as

$$P_K(x) = \sum_{k=0}^{K} \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k. \tag{4.7}$$

The compressed expression (4.7) allows to work with the Kth degree Taylor polynomial on a computer, using only little extra effort compared with the first-degree Taylor polynomials.

Example 4.2.1

Consider the function

$$f(x) = \mathrm{e}^x,\ \ x \in \mathbb{R}.$$

The function is arbitrarily often differentiable, and

$$f'(x) = f^{(2)}(x) = \cdots = f^{(k)}(x) = \cdots = \mathrm{e}^x\,.$$

For $K \in \mathbb{N}$ the Kth Taylor polynomial at $x_0 = 0$ is given by

$$\begin{aligned} P_K(x) &= f(0) + \frac{f'(0)}{1!}x + \frac{f''(0)}{2!}x^2 + \cdots + \frac{f^{(K)}(0)}{K!}x^K \\ &= 1 + x + \frac{1}{2}x^2 + \frac{1}{3!}x^3 + \cdots + \frac{1}{K!}x^K \\ &= \sum_{k=0}^{K} \frac{1}{k!}x^k. \end{aligned} \tag{4.8}$$

Figure 4.4 shows that the fifth-degree Taylor polynomial at $x_0 = 1$ gives a very good approximation to the function in a quite large interval containing $x_0 = 1$.

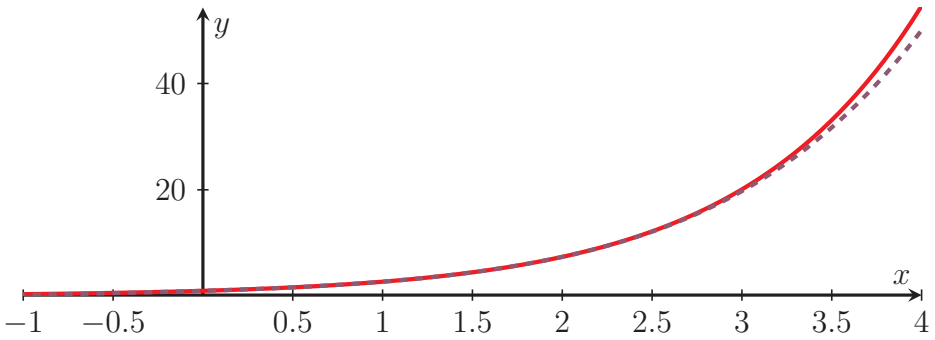


Figure 4.4: The function $f(x) = \mathrm{e}^x$ and its fifth-degree Taylor polynomial at $x_0 = 1$ (dashed).

Example 4.2.2

For a function like in Example 4.1.2,

$$f(x) = \frac{1}{x}\sin\left(\frac{1}{x}\right),\ x \neq 0, \tag{4.9}$$

it is cumbersome to calculate higher order derivatives by hand, but for a mathematical computer program it is easy to do so and also to plot the graph. Figure 4.5 shows the 4th degree Taylor polynomial at the point $x_0 = 0.07$. Compared with the first-degree Taylor polynomial shown on Figure 4.2, we see that we get a good approximation of the function f over a significantly larger interval.

Remark 4.2.1

Note the very particular form of the "correction terms" in (4.6): each term of the form $(x - x_0)^k,\ k = 0, 1, \ldots, K$, is multiplied with the corresponding factor $\frac{f^{(k)}(x_0)}{k!}$. This particular structure implies (compare with Remark 4.1.1 on page 118) that the function value and the value of the first K derivatives of the polynomial P_K and the function f *agree* at the point x_0, i.e.,

$$\begin{aligned} P_K(x_0) &= f(x_0), \\ P_K'(x_0) &= f'(x_0), \\ P_K^{(2)}(x_0) &= f^{(2)}(x_0) \\ &\ldots, \\ P_K^{(K)}(x_0) &= f^{(K)}(x_0). \end{aligned}$$

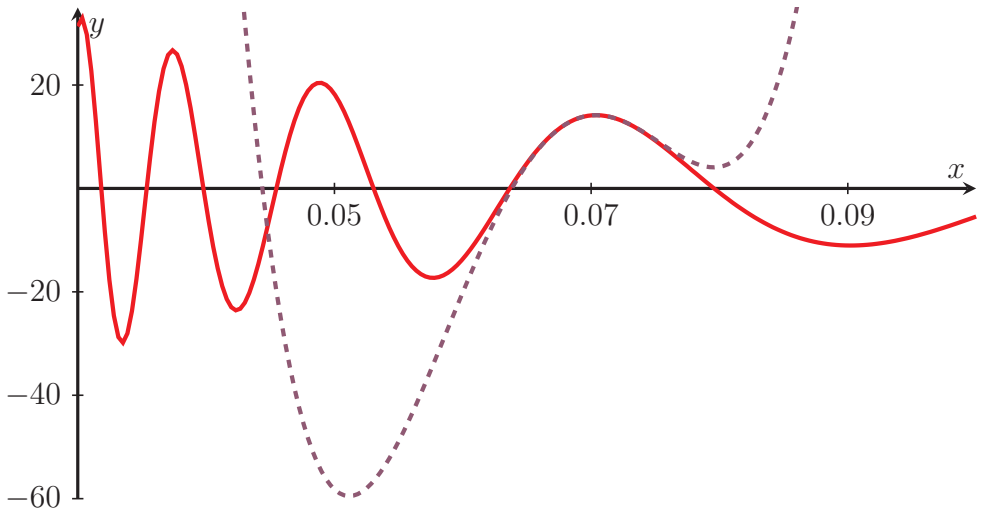


Figure 4.5: The function f in (4.9) and its fourth-degree Taylor polynomial at $x = 0.07$.

4.3 Taylor's formula for functions of one variable

Examples 4.2.1 and 4.2.2 indicate that we indeed improve the approximation quality by considering higher degree Taylor polynomials. However, in order to apply the theory we need to have more exact knowledge about the approximation quality. In other words: we need to be able to estimate the difference between the function value $f(x)$ and the corresponding approximations $P_K(x)$ for x belonging to a suitable interval containing the point x_0. In order to do so, we define the so-called *remainder term* by

$$R_K(x) := f(x) - P_K(x) = f(x) - \sum_{k=0}^{K} \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k. \tag{4.10}$$

Our goal is to develop methods to answer the following type of questions:

- Given a fixed value of $K \in \mathbb{N}$, how do we estimate the maximal deviation $|f(x) - P_K(x)|$ between $f(x)$ and $P_K(x)$ over a given interval I?
- Given a fixed value of $K \in \mathbb{N}$, how do we specify an interval containing the point x_0 such that the maximal deviation between $f(x)$ and $P_k(x)$, measured over the interval, is below a certain prescribed threshold?

- Given a fixed interval containing x_0, how large do we have to choose the degree K of the Taylor polynomial P_K such that the maximal deviation between $f(x)$ and $P_K(x)$ is below a certain threshold, measured over the entire interval?

The technical tool to answer such questions is known under the name *Taylor's formula*.

Lemma 4.3.1 Taylor's formula

Let $I \subset \mathbb{R}$ denote an interval and assume that the function $f : I \to \mathbb{R}$ is arbitrarily often differentiable. Let $x_0 \in I$ and $K \in \mathbb{N}$. Then, for each $x \in I$ there exists a scalar ξ between x and x_0 such that

$$R_K(x) = f(x) - P_K(x) = \frac{f^{(K+1)}(\xi)}{(K+1)!}(x - x_0)^{K+1}. \tag{4.11}$$

The proof of Lemma 4.3.1 is quite lengthy and is given in Appendix A.1, see Lemma A.1.3. Note that Taylor's formula does not provide us with a completely explicit expression for the difference $f(x) - P_K(x)$, it simple says that when we *fix* $x \in I$, *there exists* a corresponding ξ such that (4.11) holds. Taylor's formula does not tell us how we can choose ξ; it just tells us that such a ξ exists. In applications of Lemma 4.3.1 it is typically a technical challenge that different choices of x also lead to different values of ξ. Before we show how to apply Taylor's formula, let us restate the result:

Lemma 4.3.2 Taylor's formula

Let $I \subset \mathbb{R}$ denote an interval and assume that the function $f : I \to \mathbb{R}$ is arbitrarily often differentiable. Let $x_0 \in I$ and $K \in \mathbb{N}$. Then there exists a function $\varepsilon_K : \mathbb{R} \to \mathbb{R}$ such that

$$R_K(x) = f(x) - P_K(x) = \varepsilon_K(x - x_0)\,(x - x_0)^K, \quad \text{for all } x \in I, \tag{4.12}$$

and $\varepsilon_K(x - x_0) \to 0$ as $x - x_0 \to 0$, i.e., $x \to x_0$.

Proof. Fix some $x_0 \in I$. The scalar ξ in the expression (4.11) actually depends on the considered $x \in I$ *and* the parameter K so for the sake of clarity of the proof we will denote it by $\xi_{x,K}$. Thus,

$$R_K(x) = \frac{f^{(K+1)}(\xi_{x,K})}{(K+1)!}(x - x_0)^{K+1},$$

and hence

$$R_K(x) = (x - x_0)^K \left[\frac{f^{(K+1)}(\xi_{x,K})}{(K+1)!}(x - x_0) \right].$$

This corresponds precisely to (4.12) with

$$\varepsilon_K(x - x_0) := \frac{f^{(K+1)}(\xi_{x,K})}{(K+1)!}(x - x_0).$$

Also, when $x \to x_0$, it follows that $\xi_{x,K} \to x_0$. The function $f^{(K+1)}$ is differentiable by assumption, and hence continuous; therefore $f^{(K+1)}(\xi_x) \to f^{(K+1)}(x_0)$ as $x \to x_0$. This implies that

$$\varepsilon_K(x - x_0) = \frac{f^{(K+1)}(\xi_{x,K})}{(K+1)!}(x - x_0) \to \frac{f^{(K+1)}(x_0)}{(K+1)!} \cdot 0 = 0,$$

as $x \to x_0$, i.e., as $x - x_0 \to 0$, as claimed.

■

We are now ready to address the question of how to apply Taylor's formula to estimate the maximal deviation $|f(x) - P_K(x)|$ over a given interval. This is done via *Taylor's theorem*, stated next. It shows that under certain conditions, we can find a polynomial which is as close to f in a bounded interval as we wish:

Theorem 4.3.3 Taylor's theorem

Let $I \subset \mathbb{R}$ be an interval. Assume that the function $f : I \to \mathbb{R}$ is arbitrarily often differentiable and that there exists a constant $C > 0$ such that

$$\left| f^{(k)}(x) \right| \leq C, \quad \text{for all } k \in \mathbb{N} \text{ and all } x \in I. \tag{4.13}$$

Let $x_0 \in I$. Then for all $K \in \mathbb{N}$ and all $x \in I$,

$$\left| f(x) - \sum_{k=0}^{K} \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k \right| \leq \frac{C}{(K+1)!} |x - x_0|^{K+1}. \tag{4.14}$$

In particular, if I is a bounded interval, there exists for arbitrary $\epsilon > 0$ an $K_0 \in \mathbb{N}$ such that

$$\left| f(x) - \sum_{k=0}^{K} \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k \right| \leq \epsilon \quad \text{for all } x \in I \text{ and all } K \geq K_0. \tag{4.15}$$

The proof of Theorem 4.3.3 is given in Appendix A.1 on page 211.

Note that in case we can choose $x_0 = 0$, Taylor's theorem takes a slightly simpler form. Under the assumptions in Theorem 4.3.3 we can then approximate f near $x_0 = 0$ as well as we like with a polynomial

$$P_K(x) = \sum_{k=0}^{K} \frac{f^{(k)}(0)}{k!} x^k, \tag{4.16}$$

by choosing the degree K sufficiently high.

The inequality (4.14) can be used to decide how many terms we should include in the Taylor polynomial in order to reach a certain approximation of a given function on a fixed interval. The following example illustrates this for the exponential function, and also shows how the number of terms depend on the considered interval and on how well we want the function to be approximated.

Example 4.3.1

Consider the function

$$f(x) = \mathrm{e}^x, \ x \in \mathbb{R}.$$

We saw in Example 4.2.1 that the Kth Taylor polynomial at $x_0 = 0$ is given by

$$\begin{aligned} P_K(x) &= 1 + x + \frac{1}{2}x^2 + \frac{1}{3!}x^3 + \cdots + \frac{1}{K!}x^K \\ &= \sum_{k=0}^{K} \frac{1}{k!} x^k. \end{aligned} \tag{4.17}$$

We now aim at finding three values of $K \in \mathbb{N}$, namely:

(a) $K \in \mathbb{N}$ such that

$$\left| \mathrm{e}^x - \sum_{k=0}^{K} \frac{x^k}{k!} \right| \leq 0.015 \quad \text{for all } x \in [-1, 2]. \tag{4.18}$$

(b) $K \in \mathbb{N}$ such that

$$\left| \mathrm{e}^x - \sum_{k=0}^{K} \frac{x^k}{k!} \right| \leq 0.0015 \quad \text{for all } x \in [-1, 2]. \tag{4.19}$$

(c) $K \in \mathbb{N}$ such that

$$\left| \mathrm{e}^x - \sum_{k=0}^{K} \frac{x^k}{k!} \right| \leq 0.015 \quad \text{for all } x \in [-5, 5]. \tag{4.20}$$

In order to do so, we first observe that the estimate (4.14) involves the constant C, which we can choose as the maximum of all derivatives of f on the considered interval. For the function f we have that $f^{(k)}(x) = \mathrm{e}^x$, i.e., all derivatives equal the function itself.

For Item (a) we can thus take $C = \mathrm{e}^2$. Now (4.14) shows that (4.18) is obtained if we choose K such that

$$\frac{\mathrm{e}^2}{(K+1)!}2^{K+1} \leq 0.015; \tag{4.21}$$

this is satisfied for all $K \geq 8$.

In Item (b), we obtain an appropriate value for $K \in \mathbb{N}$ by replacing the number 0.015 on the right-hand side of (4.21) by 0.0015; the obtained inequality is satisfied for all $K \geq 10$.

Note than in Item (c), we enlarge the interval on which we want to approximate f. In order to find an appropriate K-value in Item (a), we need to replace the previous value for C with $C = \mathrm{e}^5$; we obtain the inequality

$$\frac{\mathrm{e}^5}{(K+1)!}5^{K+1} \leq 0.015,$$

which is satisfied for $K \geq 19$.

4.4 The tangent plane for functions of several variables

We will now take the first steps to generalize the theory for Taylor polynomials to higher dimensions. In this section we focus on first-degree Taylor polynomials; the case of second degree Taylor polynomials will be treated in Section 4.5.

Consider a function $f : U \to \mathbb{R}$, where U is an open set in $\mathbb{R}^n$, and fix a vector $\boldsymbol{x}_0 \in U$. Similarly to our generalizations of the concepts of continuity and differentiability, our task is to identify a general definition of the Taylor polynomial of first degree at the point $\boldsymbol{x}_0$; the definition must correspond to Definition 4.1.1 in the case $n = 1$.

Looking now at (4.2), the difference $x - x_0$ will then be replaced by the difference $\boldsymbol{x} - \boldsymbol{x}_0$, where $\boldsymbol{x} \in \mathbb{R}^n$. For a function of n variables it is natural to replace the derivative $f'(x_0)$ in (4.2) by the gradient vector at the point

$\boldsymbol{x}_0$. The gradient vector was defined in Equation (3.23) on page 96 as

$$\nabla f(\boldsymbol{x}_0) = \left(\frac{\partial f}{\partial x_1}(\boldsymbol{x}_0), \frac{\partial f}{\partial x_2}(\boldsymbol{x}_0), \ldots, \frac{\partial f}{\partial x_n}(\boldsymbol{x}_0) \right). \tag{4.22}$$

With these observations, the expression in (4.2) indeed has a generalization to $\mathbb{R}^n$ if we consider the product between $f'(x_0)$ and $(x - x_0)$ as the inner product in $\mathbb{R}$. We thus arrive at the following definition:

Definition 4.4.1 Taylor polynomial of first degree

Let U denote an open set in $\mathbb{R}^n$ and consider a function $f : U \to \mathbb{R}$ for which all partial derivatives exist on the set U. Fix a point $\boldsymbol{x}_0 \in U$. Then the *Taylor polynomial of first degree* at the point $\boldsymbol{x}_0$ is defined by

$$P_{1,f,\boldsymbol{x}_0}(\boldsymbol{x}) = P_1(\boldsymbol{x}) = f(\boldsymbol{x}_0) + \langle (\boldsymbol{x} - \boldsymbol{x}_0), \nabla f(\boldsymbol{x}_0) \rangle, \quad \boldsymbol{x} \in \mathbb{R}^n. \tag{4.23}$$

The graph of the function P_1, i.e., the set of points $(\boldsymbol{x}, P_1(\boldsymbol{x}))$, $\boldsymbol{x} \in \mathbb{R}^n$, is called the *tangent plane* of the function f at the point $\boldsymbol{x}_0$.

We will often consider functions of two variables. In such cases we can apply the following explicit expression for the first-degree Taylor polynomial:

Lemma 4.4.1

Let U denote an open set in $\mathbb{R}^2$ and consider a differentiable function $f : U \to \mathbb{R}$. Fix a point $(x_0, y_0) \in U$. Then the Taylor polynomial of first degree at the point (x_0, y_0) is

$$P_1(x, y) = f(x_0, y_0) + \frac{\partial f}{\partial x}(x_0, y_0)(x - x_0) + \frac{\partial f}{\partial y}(x_0, y_0)(y - y_0), \tag{4.24}$$

where $(x, y) \in \mathbb{R}^2$.

Proof. Write $\boldsymbol{x} = (x, y)$ and $\boldsymbol{x}_0 = (x_0, y_0)$. Then (4.23) shows that

$$\begin{aligned} P_1(x, y) &= f(\boldsymbol{x}_0) + \langle (\boldsymbol{x} - \boldsymbol{x}_0), \nabla f(\boldsymbol{x}_0) \rangle \\ &= f(x_0, y_0) + \begin{bmatrix} \frac{\partial f}{\partial x}(x_0, y_0) & \frac{\partial f}{\partial y}(x_0, y_0) \end{bmatrix} \begin{bmatrix} x - x_0 \\ y - y_0 \end{bmatrix} \\ &= f(x_0, y_0) + \frac{\partial f}{\partial x}(x_0, y_0)(x - x_0) + \frac{\partial f}{\partial y}(x_0, y_0)(y - y_0), \end{aligned}$$

as claimed.

■

Remark 4.4.1

Lemma 4.4.1 shows that the first-degree Taylor polynomial at a point (x_0, y_0) for a function of two variables satisfies that

$$P_1(x_0, y_0) = f(x_0, y_0),\ \frac{\partial P_1}{\partial x}(x_0, y_0) = \frac{\partial f}{\partial x}(x_0, y_0),\ \frac{\partial P_1}{\partial y}(x_0, y_0) = \frac{\partial f}{\partial y}(x_0, y_0).$$

That is, the function f and its first-degree Taylor polynomial P_1 coincide at the point (x_0, y_0), and the gradient of f and the gradient of P_1 also coincide at (x_0, y_o). This property corresponds to what we have for the first-degree Taylor polynomial for a function of one variable; see Remark 4.1.1. The similar result (with a similar calculation) holds for a function of n variables.

Example 4.4.1

Let us return to the function

$$f : \mathbb{R}^2 \to \mathbb{R},\ \ f(x, y) = \sin(x^2 + y),$$

which we considered in Example 3.3.2. The calculations in Example 3.3.2 show that

$$\nabla f(x, y) = (2x\cos(x^2 + y), \cos(x^2 + y)).$$

Thus, at the point $(x_0, y_0) = (\sqrt{\pi}, \pi)$ we have

$$\nabla f(\sqrt{\pi}, \pi) = (2\sqrt{\pi}\cos(\pi + \pi), \cos(\pi + \pi)) = (2\sqrt{\pi}, 1).$$

Therefore the Taylor polynomial of first degree at the point $(x_0, y_0) = (\sqrt{\pi}, \pi)$ is

$$\begin{aligned} P_1(x, y) &= f(\sqrt{\pi}, \pi) + \left\langle \begin{bmatrix} x - \sqrt{\pi} \\ y - \pi \end{bmatrix}, \begin{bmatrix} 2\sqrt{\pi} \\ 1 \end{bmatrix} \right\rangle \\ &= \sin(\pi + \pi) + (x - \sqrt{\pi})2\sqrt{\pi} + (y - \pi) \cdot 1 \\ &= 2\sqrt{\pi}x + y - 3\pi,\ \ (x, y) \in \mathbb{R}^2. \end{aligned}$$

4.5 Taylor polynomials for functions of several variables

As for functions of one variable, we can define higher-degree Taylor polynomials by adding higher-degree terms to the formula for the Taylor polynomial of first degree. We will here only consider the Taylor polynomial

of second degree as the notation for higher degree polynomials becomes rather complicated as one usually formulate such higher degree Taylor polynomials using tensors or multi-indices. We refer the interested reader to the additional notes. To define second-degree Taylor polynomials of functions of several variables, we need the Hessian matrix $\boldsymbol{H}_f(\boldsymbol{x}_0)$ introduced in Equation (3.27) on page 102.

Definition 4.5.1 Taylor polynomial of second degree

Let U denote an open set in $\mathbb{R}^n$ and consider a function $f : U \to \mathbb{R}$ for which all partial derivatives of order one and two exist on the set U. Fix a point $\boldsymbol{x}_0 \in U$. Then the *Taylor polynomial of second degree* $P_{2,f,\boldsymbol{x}_0} = P_2$ at the point $\boldsymbol{x}_0$ is defined by

$$P_2(\boldsymbol{x}) = f(\boldsymbol{x}_0) + \langle (\boldsymbol{x} - \boldsymbol{x}_0), \nabla f(\boldsymbol{x}_0) \rangle + \frac{1}{2} \langle (\boldsymbol{x} - \boldsymbol{x}_0), \boldsymbol{H}_f(\boldsymbol{x}_0)(\boldsymbol{x} - \boldsymbol{x}_0) \rangle \quad (4.25)$$

for $\boldsymbol{x} \in \mathbb{R}^n$.

If, in addition to the assumptions of Definition 4.5.1, all second-order partial derivatives are continuous in U, then the Hessian matrix is a real *symmetric* $n \times n$ matrix by Theorem 3.6.5 on page 108.

For functions of just two variables we can apply the following explicit expression for the second-degree Taylor polynomial:

Lemma 4.5.1

Let U denote an open set in $\mathbb{R}^2$ and consider a function $f : U \to \mathbb{R}$ for which all partial derivatives of second order exist *and are continuous* on the set U. Fix a point $(x_0, y_0) \in U$. Then the Taylor polynomial of second order at the point (x_0, y_0) is

$$\begin{aligned} P_2(x, y) = {} & f(x_0, y_0) + \frac{\partial f}{\partial x}(x_0, y_0)(x - x_0) + \frac{\partial f}{\partial y}(x_0, y_0)(y - y_0) \quad (4.26) \\ & + \frac{1}{2}\frac{\partial^2 f}{\partial x^2}(x_0, y_0)(x - x_0)^2 + \frac{1}{2}\frac{\partial^2 f}{\partial y^2}(x_0, y_0)(y - y_0)^2 \\ & + \frac{\partial^2 f}{\partial x \partial y}(x_0, y_0)(x - x_0)(y - y_0), \ (x, y) \in \mathbb{R}^2. \quad (4.27) \end{aligned}$$

Proof. The terms in the first line (4.26) corresponds exactly to the first-degree Taylor polynomial $f(\boldsymbol{x}_0) + \langle (\boldsymbol{x} - \boldsymbol{x}_0), \nabla f(\boldsymbol{x}_0) \rangle$, so we now focus on the new term

$$\frac{1}{2} \langle (\boldsymbol{x} - \boldsymbol{x}_0), \boldsymbol{H}_f(\boldsymbol{x}_0)(\boldsymbol{x} - \boldsymbol{x}_0) \rangle = \frac{1}{2} \langle \boldsymbol{H}_f(\boldsymbol{x}_0)(\boldsymbol{x} - \boldsymbol{x}_0), (\boldsymbol{x} - \boldsymbol{x}_0) \rangle,$$

where we have used that $\boldsymbol{H}_f(\boldsymbol{x}_0)$ is a symmetric matrix by Theorem 3.6.5. Using the expression in (3.28) for the Hessian matrix at the point (x_0, y_0), we see that in the two-dimensional case,

$$\begin{aligned}\boldsymbol{H}_f(\boldsymbol{x}_0)(\boldsymbol{x}-\boldsymbol{x}_0) &= \boldsymbol{H}_f(x_0,y_0)\begin{bmatrix}x-x_0\\y-y_0\end{bmatrix}\\ &= \begin{bmatrix}\frac{\partial^2 f}{\partial x^2}(x_0,y_0) & \frac{\partial^2 f}{\partial x\partial y}(x_0,y_0)\\ \frac{\partial^2 f}{\partial y\partial x}(x_0,y_0) & \frac{\partial^2 f}{\partial y^2}(x_0,y_0)\end{bmatrix}\begin{bmatrix}x-x_0\\y-y_0\end{bmatrix}\\ &= \begin{bmatrix}\frac{\partial^2 f}{\partial x^2}(x_0,y_0)(x-x_0)+\frac{\partial^2 f}{\partial x\partial y}(x_0,y_0)(y-y_0)\\ \frac{\partial^2 f}{\partial y\partial x}(x_0,y_0)(x-x_0)+\frac{\partial^2 f}{\partial y^2}(x_0,y_0)(y-y_0)\end{bmatrix}.\end{aligned}$$

Therefore

$$\begin{aligned}&\frac{1}{2}\langle \boldsymbol{H}_f(\boldsymbol{x}_0)(\boldsymbol{x}-\boldsymbol{x}_0),(\boldsymbol{x}-\boldsymbol{x}_0)\rangle =\\ &\quad\frac{1}{2}\begin{bmatrix}x-x_0 & y-y_0\end{bmatrix}\begin{bmatrix}\frac{\partial^2 f}{\partial x_1}(x_0,y_0)(x-x_0)+\frac{\partial^2 f}{\partial x\partial y}(x_0,y_0)(y-y_0)\\ \frac{\partial^2 f}{\partial y\partial x}(x_0,y_0)(x-x_0)+\frac{\partial^2 f}{\partial y^2}(x_0,y_0)(y-y_0)\end{bmatrix}\\ &\qquad=\frac{1}{2}\frac{\partial^2 f}{\partial x^2}(x_0,y_0)(x-x_0)^2+\frac{1}{2}\frac{\partial^2 f}{\partial y^2}(x_0,y_0)(y-y_0)^2\\ &\qquad\qquad+\frac{\partial^2 f}{\partial x\partial y}(x_0,y_0)(x-x_0)(y-y_0)\end{aligned}$$

as claimed.

■

The second-degree Taylor polynomial $P_{2,f,\boldsymbol{x}_0}(\boldsymbol{x})$ defined in Equation (4.25) on the preceding page is actually a class of functions we understand very well. In fact, $P_{2,f,\boldsymbol{x}_0}$ is a *quadratic form*. We ask the reader to prove this fact in the next exercise.

Exercise 4.5.1

Show that the second-degree Taylor polynomial $P_{2,f,\boldsymbol{x}_0}(\boldsymbol{x})$ defined in Equation (4.25) on the previous page is a quadratic form as defined in Definition 1.2.1 on page 16. You should express the matrix A, the column vector $\boldsymbol{b}$ and the constant c in terms of f and its derivatives and $\boldsymbol{x}_0$. *Hint:* Start by expanding all inner products in (4.25) using linearity of the inner product. Collect *all* constant terms (those not depending on $\boldsymbol{x}$) and set them equal to c. Proceed to the terms linear in $\boldsymbol{x}$.

Example 4.5.2

For the function

$$f : \mathbb{R}^2 \to \mathbb{R}, \;\; f(x, y) = \sin(x^2 + y),$$

we saw in Example 4.4.1 that, at the point $(x_0, y_0) = (\sqrt{\pi}, \pi)$, the Taylor polynomial of first degree is

$$P_1(x, y) = 2\sqrt{\pi}x + y - 3\pi, \;\; (x, y) \in \mathbb{R}^2.$$

The second-order partial derivatives were calculated in Example 3.5.2, where we saw that

$$\begin{aligned}
\frac{\partial^2 f}{\partial x}(x, y) &= 2\cos(x^2 + y) - 4x^2 \sin(x^2 + y) \\
\frac{\partial^2 f}{\partial x \partial y}(x, y) &= -2x \sin(x^2 + y) \\
\frac{\partial^2 f}{\partial y \partial x}(x, y) &= -2x \sin(x^2 + y) \\
\frac{\partial^2 f}{\partial y^2}(x, y) &= -\sin(x^2 + y).
\end{aligned}$$

Thus, at the point $(x_0, y_0) = (\sqrt{\pi}, \pi)$,

$$\begin{aligned}
\frac{\partial^2 f}{\partial x}(\sqrt{\pi}, \pi) &= 2\cos(\pi + \pi) - 4\pi^2 \sin(\pi + \pi) = 2, \\
\frac{\partial^2 f}{\partial x \partial y}(\sqrt{\pi}, \pi) &= -2\sqrt{\pi} \sin(\pi + \pi) = 0, \\
\frac{\partial^2 f}{\partial y \partial x}(\sqrt{\pi}, \pi) &= -2\sqrt{\pi} \sin(\pi + \pi) = 0,
\end{aligned}$$

and

$$\frac{\partial^2 f}{\partial y^2}(\sqrt{\pi}, \pi) = -\sin(\pi + \pi) = 0.$$

By Lemma 4.5.1, the second-degree Taylor polynomial at the point $(x_0, y_0) = (\sqrt{\pi}, \pi)$ is

$$\begin{aligned}
P_2(x, y) &= 2\sqrt{\pi}x + y - 3\pi + (x - \sqrt{\pi})^2 \\
&= x^2 + y - 2\pi.
\end{aligned}$$

Example 4.5.3

For the function

$$f : \mathbb{R}^2 \to \mathbb{R}, \ f(x, y) = \sin(x^2 + y^2),$$

direct calculation (try to do it) shows that the first-degree Taylor polynomial at the point $(x_0, y_0) = (0, 0)$ is

$$P_1(x, y) = 0,$$

and that the second-degree Taylor polynomial is

$$P_2(x, y) = x^2 + y^2.$$

Figure 4.6 shows the function f and the two Taylor polynomials.

4.6 Taylor's formula for functions of several variables

The Taylor polynomial P_2 of second degree is typically a better approximation of a given function $f : \mathbb{R}^n \to \mathbb{R}$ than the first-degree Taylor approximation, but still it is just an approximation. That is, we can write

$$f(\boldsymbol{x}) = P_2(\boldsymbol{x}) + R_2(\boldsymbol{x}),$$

where the function $R_2 : \mathbb{R}^n \to \mathbb{R}$ is the so called *remainder term*. Alternatively, we may write

$$R_2(\boldsymbol{x}) = f(\boldsymbol{x}) - P_2(\boldsymbol{x}).$$

In general we do not know exactly how the function R_2 look like, but as for the case of functions of one variable the important *Taylor's formula* quantifies its behavior:

Theorem 4.6.1 Taylor's formula

In the setup of Definition 4.5.1, with the additional assumption that f is a C^2-function, the remainder-term R_2 has the form

$$R_2(\boldsymbol{x}) = f(\boldsymbol{x}) - P_2(\boldsymbol{x}) = \varepsilon(\boldsymbol{x} - \boldsymbol{x}_0)\|\boldsymbol{x} - \boldsymbol{x}_0\|^2 \tag{4.28}$$

for some function $\varepsilon : \mathbb{R}^n \to \mathbb{R}$ for which $\varepsilon(\boldsymbol{x} - \boldsymbol{x}_0) \to 0$ whenever $\boldsymbol{x} \to \boldsymbol{x}_0$.

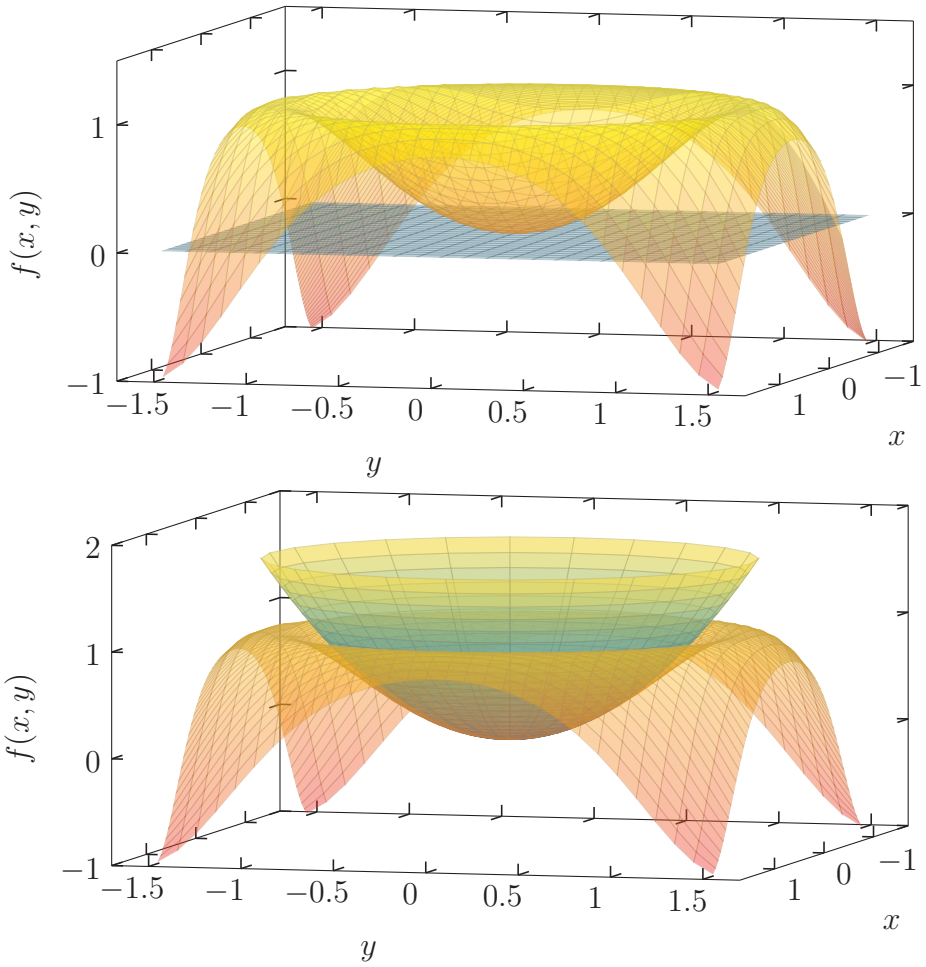


Figure 4.6: The function $f(x, y) = \sin(x^2 + y^2)$ and its Taylor polynomials of first degree and second degree at the point $(0, 0)$.

Theorem 4.6.1 can be proved by a number of tricks, reducing the problem to the one-dimensional case considered in Lemma 4.3.1; this leads to a result of a similar type (but of higher complexity) as in Lemma 4.3.1. Furthermore, there exist definitions of Taylor polynomials of arbitrary high degree, and corresponding error estimates. It is outside the scope of the current presentation to go into an analysis of this.

4.7 Taylor polynomials for vector functions of several variables

We will now turn to Taylor polynomials of a *vector* function $\boldsymbol{f} = (f_1, \ldots, f_k) : U \to \mathbb{R}^k$, where U, as usual, is an open set in $\mathbb{R}^n$. Luckily, the situation is not much more complicated than for scalar functions of several variables. Recall that each coordinate function $f_i : U \to \mathbb{R}$ is a scalar function of n variables. So to get the Kth degree Taylor polynomial for $\boldsymbol{f}$ at $\boldsymbol{x}_0$, we "just" need to find the Kth degree Taylor polynomial of each coordinate function f_i at the same point $\boldsymbol{x}_0$. This leads to the following definition.

Definition 4.7.1 Taylor polynomial for vector functions

Let U denote an open set in $\mathbb{R}^n$ and consider a function $\boldsymbol{f} = (f_1, \ldots, f_k) : U \to \mathbb{R}^k$. Suppose all partial derivatives up to degree K of each coordinate function f_i, $i = 1, \ldots, k$, exist on the set U. Fix a point $\boldsymbol{x}_0 \in U$. Then the *Taylor polynomial of K degree* $\boldsymbol{P}_{K,\boldsymbol{f},\boldsymbol{x}_0} = \boldsymbol{P}_K$ at the point $\boldsymbol{x}_0$ is defined as the vector function $\boldsymbol{P}_K : \mathbb{R}^n \to \mathbb{R}^k$ whose ith coordinate function is $P_{K,f_i,\boldsymbol{x}_0}$.

In Definition 4.7.1, one usually assumes that $\boldsymbol{f}$ is a C^K vector function, see the discussion just below Definition 3.8.2 on page 114, as it guarantees that all the partial derivatives exist and are continuous. Note that Definition 4.7.1 is not very explicit, and we have only discussed how to write down $P_{K,f_i,\boldsymbol{x}_0}$ for $K = 1$ and $K = 2$ in case $n > 1$. For $K = 1$ there is a simple formula for $\boldsymbol{P}_{1,\boldsymbol{f},\boldsymbol{x}_0}$ in terms of the Jacobian of $\boldsymbol{f}$. Suppose that $\boldsymbol{f} : U \to \mathbb{R}^k$ is a differentiable function at $\boldsymbol{x}_0 \in U$. Then the first-degree Taylor polynomial at the point $\boldsymbol{x}_0$ is given by:

$$\boldsymbol{P}_{1,\boldsymbol{f},\boldsymbol{x}_0}(\boldsymbol{x}) = \boldsymbol{f}(\boldsymbol{x}_0) + \boldsymbol{J}_{\boldsymbol{f}}(\boldsymbol{x}_0)(\boldsymbol{x} - \boldsymbol{x}_0). \tag{4.29}$$

We actually derived this first-degree Taylor polynomial in Equation (4.1) on page 117 from Definition 3.8.1 on page 112. Moreover, we know the remainder term satisfies:

$$\boldsymbol{R}_{1,\boldsymbol{f},\boldsymbol{x}_0}(\boldsymbol{x}) = \boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{P}_{1,\boldsymbol{f},\boldsymbol{x}_0}(\boldsymbol{x}) = \boldsymbol{\varepsilon}(\boldsymbol{x} - \boldsymbol{x}_0)\|\boldsymbol{x} - \boldsymbol{x}_0\|,$$

where $\boldsymbol{\varepsilon}(\boldsymbol{x} - \boldsymbol{x}_0) \to \boldsymbol{0}$ as $\boldsymbol{x} \to \boldsymbol{x}_0$.

Exercise 4.7.1

Suppose that $\boldsymbol{f} = (f_1, \ldots, f_k) : U \to \mathbb{R}^k$ is a differentiable function at $\boldsymbol{x}_0 \in U$, where $U \subseteq \mathbb{R}^n$ is an open set. Derive the formula in Equation (4.29) by applying the formula in Equation (4.23) on page 129 for each coordinate

function f_i, $i = 1, \ldots, k$. *Hint:* Remember that the rows of the Jacobian matrix contain the gradient vector of each coordinate function.

It is rather technical to write down $\boldsymbol{P}_{K,\boldsymbol{f},\boldsymbol{x}_0}(\boldsymbol{x})$ explicitly for $K > 1$, and it is beyond the scope of this text.

CHAPTER 5

Local and Global Extrema of Functions

Many important problems in engineering deals with *optimization*. For example, how can we – under given physical conditions – minimize the power consummation of an electric device? Or, how can we maximize the output from a wind mill or a solar panel? Typically, the value of the "item" to be optimized can be expressed as a function of a number of variables representing physical quantities. In such cases we are left with a purely mathematical question of how to maximize/minimize a function in a certain domain that is determined by the physical limitations for the involved variables. In this chapter we will discuss methods for optimization for *scalar* functions of several variables. We begin by reminding the reader of the one-dimensional case in Section 5.1. The general case of a scalar function of several variables is then considered in Section 5.2.

We will not consider vector functions in this chapter. In fact, one cannot optimize vector functions since such functions takes values in $\mathbb{R}^k$. Recall that it does not make sense to ask if a vector $\boldsymbol{y}_1 \in \mathbb{R}^k$ is smaller than or larger than $\boldsymbol{y}_2 \in \mathbb{R}^k$. Phrased mathematically, there is *no ordering* on $\mathbb{R}^k$ so we *cannot* ask for the "smallest" or "largest" value of a function $\boldsymbol{f} : \mathrm{dom}(\boldsymbol{f}) \to \mathbb{R}^k$. We already see this issue for complex-valued scalar function, again, because there is no ordering on $\mathbb{C}$. However, the solution is evident: we optimize the norm of the function values instead. Hence, if $\boldsymbol{f} : \mathrm{dom}(\boldsymbol{f}) \to \mathbb{F}^k$, we consider the function $\boldsymbol{x} \mapsto \|\boldsymbol{f}(\boldsymbol{x})\|, \mathrm{dom}(\boldsymbol{f}) \to [0, \infty[$, which is a real-valued scalar function. The norm $\|\cdot\|$ is not necessarily the standard norm $\mathbb{F}^k$, and the choice often depends on the application in question.

For (real-valued) scalar functions, we will consider two types of optimization, a *global optimization* and a *local optimization*. Global

optimization aims at finding the maximum/minimum value of a given function over the entire domain. On the other hand, *local optimization* deals with an analysis of a given function in a small neighborhood of a given point. It provides methods to decide whether the given point at least yields the maximum/minimum value within a small neighborhood of the point. The main tool here is to approximate the function locally by a *quadratic form*, namely, a second-degree Taylor polynomial. Both local and global optimization are highly relevant in practice.

5.1 The range of functions of one variable

In this section we will remind the reader about standard results concerning the *image* or *range* of a function f that is defined on a closed and bounded interval $[a, b]$ in $\mathbb{R}$. The image set $\mathrm{im}(f)$ of $f : [a, b] \to \mathbb{R}$ is also written as $f([a, b])$, and it is the set given by $\{f(x) \,\big|\, x \in [a, b]\}$. One of the fundamental theorems of analysis states that, if the function f defined on a closed and bounded interval is continuous, then the image set is itself a closed and bounded interval.

Theorem 5.1.1

Consider a continuous function $f : [a, b] \to \mathbb{R}$ defined on a bounded and closed interval $[a, b]$. Then the image of f is bounded and has a minimal value $m \in \mathbb{R}$ and maximal value $M \in \mathbb{R}$, that is,

$$\exists m, M \in \mathrm{im}(f)\, \forall y \in \mathrm{im}(f) : m \leq y \leq M.$$

Moreover, the image set has the form

$$f([a, b]) = [m, M]. \tag{5.1}$$

The proof of Theorem 5.1.1 is beyond the scope of this text, and we refer the reader to Theorem 2.9 in the additional notes for a proof.

The value m in Theorem 5.1.1 is called the *minimum* of the function f, and the value M is called the *maximum*. We use *extremum* to denote either the minimum or the maximum. Theorem 5.1.1 tells us that there indeed exist $x_1 \in [a, b]$ such that $f(x_1) = m$, and we say that the function f *attains its minimum* in x_1. Similarly, there exists $x_2 \in [a, b]$ such that $f(x_2) = M$, that is, f *attains its maximum* in x_2. Furthermore, Theorem 5.1.1 says that any y in the interval between m and M there exists an $x \in [a, b]$ so that $f(x) = y$.

For the validity of Theorem 5.1.1 it is essential that all the assumptions are satisfied. The following examples illustrate this.

Example 5.1.1 (a) Consider the continuous function

$$f :]0,1[\to \mathbb{R},\ f(x) = \ln(x).$$

Then $\mathrm{im}(f) = f(]0,1[) =]-\infty, 0[$. That is, the image set is not bounded, and the function f neither has a minimum nor a maximum. Note that the domain $]0,1[$ is non-closed, i.e., the conclusion does not contradict Theorem 5.1.1. Note that $\ln(1) = 0$ is not a maximum of $\mathrm{im}(f)$. It is, however, the supremum of $\mathrm{im}(f)$ as the least upper bound of $]-\infty, 0[$. The function f does not attain a maximum since there is no $x \in \mathrm{dom}(f)$ such that $f(x) = 1$.

(b) Consider the continuous function

$$f : [0,\infty[\to \mathbb{R},\ f(x) = 3x^2.$$

Then $\mathrm{im}(f) = f([0,\infty[) = [0,\infty[$. Hence, the image set is not a bounded interval, and the function does not have a maximum value. Note that the domain $[0,\infty[$ is not a bounded set so the conclusion does not contradict Theorem 5.1.1.

(c) Consider the function

$$f : [-2,-1] \cup [1,2] \to \mathbb{R},\ f(x) = \begin{cases} -1 & \text{whenever } x \in [-2,-1], \\ 1 & \text{whenever } x \in [1,2]. \end{cases}$$

Then the image set only consists of the two points $\{-1, 1\}$, and is thus not an interval. This does not contradict Theorem 5.1.1; the function f is continuous (why?), and the domain $[-2,-1] \cup [1,2]$ is closed and bounded, but it is not an interval.

(d) Consider the function

$$f : [-1,1] \to \mathbb{R},\ f(x) = \begin{cases} \frac{1}{x} & \text{whenever } x \neq 0, \\ 0 & \text{whenever } x = 0. \end{cases}$$

Then the image set is $f([-1,1]) =]-\infty,-1] \cup \{0\} \cup [1,\infty[$. That is, even though the function f is defined on the closed and bounded interval $[-1,1]$, the image set is not an interval; furthermore there is neither a minimum or a maximum. This does not contradict Theorem 5.1.1 since the function f is not continuous.

Note that Theorem 5.1.1 is a theoretical result: it tells us that a function f that satisfies the stated conditions has a minimum and a maximum, but not how to find these values. For this purpose we need the following result.

Theorem 5.1.2

Consider a continuous function $f : [a, b] \to \mathbb{R}$. Then, if f attains its minimum or maximum at the point x_0, one of the following possibilities occur:

(i) $x_0 = a$;

(ii) $x_0 = b$;

(iii) x_0 is a point where f is not differentiable;

(iv) x_0 is a point where f is differentiable, and $f'(x_0) = 0$.

Proof. The result follows from a more general result proved in Theorem 5.2.2. ■

In other words: in order to find the minimum and the maximum for the function f, it is enough to search among the *end points* a, b of the domain *and* the points x_0 where f is either not differentiable, or $f'(x_0) = 0$.

Example 5.1.2

Consider the function

$$f : [0, 2] \to \mathbb{R}, \ f(x) = x^3 - x.$$

Then f satisfies the conditions in Theorem 5.1.1. In order to find the minimum and the maximum, we now apply Theorem 5.1.2. First, the function f is differentiable for all $x \in]0, 2[$, and

$$f'(x) = 3x^2 - 1.$$

Note that the solutions to the equation $3x^2 - 1 = 0$ are $x = \pm 1/\sqrt{3}$. However, $x = -1/\sqrt{3}$ does not belong to the considered domain $[0, 2]$ and will thus be discarded. According to Theorem 5.1.2, the minimum and maximum of the function f are among the values

$$f(0) = 0, \qquad f(2) = 6, \qquad f(\frac{1}{3^{1/2}}) = \frac{1}{3^{3/2}} - \frac{1}{3^{1/2}} = -\frac{2}{3^{3/2}}.$$

We conclude that the minimum is $f(\frac{1}{3^{1/2}}) = -\frac{2}{3^{3/2}}$, the maximum is $f(2) = 6$, and the image set is $[-\frac{2}{3^{3/2}}, 6]$.

Note that Theorem 5.1.2 just tells us that in order to find the minimum and the maximum, we should examine the points $x \in [a, b]$ where $f'(x) = 0$; it can very well be that $f'(x) = 0$ and that the function f neither assumes a minimum nor a maximum at the point x. The following example illustrates this.

Example 5.1.3

Consider the function

$$f : [-2, 2] \to \mathbb{R}, \ f(x) = x^3.$$

The function f is increasing, so the image set is $f([-2, 2]) = [f(-2), f(2)] = [-8, 8]$. We see that $f'(x) = 3x^2$ and that $f'(0) = 0$; however, the function f neither has a minimum nor a maximum at the point $x = 0$. A point were the derivative is zero, but where f neither has a minimum nor a maximum, is called a *saddle point*.

Recall that the *second derivative* f'' can help us to decide whether a point x_0 with $f'(x_0) = 0$ qualifies as a candidate to give a minimum value or a maximum value.

Theorem 5.1.3 Second derivative test

Consider a function $f : [a, b] \to \mathbb{R}$ which is two times differentiable on the open interval $]a, b[$. Assume that $f'(x_0) = 0$ for some $x_0 \in]a, b[$. Then the following assertions hold:

(i) If $f''(x_0) < 0$, there exists an open interval $I \subset]a, b[$ containing x_0 such that $f(x) < f(x_0)$ for all $x \in I \setminus \{x_0\}$.

(ii) If $f''(x_0) > 0$, there exists an open interval $I \subset]a, b[$ containing x_0 such that $f(x) > f(x_0)$ for all $x \in I \setminus \{x_0\}$.

Proof. The result follows from a more general result proved in Theorem 5.2.4. ■

Note that if the assumption in Item (i) of Theorem 5.1.3 is satisfied, the conclusion only says that the function value $f(x_0)$ is larger than the function values in a small interval containing x_0. That is, we can not conclude yet that $f(x_0)$ is a maximal value measured over the entire interval $[a, b]$. We say that $f(x_0)$ is a (strict) *local maximum value*. Similarly, in Item (ii) of Theorem 5.1.3, the conclusion is phrased by saying that $f(x_0)$ is a (strict) *local minimum value*.

In case $f'(x_0) = 0$ and also $f''(x_0) = 0$, Theorem 5.1.3 does not give us any information. In this case it is necessary with a closer examination to find out whether x_0 can lead to a minimum value or a maximum value. The following example illustrates this.

Example 5.1.4

Consider the following three functions, all of them on the interval $[-1, 1]$

$$f(x) = -x^4, \qquad g(x) = x^4, \qquad h(x) = x^3.$$

Then

$$f'(x) = -4x^3, \qquad g'(x) = 4x^3, \qquad h'(x) = 3x^2$$

and

$$f''(x) = -12x^2, \qquad g''(x) = 12x^2, \qquad h''(x) = 6x,$$

so

$$f'(0) = f''(0) = g'(0) = g''(0) = h'(0) = h''(0) = 0.$$

By inspection, we see that $x = 0$ leads to a maximum for the function f and a minimum for the function g. For the function h, the point neither yields a minimum value nor a maximum value.

5.2 The range of functions of several variables

We will now consider the image set for functions of n variables. Our first goal is to generalizes Theorem 5.1.1 to functions $f : B \to \mathbb{R}$, where B is a subset of $\mathbb{R}^n$. In order to do so, we need to find the correct higher-dimensional variants of the conditions in Theorem 5.1.1, such that we avoid the issues encountered in Example 5.1.1. Let us look at the conditions in Theorem 5.1.1, one by one:

(i) Example 5.1.1 (i) shows that in order for the result in Theorem 5.1.1 to hold, it is essential that the function f is defined on an interval $[a, b]$ and *not* on $]a, b[$; that is, the domain must contain the end points of the interval, or, in other words, the domain must be closed. For a function of n variables, $f : B \to \mathbb{R}$, we will therefore require that the domain B is *closed*.

(ii) Example 5.1.1 (ii) shows that in order for the result in Theorem 5.1.1 to hold, it is essential that the function f is defined on a bounded set.

For a function of n variables, $f : B \to \mathbb{R}$, we will therefore require that B is a *bounded* set; see Definition 2.2.6.

(iii) Example 5.1.1 (iii) shows that in order for the result in Theorem 5.1.1 to hold, it is essential that the function f is defined on an interval; that is, the result might not hold for functions that are defined on a union of two disjoint intervals. The next definition translates this into a condition on sets in $\mathbb{R}^n$.

Definition 5.2.1 Connected set

A set $B \subseteq \mathbb{R}^n$ is said to be connected if, for each choice of points $\boldsymbol{x}_1, \boldsymbol{x}_2 \in B$, there exists a continuous function $\boldsymbol{r} : [0, 1] \to B$ such that $\boldsymbol{r}(0) = \boldsymbol{x}_1$ and $\boldsymbol{r}(1) = \boldsymbol{x}_2$.

Since the function $\boldsymbol{r}$ in Definition 5.2.1 is a curve, one sometimes speak of *curve connected sets*. The key point in Definition 5.2.1 is that the curve r runs *entirely inside* the set B. In the one-dimensional setting, this condition excludes the possibility that the function is defined on the union of two disjoint intervals. It turns out that we now have identified the precise conditions that are necessary in order to generalize Theorem 5.1.1:

Theorem 5.2.1

Let $B \subset \mathbb{R}^n$ be a closed and bounded, and consider a continuous function $f : B \to \mathbb{R}$. Then f has a minimal and maximal value. If B is, in addition, also a connected set, then the image set has the form

$$\operatorname{im}(f) = f(B) = [m, M], \tag{5.2}$$

for some $m, M \in \mathbb{R}$.

Thus, exactly like Theorem 5.1.1, the result guarantees the existence of a minimum value and a maximum value. In order to find these values, we will typically apply the following result, which generalizes Theorem 5.1.2.

Theorem 5.2.2

Let $A \subseteq \mathbb{R}^n$ be a set, and consider a function $f : A \to \mathbb{R}$. Then, if f attains its minimum or maximum at the point $\boldsymbol{x}_0 \in A$, one of the following possibilities occur:

(i) $\boldsymbol{x}_0 \in A \cap \partial A$ ($\boldsymbol{x}_0$ is a boundary point in the domain);

(ii) $\boldsymbol{x}_0 \in A^\circ$ is a point where f *is not* differentiable;

(iii) $\boldsymbol{x}_0 \in A^\circ$ is a point where f *is* differentiable, and $\nabla f(\boldsymbol{x}_0) = \mathbf{0}$.

Proof. Suppose that $\boldsymbol{x}_0 \in A$ is a point where f attains an extremum. It is obvious that $\boldsymbol{x}_0$ is either a boundary point $\boldsymbol{x}_0 \in \partial A$ or an interior point $\boldsymbol{x}_0 \in A^\circ$. If $\boldsymbol{x}_0$ is an interior point, then one of two things can happen: f *is not* differentiable at $\boldsymbol{x}_0$ or f *is* differentiable at $\boldsymbol{x}_0$. Hence, the only thing we have to prove is that if f *is* differentiable at $\boldsymbol{x}_0$, then $\nabla f(\boldsymbol{x}_0) = \mathbf{0}$. We postpone the proof of this fact to Lemma 5.2.3 on page 148. ■

Thus, under the assumptions in Theorem 5.2.2, if a function f is differentiable and we want to find its minimum value and maximum value, it is enough to examine the function values $f(\boldsymbol{x})$ for points $\boldsymbol{x}$ belonging to the boundary of the domain of the function *and* for points $\boldsymbol{x}$ in the interior of the domain for which $\nabla f(\boldsymbol{x}_0) = \mathbf{0}$. This motivates the following definition.

Definition 5.2.2 Stationary point

Let $A \subseteq \mathbb{R}^n$ be a set. A point $\boldsymbol{x}_0 \in A^\circ$ is called a *stationary point* for f if f is differentiable at $\boldsymbol{x}_0$ and $\nabla f(\boldsymbol{x}_0) = \mathbf{0}$.

Calculation of the stationary points require us to solve a number of equations:

Example 5.2.1

For the function

$$f : \mathbb{R}^2 \to \mathbb{R}, \;\; f(x, y) = xy^2 + x^2 + y^3,$$

we saw in Example 3.3.4 on page 96 that

$$\nabla f(x, y) = (y^2 + 2x, 2xy + 3y^2). \tag{5.3}$$

Thus, in order to find the stationary points, we need to solve the equations

$$y^2 + 2x = 0 \quad \wedge \quad 2xy + 3y^2 = 0. \tag{5.4}$$

Note that the first equation implies that $2x = -y^2$; inserting this in the second equation shows that $-y^3 + 3y^2 = 0$, or, $y^2(-y + 3) = 0$. Thus, we have $y = 0$ or $y = 3$. For $y = 0$, the first equation in (5.4) leads to $x = 0$, and for $y = 3$ we get $x = -9/2$. Thus, the stationary points are $(0, 0)$ and $(-9/2, 3)$.

Note that the set of stationary points might not consist of isolated points:

Example 5.2.2

For the function

$$f : \mathbb{R}^3 \to \mathbb{R},\ f(x_1, x_2, x_3) = x_1x_2 + x_2^2x_3^3,$$

we saw in Example 3.3.5 that

$$\nabla f(x_1, x_2, x_3) = (x_2, x_1 + 2x_2x_3^3, 3x_2^2x_3^2).$$

Thus, (x_1, x_2, x_3) is a stationary point if and only if the equations

$$x_2 = 0, \quad x_1 + 2x_2x_3^3 = 0, \quad \textit{and } 3x_2^2x_3^2 = 0,$$

are satisfied. Using that $x_2 = 0$, the second equation shows that also $x_1 = 0$; also, the third equation is automatically satisfied, i.e., the equations do not put any restriction on x_3. Thus, the stationary points are exactly the points of the form $(x_1, x_2, x_3) = (0, 0, t)$, where $t \in \mathbb{R}$.

Let us now consider an example, where we apply Theorem 5.2.2 to find the minimum value and the maximum value for a given function.

Example 5.2.3

Let

$$B = \{(x, y) \in \mathbb{R}^2 \mid x^2 + y^2 \leq 4\},$$

and consider the function

$$f : B \to \mathbb{R},\ f(x, y) = xy + 1.$$

In order to apply Theorem 5.2.2, we first calculate the stationary points in the interior of B. We have

$$\nabla f(x, y) = (y, x).$$

Thus, the only stationary point is $(x, y) = (0, 0)$, and $f(0, 0) = 1$.

We now need to examine the function values on the boundary of B. Note that B is a ball with radius 2. Thus, the boundary δB is the circle in $\mathbb{R}^2$ with radius $r = 2$. The points in this set are parameterized by $\boldsymbol{r}(t) = (2\cos t, 2\sin t)$, where $t \in [0, 2\pi]$. On the boundary, the function values for the function f are therefore precisely the values of the function $g : [0, 2\pi] \to \mathbb{R}, g = f \circ \boldsymbol{r}$, i.e., $g(t) = f(2\cos t, 2\sin t)$. We see that

$g(t) = f(2\cos t, 2\sin t) = 4\cos t \sin t + 1 = 2\sin(2t) + 1$ for $t \in [0, 2\pi]$. The sine function take all values in the interval $[-1, 1]$, so the image set of the function g is $g([0, 2\pi]) = [-1, 3]$; comparing this information with the function value of f in the stationary point $(0, 0)$, it follows that the minimum value for the function f is -1 and the maximum value is 3, i.e., $\mathrm{im}(f) = [-1, 3]$.

Frequently, in particular, in application with functions of many variables, it is impossible to find global extrema, and we have to settle with a local analysis of a given function f, that is, we search for extrema in a neighborhood of a certain point $\boldsymbol{x}_0$. In such cases we apply the following definition, which is formalizing the discussion after Theorem 5.1.3.

Definition 5.2.3 Local minimum and maximum

Let A be a set in $\mathbb{R}^n$ and $f : A \to \mathbb{R}$ a function.

(i) If $\boldsymbol{x}_0 \in A$ and $f(\boldsymbol{x}) \geq f(\boldsymbol{x}_0)$ for all $\boldsymbol{x} \neq \boldsymbol{x}_0$ in some ball $B(\boldsymbol{x}_0, \epsilon) \cap A$, that is,

$$\exists \epsilon > 0 \, \forall \boldsymbol{x} \in A \setminus \{\boldsymbol{x}_0\} : \|\boldsymbol{x} - \boldsymbol{x}_0\| < \epsilon \Rightarrow f(\boldsymbol{x}) \geq f(\boldsymbol{x}_0),$$

we say that the function f has a *local minimum* at the point $\boldsymbol{x}_0$.

(ii) If $\boldsymbol{x}_0 \in A$ and $f(\boldsymbol{x}) \leq f(\boldsymbol{x}_0)$ for all $\boldsymbol{x} \neq \boldsymbol{x}_0$ in some ball $B(\boldsymbol{x}_0, \epsilon) \cap A$, that is,

$$\exists \epsilon > 0 \, \forall \boldsymbol{x} \in A \setminus \{\boldsymbol{x}_0\} : \|\boldsymbol{x} - \boldsymbol{x}_0\| < \epsilon \Rightarrow f(\boldsymbol{x}) \leq f(\boldsymbol{x}_0),$$

we say that the function f has a *local maximum* at the point $\boldsymbol{x}_0$.

(iii) If the above conditions hold with strict inequalities, i.e., $f(\boldsymbol{x}) > f(\boldsymbol{x}_0)$ and $f(\boldsymbol{x}) < f(\boldsymbol{x}_0)$, we say that f has a *strict local minimum* and a *strict local maximum*, respectively.

Let us illustrate the notion of *local extrema* with an example of a function of one variable.

Example 5.2.4

Consider the function

$$f : [-2,3] \to \mathbb{R}, \quad f(x) = \begin{cases} x^2 & x \in [-2,1[, \\ 1 & x \in [1,2], \\ 4(x-5/2)^2 & x \in]2,3]. \end{cases}$$

By plotting the graph of the function, it is not difficult to conclude the following. The function has local minima at $x = 0$, $x = 3/2$ and for any $x \in]1,2[$. At $x = 0$ and $x = 3/2$ the two local minima are also strict global minima with value 0, while the local minima in $]1,2[$ are neither strict nor global. The function has local maxima at $x = -2$, $x = 3$ and for any $x \in [1,2]$. The local maxima at $x = -2$ is also a strict global maxima with value $f(-2) = 4$.

In order to find the points where a given function attains local maximum and local minimum values, we use the same approach as we saw in Theorem 5.2.2, that is, we search for points x_0, where $\nabla f(x_0) = 0$. This gives us a necessary condition for a differentiable function having a local extremum in an interior point.

Lemma 5.2.3

Let $A \subseteq \mathbb{R}^n$. If $f : A \to \mathbb{R}$ has a local extremum in an interior point $\boldsymbol{x}_0 \in A^\circ$, where f is differentiable, then $\boldsymbol{x}_0$ is a stationary point, i.e., $\nabla f(\boldsymbol{x}_0) = \mathbf{0}$.

Proof. Assume f is differentiable at $\boldsymbol{x}_0$, but $\nabla f(\boldsymbol{x}_0) \neq \mathbf{0}$. Let $\boldsymbol{e} = \nabla f(\boldsymbol{x}_0)/\|\nabla f(\boldsymbol{x}_0)\|$, and let $\boldsymbol{h} = t\boldsymbol{e}$ for some $t \in \mathbb{R}$. Then, there is a function $\varepsilon : \mathbb{R}^n \to \mathbb{R}$ satisfying $\varepsilon(\boldsymbol{h}) \to 0$ for $\boldsymbol{h} \to \mathbf{0}$ such that

$$\begin{aligned} f(\boldsymbol{x}_0 + \boldsymbol{h}) &= f(\boldsymbol{x}_0) + \nabla f(\boldsymbol{x}_0)\boldsymbol{h} + \varepsilon(\boldsymbol{h})\|\boldsymbol{h}\| \\ &= f(\boldsymbol{x}_0) + t\|\nabla f(\boldsymbol{x}_0)\| + \varepsilon(t\boldsymbol{e})|t| \\ &= f(\boldsymbol{x}_0) + t(\|\nabla f(\boldsymbol{x}_0)\| \pm \varepsilon(t\boldsymbol{e})). \end{aligned}$$

For sufficiently small values of $|t|$, the term $\|\nabla f(\boldsymbol{x}_0)\| \pm \varepsilon(t\boldsymbol{e})$ is positive, hence $f(\boldsymbol{x}_0 + \boldsymbol{e}t) = f(\boldsymbol{x}_0) + ct$ for some positive constant c. This shows that $f(\boldsymbol{x}_0)$ cannot be a local extremum.

■

Note that $\nabla f(\boldsymbol{x}_0) = \mathbf{0}$ in Lemma 5.2.3 is only a necessary condition. Indeed, the function $f : \mathbb{R} \to \mathbb{R}$, $f(x) = x^3$ has a stationary point at $x_0 = 0$, but $f(x_0)$ is neither a local minimum nor local maximum. This motivates the following definition.

Definition 5.2.4 Saddle point

A stationary point of $f : A \to \mathbb{R}$, $A \subseteq \mathbb{R}^n$, is called a *saddle point* if it is neither a local minimum nor a local maximum.

The name *saddle point* comes from the fact that the graph of $f(x, y) = x^2 - y^2$ near the stationary point $(0, 0)$ looks like a horse saddle.

Let us finally notice that under certain circumstances the Hessian matrix can be used to argue whether a stationary point is a candidate to yield a minimum value of a maximum value. It is here assumed that the function f is twice continuously differentiable, i.e., that f is a C^2 function. Recall that this means that all partial derivatives appearing in the Hessian matrix exist and are continuous, or, equivalently, that ∇f is a C^1 vector function (recall Definition 3.8.2 on page 114).

Theorem 5.2.4 Second partial derivative test

Let $U \subseteq \mathbb{R}^n$ be an open set, and assume that $\boldsymbol{x}_0 \in U$ is a stationary point for a function $f : U \to \mathbb{R}$. Suppose that f is twice continuously differentiable. Denote the Hessian matrix at $\boldsymbol{x}_0$ by $\boldsymbol{H}_f(\boldsymbol{x}_0)$. Then the following hold:

(i) If $\boldsymbol{H}_f(\boldsymbol{x}_0)$ is positive definite (i.e., all eigenvalues are positive), then $\boldsymbol{x}_0$ is a strict local minimum.

(ii) If $\boldsymbol{H}_f(\boldsymbol{x}_0)$ is negative definite (i.e., all eigenvalues are negative), then $\boldsymbol{x}_0$ is a strict local maximum.

(iii) If $\boldsymbol{H}_f(\boldsymbol{x}_0)$ has both positive and negative eigenvalues, then $\boldsymbol{x}_0$ is a saddle point, that is, $f(\boldsymbol{x}_0)$ is neither a minimum nor maximum value (not even in a small neighborhood of $\boldsymbol{x}_0$).

(iv) If $\boldsymbol{H}_f(\boldsymbol{x}_0)$ is singular (i.e., has $\lambda = 0$ as an eigenvalue) *and* all non-zero eigenvalues have the same sign, then a detailed examination is necessary in order to decide whether $f(\boldsymbol{x}_0)$ is a minimum value, a maximum value, or a saddle point.

Proof. Since $\boldsymbol{x}_0 \in U$ is a stationary point, i.e., $\nabla f(\boldsymbol{x}_0) = \boldsymbol{0}$, it follows from Taylor's formula for $P_{2,f,\boldsymbol{x}_0}$ in Theorem 4.6.1 with $\boldsymbol{h} = \boldsymbol{x} - \boldsymbol{x}_0$ that

$$\begin{aligned} f(\boldsymbol{x}_0 + \boldsymbol{h}) &= f(\boldsymbol{x}_0) + \langle \boldsymbol{h}, \nabla f(\boldsymbol{x}_0) \rangle + \frac{1}{2} \langle \boldsymbol{h}, \boldsymbol{H}_f(\boldsymbol{x}_0) \boldsymbol{h} \rangle + \varepsilon(\boldsymbol{h}) \|\boldsymbol{h}\|^2 \\ &= f(\boldsymbol{x}_0) + \frac{1}{2} \langle \boldsymbol{h}, \boldsymbol{H}_f(\boldsymbol{x}_0) \boldsymbol{h} \rangle + \varepsilon(\boldsymbol{h}) \|\boldsymbol{h}\|^2, \end{aligned}$$

where $\varepsilon(\boldsymbol{h}) \to 0$ as $\boldsymbol{h} \to \boldsymbol{0}$. Since $\boldsymbol{H}_f(\boldsymbol{x}_0)$ is real and symmetric, it follows from the spectral theorem that $\boldsymbol{H}_f(\boldsymbol{x}_0) = Q \Lambda Q^T$ where Q is real orthogonal

and $\Lambda = \operatorname{diag}(\lambda_1, \dots, \lambda_m)$ contains the eigenvalues of the Hessian matrix. Let $\tilde{\boldsymbol{h}} = Q^T \boldsymbol{h}$, and let $\tilde{h}_k$ be the kth entry of $\tilde{\boldsymbol{h}}$. Then

$$\langle \boldsymbol{h}, \boldsymbol{H}_f(\boldsymbol{x}_0)\boldsymbol{h} \rangle = \langle \tilde{\boldsymbol{h}}, \Lambda \tilde{\boldsymbol{h}} \rangle = \sum_{k=1}^{n} \lambda_k |\tilde{h}_k|^2$$

To prove Item (i), we assume that $\boldsymbol{H}_f(\boldsymbol{x}_0)$ is positive definite, i.e., $\lambda_k > 0$ for all k. Let $\lambda_{\min} > 0$ denote the smallest eigenvalue. Then

$$\langle \boldsymbol{h}, \boldsymbol{H}_f(\boldsymbol{x}_0)\boldsymbol{h} \rangle = \sum_{k=1}^{n} \lambda_k |\tilde{h}_k|^2 \geq \lambda_{\min} \sum_{k=1}^{n} |\tilde{h}_k|^2 = \lambda_{\min} \|\tilde{\boldsymbol{h}}\|^2 = \lambda_{\min} \|\boldsymbol{h}\|^2$$

Hence, for any $\boldsymbol{h} \in \mathbb{R}^n$ with $\|\boldsymbol{h}\| \neq 0$ sufficiently small, we have

$$f(\boldsymbol{x}_0 + \boldsymbol{h}) \geq f(\boldsymbol{x}_0) + \tfrac{1}{2}\lambda_{\min} \|\boldsymbol{h}\|^2 + \varepsilon(\boldsymbol{h}) \|\boldsymbol{h}\|^2 = \\ f(\boldsymbol{x}_0) + \|\boldsymbol{h}\|^2 \big(\lambda_{\min}/2 + \varepsilon(\boldsymbol{h}) \big) > f(\boldsymbol{x}_0),$$

which shows that $f(\boldsymbol{x}_0)$ is a strict local minimum.

The proofs of the other assertions are similar, and we leave the details to the reader.

■

Note that the complication in Item (iv) in case of a non-invertible Hessian matrix corresponds precisely to the information we got out of Example 5.1.4 in the one-dimensional case. The complication is due to the fact the a second order Taylor approximation is not a sufficiently high order polynomial to examine the stationary point.

Finally, let us remark that the Hessian matrix may be very expensive to compute in case of functions with many variables, e.g., in large neural networks. In Example 3.7.2 on page 111 the Hessian of the overall loss function $R(\theta)$ would be an $N \times N$ matrix, where N is the number of parameters in the network, which easily can be in the millions. In such cases one often has to resort to first order methods only (e.g., using only the gradient).

CHAPTER 6

Integration

Integration is one of the fundamental concepts in mathematics. Several definitions of integration exist in the mathematical literature, each tailored for specific purposes. Also, numerous problems in physics and engineering can be formulated and solved in terms of integrals. In this chapter we aim at a presentation of the classical Riemann integral, often just called *the integral*, as the other types of integrals only occur in more advanced literature. We begin in Section 6.1 by considering integrals of scalar functions over bounded intervals $[a, b]$ in $\mathbb{R}$. In Section 6.3 we first mimic the procedure to introduce integration over rectangles in $\mathbb{R}^2$, and then show how to integrate scalar functions of two variables over more complicated domains in $\mathbb{R}^2$. In Section 6.4 we introduce a technique that can often be used to transform a "complicated integral" into a "simpler integral"; a special and particularly useful version of this technique is discussed in Section 6.5. A natural generalization of integration in $\mathbb{R}^2$ to $\mathbb{R}^n$ is presented in Section 6.6. Finally, in Section 6.7 we consider integration of vector functions of n variables.

6.1 The Riemann integral of functions of one variable

As prelude to the later and more advanced sections in this chapter, we will start by considering integration of functions $f : [a, b] \to \mathbb{R}$, where $[a, b]$ is a bounded and closed interval in $\mathbb{R}$. The integral can be defined in two equivalent ways, namely, either as a limit for certain sums (to be specified below) or as the anti-derivative of the function f. We will describe both ways. We expect the reader to have some basic knowledge about the topic, but maybe at a less technical level.

To get started, let us assume that the function $f : [a, b] \to \mathbb{R}$ is continuous

and positive, that is, $f(x) \geq 0$ for all $x \in [a, b]$. Let us imagine that we want to find an approximate value for the area of the subset in $\mathbb{R}^2$ that is bounded by the lines $x = a, x = b$, the x-axis, and the graph of the function f. In order to do this, fix some $J \in \mathbb{N}$ and split the interval $[a, b]$ into J subintervals $Q_j, j = 1, \ldots, J$, given by

$$Q_1 = [x_0, x_1], Q_2 = [x_1, x_2], \ldots, Q_J = [x_{J-1}, x_J], \tag{6.1}$$

where $x_0, x_1, \ldots, x_J \in [a, b]$ satisfy:

$$a = x_0 < x_1 < x_2 < \cdots < x_{J-1} < x_J = b. \tag{6.2}$$

The length of each interval $Q_j = [x_{j-1}, x_j]$ is denoted by $\Delta x_j = x_j - x_{j-1}$. One such choice is to divide $[a, b]$ in J intervals of equal length $\Delta x = \Delta x_j = \frac{b-a}{J}$:

$$Q_1 = \left[a, a + \tfrac{b-a}{J}\right], Q_2 = \left[a + \tfrac{b-a}{J}, a + 2\tfrac{b-a}{J}\right], \ldots, Q_J = \left[a + (J-1)\tfrac{b-a}{J}, b\right].$$

For each interval Q_j, we pick an arbitrary point $\xi_j \in Q_j$. Then the associated *Riemann sum*, to be denoted by S_J, is defined as

$$S_J = \sum_{j=1}^{J} f(\xi_j)\Delta x_j. \tag{6.3}$$

Intuitively, if J is "large", the value S_J is a "good approximation" of the desired area; furthermore, typically we obtain "better approximations" by increasing J, i.e., by splitting the interval $[a, b]$ into finer intervals. See Figure 6.1 for an illustration of this.

The Riemann sums S_J indeed have a limit, denoted $I \in \mathbb{R}$, as $J \to \infty$ regardless of how we pick the subintervals Q_j and $\xi_j \in Q_j$; the limit I can be considered as the precise definition of the area of the considered domain. If the function f is not necessarily positive (but still continuous), one can prove that the Riemann sums S_J still converge to a real scalar I as $J \to \infty$, and the limit I is the "signed" area under the graph, see Theorem 6.1.1 on page 154.

However, we often need to integrate functions that are *not* continuous on their entire domain. This leads to the following definition.

Definition 6.1.1 Riemann integral in $\mathbb{R}$

A function $f : [a, b] \to \mathbb{R}$ is said to be *Riemann integrable* if there exists a number $I \in \mathbb{R}$ with the following property: For any $\epsilon > 0$ there exists $\delta > 0$ such that for any subdivision of $[a, b]$ into J intervals Q_j, $j = 1, \ldots, J$, with

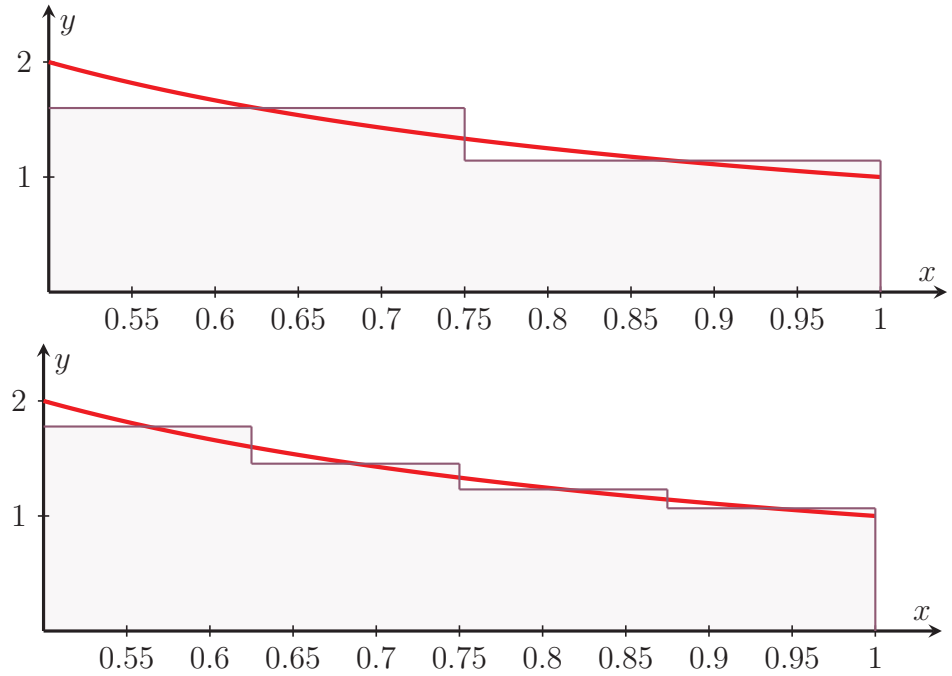


Figure 6.1: Riemann sums of the function $f(x) = x^{-1}$ on $x \in [1/2, 1]$. On the first figure, the interval $[1/2, 1]$ is split into two intervals (corresponding to $J = 2$.) The area "under the staircase" corresponds to the Riemann sum S_2 for the choice $\xi_1 = 5/8, \xi_2 = 7/8$. On the second figure, the interval $[1/2, 1]$ is split into four intervals (corresponding to $J = 4$.) The area "under the staircase" corresponds to the Riemann sum S_4 for the choice $\xi_1 = 9/16, \xi_2 = 11/16, \xi_3 = 13/16, \xi_4 = 15/16$. We see that already S_4 is a much better approximation to the area under the graph of the function f than S_2.

$\max_j \Delta x_j < \delta$ and for any $\xi_j \in Q_j$, the Riemann sum $S_J = \sum_{j=1}^{J} f(\xi_j)\Delta x_j$ satisfies:

$$\left| I - S_J \right| < \epsilon.$$

In case f is Riemann integrable over the interval $[a, b]$, the real scalar I is called *the integral of* f *over* $[a, b]$, and it is written as

$$\int_a^b f(x)\,\mathrm{d}x = I. \tag{6.4}$$

Often we will skip the name "Riemann" and simply speak about the integral of the function f over the interval $[a, b]$. It is important to remember that the symbol $\int_a^b f(x)\,\mathrm{d}x$ has only one "meaning": it is just the way we

denote the limit for the Riemann sums S_J as $J \to \infty$, indeed,

$$\int_a^b f(x)\,\mathrm{d}x = \lim_{J\to\infty} S_J = \lim_{J\to\infty} \sum_{j=1}^{J} f(\xi_j)\Delta x_j.$$

One can show that the number I from Definition 6.1.1 is uniquely determined. It is also important to note that although different choices of the points x_j and ξ_j, $j = 1, \ldots, J$, lead to different values for the Riemann sums S_J, the limit I is *independent* of the choice of these points.

As we already remarked, the Riemann sums always converge for continuous functions. Let us state this useful result as a theorem.

Theorem 6.1.1

Let $f : [a, b] \to \mathbb{R}$ be a function.

(i) If f is Riemann integrable, then f is bounded, i.e., the image $\mathrm{im}(f)$ is a bounded set in $\mathbb{R}$.

(ii) If f is continuous, then f is Riemann integrable.

Let us state, again without proof, a number of useful properties of the Riemann integral.

Theorem 6.1.2 Rules of Riemann integration

Let $f, g : [a, b] \to \mathbb{R}$ be functions.

(i) *(linearity)* Let $c, d \in \mathbb{R}$. If f and g are Riemann integrable functions, then so is $cf + dg$ and

$$\int_a^b (cf(x) + dg(x))\,\mathrm{d}x = c\int_a^b f(x)\,\mathrm{d}x + d\int_a^b g(x)\,\mathrm{d}x. \tag{6.5}$$

(ii) *(monotonicity)* If f and g are Riemann integrable functions and $f(x) \le g(x)$ for all $x \in [a, b]$, then

$$\int_a^b f(x)\,\mathrm{d}x \le \int_a^b g(x)\,\mathrm{d}x. \tag{6.6}$$

(iii) *(triangle inequality)* If f and $|f|$ are Riemann integrable functions, then

$$\left|\int_a^b f(x)\,\mathrm{d}x\right| \le \int_a^b |f(x)|\,\mathrm{d}x. \tag{6.7}$$

(iv) *(insertion rule)* Let $x_0 \in]a, b[$. If f is Riemann integrable over $[a, x_0]$ and over $[x_0, b]$, then f is Riemann integrable over $[a, b]$ and

$$\int_a^b f(x)\,\mathrm{d}x = \int_a^{x_0} f(x)\,\mathrm{d}x + \int_{x_0}^b f(x)\,\mathrm{d}x. \tag{6.8}$$

It turns out that it is useful to define $\int_a^b f(x)\,\mathrm{d}x$ not only for $b > a$ but also for $a < b$. We do this by setting, for $b \geq a$:

$$\int_b^a f(x)\,\mathrm{d}x = \begin{cases} -\int_a^b f(x)\,\mathrm{d}x & b > a, \\ 0 & b = a. \end{cases} \tag{6.9}$$

With this notation, we can phrase the insertion rule (6.8), given that each of the three integrals exist, as

$$\int_a^b f(x)\,\mathrm{d}x = \int_a^c f(x)\,\mathrm{d}x + \int_c^b f(x)\,\mathrm{d}x,$$

where $f : I \to \mathbb{R}$, $I \subseteq \mathbb{R}$ is any interval, and $a, b, c \in I$.

6.2 Anti-derivatives of functions of one variable

In general, it is cumbersome to calculate the integral $\int_a^b f(x)\,\mathrm{d}x$ as in Equation (6.4) on page 153, even for simple functions like $\cos(x)$ and x^n (and one often has to approximate it on a computer using numerical integration). This is the point, where the second interpretation of the integral comes in handy. It is based on the following concept.

Definition 6.2.1 Anti-derivative

Let $I \subseteq \mathbb{R}$ be an interval, and let $f : I \to \mathbb{R}$ be a function. A function $F : I \to \mathbb{R}$ is an *anti-derivative* of f on I if F is differentiable on I with

$$F'(x) = f(x), \quad \text{for all } x \in I. \tag{6.10}$$

Anti-derivatives are often denoted by a so-called *indefinite integral*:

$$F(x) = \int f(x)\,\mathrm{d}x, \tag{6.11}$$

while the Riemann integral is called a *definite integral*.

Anti-derivatives are known for a large class of standard functions in the mathematical analysis:

Example 6.2.1

Let us list a number of functions from the classical mathematical analysis and their corresponding anti-derivatives. Note that in each case we only mention one anti-derivative.

$$\int x^n \,\mathrm{d}x = \frac{1}{n+1}x^{n+1}, \quad \text{for } n \in \mathbb{R} \setminus \{-1\}, \tag{6.12}$$

$$\int \frac{1}{x} \,\mathrm{d}x = \ln(x), \tag{6.13}$$

$$\int \cos(x) \,\mathrm{d}x = \sin(x), \tag{6.14}$$

$$\int \sin(x) \,\mathrm{d}x = -\cos(x), \tag{6.15}$$

$$\int \mathrm{e}^x \,\mathrm{d}x = \mathrm{e}^x\,. \tag{6.16}$$

Be careful not to confuse $\int f(x)\,\mathrm{d}x$ with the Riemann integral. The symbol $\int f(x)\,\mathrm{d}x$ means a *function* whose derivative is f, while $\int_a^b f(x)\,\mathrm{d}x$ is a real *scalar*. However, the anti-derivative is, in fact, closely related to the integral defined in Definition 6.1.1 on page 152. More precisely, we have the following fundamental result.

Theorem 6.2.1 Fundamental Theorem of Calculus

Suppose $f : I \to \mathbb{R}$ is a continuous function. Then f has an anti-derivative on I given by

$$F(x) = \int_{x_0}^{x} f(y)\,\mathrm{d}y \quad \text{for } x \in I, \tag{6.17}$$

where $x_0 \in I$ is a fixed, but arbitrary point in I.

It is easy to show using the mean value theorem that if F is an anti-derivative of the function f, then all anti-derivative are given by the formula $F(x) + C$, where C is an arbitrary constant.

Hence, if a function has an anti-derivative, it actually has infinitely many anti-derivatives. On the other hand, there are many Riemann integrable functions that do not have an anti-derivative. It follows from Theorem 6.2.1 that such functions cannot be continuous. A simple example of Riemann integrable function with no anti-derivative is considered in the next exercise.

Exercise 6.2.2

Consider the function $f : \mathbb{R} \to \mathbb{R}$ given by

$$f(x) = \begin{cases} 1 & x \in [c,d], \\ 0 & x \notin [c,d], \end{cases}$$

where $d > c$. Show first that f is Riemann integrable. Show then that f cannot have an anti-derivative.

Corollary 6.2.2

Let F denote an anti-derivative of a continuous function $f : I \to \mathbb{R}$. Then, for any $a, b \in I$,

$$\int_a^b f(x)\,\mathrm{d}x = F(b) - F(a). \tag{6.18}$$

Often, the result in (6.18) is written in a slightly different form, namely as

$$\int_a^b f(x)\,\mathrm{d}x = [F(x)]_a^b\,; \tag{6.19}$$

this form shows directly that we calculate the integral by *first* calculating an anti-derivative of the function f, and *then* insert the endpoints a and b of the considered interval.

Typically, application of Corollary 6.2.2 simplifies the calculation of the Riemann integral. Let us demonstrate this with a few examples.

Example 6.2.3

Using Corollary 6.2.2 and (6.14), we directly see that

$$\int_0^\pi \cos(x)\,\mathrm{d}x = [\sin(x)]_0^\pi = \sin(\pi) - \sin(0) = 0.$$

Similarly, using (6.16),

$$\int_1^2 \mathrm{e}^x\,\mathrm{d}x = [\mathrm{e}^x]_1^2 = \mathrm{e}^2 - \mathrm{e}.$$

Many complicated integrals can be simplified (basically to an application of some of the results in Example 6.2.1) by using certain rules, stated next. For the precise statement of the results, we need to assume that functions, whose derivative appears inside an integral, are continuously differentiable as defined in Definition 3.1.2 on page 84.

Theorem 6.2.3 Rules of indefinite integration

(i) *(Partial integration)* If f is a continuous function, F an anti-derivative of f, and g is continuously differentiable, then

$$\int f(x)g(x)\,\mathrm{d}x = F(x)g(x) - \int F(x)g'(x)\,\mathrm{d}x. \tag{6.20}$$

(ii) *(Integration by substitution)* If f is a continuous function, F an anti-derivative of f, and g a continuously differentiable function such that its range is contained in the domain of the function f, then

$$\int f(g(x))g'(x)\,\mathrm{d}x = F(g(x)). \tag{6.21}$$

Obviously, Theorem 6.2.3 can also be formulated for Riemann integration. In particular, Equation (6.20) becomes

$$\int_a^b f(x)g(x)\,\mathrm{d}x = F(b)g(b) - F(a)g(a) - \int_a^b F(x)g'(x)\,\mathrm{d}x.$$

Note the particular structure the right hand side: it does not give the "final result" for $\int_a^b f(x)g(x)\,\mathrm{d}x$ since we still have to calculate $\int_a^b F(x)g'(x)\,\mathrm{d}x$ subsequently. In other words, partial integration is only a useful technique if the particular structure of the functions f and g implies that it is easier to calculate the integral $\int F(x)g'(x)\,\mathrm{d}x$ than the integral $\int f(x)g(x)\,\mathrm{d}x$.

The difficult in integration is typically to identify an integration rule that can simplify the given integral. Worse than that, there exist integrals that can not be expressed in terms of the standard functions from mathematical analysis. Let us illustrate Theorem 6.2.3 with a number of examples.

Example 6.2.4

We would like to calculate the integral

$$\int_1^2 2x\ln(x)\,\mathrm{d}x. \tag{6.22}$$

In order to do so, we first aim at determining

$$\int 2x\ln(x)\,\mathrm{d}x, x > 0. \tag{6.23}$$

We immediately see that the integrand in (6.23) can be considered as the product of two functions, namely $2x$ and $\ln(x)$. This makes it natural to try

to apply partial integration as in (6.20). First, let $f(x) = 2x, g(x) = \ln(x)$. Then $F(x) = x^2$ is an anti-derivative of f and $g'(x) = x^{-1}$. Thus, using Theorem 6.2.3(i),

$$\int 2x \ln(x)\,\mathrm{d}x = x^2 \ln(x) - \int x^2 x^{-1}\,\mathrm{d}x = x^2 \ln(x) - \int x\,\mathrm{d}x = x^2 \ln(x) - \frac{1}{2}x^2.$$

In this particular case we could actually also take the "opposite choice" of the functions f and g! Indeed, letting $f(x) = \ln(x), g(x) = 2x$, the function $F(x) = x\ln(x) - x$ is an anti-derivative of f, and $g'(x) = 2$. Thus,

$$\begin{aligned}\int 2x \ln(x)\,\mathrm{d}x &= (x\ln(x) - x)2x - \int (x\ln(x) - x)2\,\mathrm{d}x \qquad (6.24)\\ &= 2x^2\ln(x) - 2x^2 - \int 2x\ln(x)\,\mathrm{d}x + \int 2x\,\mathrm{d}x\\ &= 2x^2\ln(x) - x^2 - \int 2x\ln(x)\,\mathrm{d}x. \qquad (6.25)\end{aligned}$$

Note that at first sight, this calculation does not seems to be useful because it returns exactly the same integral in (6.25) as the one we start with in (6.24). However, exactly this observation allow us to *collect* the terms and write that

$$2\int 2x\ln(x)\,\mathrm{d}x = 2x^2\ln(x) - x^2,$$

again showing that

$$\int 2x\ln(x)\,\mathrm{d}x = x^2\ln(x) - \frac{1}{2}x^2.$$

Finally, after having calculated the integral in (6.23), we can apply (6.19) to obtain that

$$\begin{aligned}\int_1^2 2x\ln(x)\,\mathrm{d}x &= \left[x^2\ln(x) - \tfrac{1}{2}x^2\right]_1^2\\ &= (4\ln(2) - 2) - \left(\ln(1) - \tfrac{1}{2}\right) = 4\ln(2) - \tfrac{3}{2}.\end{aligned}$$

Example 6.2.5

Fix some $a, b \in \mathbb{R}$ with $a \neq 0$ and consider the integral

$$\int \cos(ax + b)\,\mathrm{d}x. \qquad (6.26)$$

The function $\cos(ax + b)$ can natural be considered as a composition of the functions $\cos(x)$ and $ax + b$, so it is natural to take $f(x) = \cos(x)$ and

$g(x) = ax + b$, and aim at an application of the rule in Theorem 6.2.3(ii). With the mentioned choice of the functions $f(x)$ and $g(x)$ we indeed obtain that $f(g(x)) = \cos(ax + b)$. However, the integral does not get exactly the form in (6.21), as the term $g'(x)$ is missing. But in the current case we can get away with this problem by observing that $g'(x) = a$, i.e., the missing term is just a constant! Therefore we can write the integral (6.26) as

$$\begin{aligned}\int \cos(ax+b)\,\mathrm{d}x &= \frac{1}{a}\int \cos(ax+b)a\,\mathrm{d}x = \frac{1}{a}\int f(g(x))g'(x)\,\mathrm{d}x \\ &= \frac{1}{a}F(g(x)) = \frac{1}{a}\sin(ax+b).\end{aligned}$$

Note that this calculation used the linearity rule in Theorem 6.1.2(i) in the first step!

Example 6.2.6

Using the results in Examples 6.2.4 and 6.2.5,

$$\begin{aligned}\int \left(10x\ln(x) + 3\cos(2x+\tfrac{\pi}{3})\right)\mathrm{d}x &= 5\int 2x\ln(x)\,\mathrm{d}x + 3\int \cos(2x+\tfrac{\pi}{3})\,\mathrm{d}x \\ &= 5\left(x^2\ln(x) - \tfrac{1}{2}x^2\right) + \tfrac{3}{2}\sin(2x+\tfrac{\pi}{3}) \\ &= 5x^2\ln(x) - \tfrac{5}{2}x^2 + \tfrac{3}{2}\sin(2x+\tfrac{\pi}{3}).\end{aligned}$$

6.3 The Riemann integral of functions of two variables

Integration in $\mathbb{R}^2$ (and $\mathbb{R}^n$) is in many regards completely similar to integration in $\mathbb{R}$. However, the notion of an anti-derivative does not generalize well to functions of several variables, and we will not return to this concept until Chapter 7. Moreover, even when defining the Riemann integral of a function of two variables, we will encounter technical issues arising from the fact that general sets in $\mathbb{R}^2$ can be significantly more complicated than just intervals in $\mathbb{R}$. In order to avoid these difficulties we will first consider integration over rectangles in $\mathbb{R}^2$.

A *rectangle* in $\mathbb{R}^2$ is a set of the form

$$Q = \left\{(x,y) \in \mathbb{R}^2 \;\middle|\; a_1 \le x \le b_1 \;\wedge\; a_2 \le y \le b_2\right\}. \tag{6.27}$$

The rectangle will also be written as $Q = [a_1, b_1] \times [a_2, b_2]$. Consider now a continuous function $f : Q \to \mathbb{R}$, and assume first that all function values

are positive, i.e., $f(x) \geq 0$ for all $x \in Q$. Our goal is to define the integral of f over Q; parallel to our definition of the integral in $\mathbb{R}$, it should be interpreted as the *volume* of the region between the graph of the function $f : Q \to \mathbb{R}$ and the xy-plane.

We will follow the procedure for the Riemann integral in $\mathbb{R}$ and first split the rectangle Q into smaller rectangles. Technically, this is done by fixing an $J \in \mathbb{N}$ and splitting the intervals $[a_1, b_1]$ and $[a_2, b_2]$ into J intervals of equal size. Denote the splitting points by $x_1, x_2, \ldots, x_{J-1}$ resp. $y_1, y_2, \ldots, y_{J-1}$, and let, for notational convenience, $x_0 \coloneqq a$, $x_J \coloneqq b$, $y_0 \coloneqq a_2$, and $y_J \coloneqq b$; then

$$a_1 = x_0 < x_1 < x_2 < \cdots < x_{J-1} < x_J = b_1, \tag{6.28}$$

and

$$a_2 = y_0 < y_1 < y_2 < \cdots < y_{J-1} < y_J = b_2. \tag{6.29}$$

The partition of the rectangle Q into smaller rectangles is now defined via

$$Q_{i,j} = [x_{i-1}, x_i] \times [y_{j-1}, y_j], \;\; i, j = 1, 2, \ldots, J. \tag{6.30}$$

Note that the length of each subinterval $[x_{i-1}, x_i]$ is $\Delta x \coloneqq \frac{b_1 - a_1}{J}$, and the length of each subinterval $[y_{j-1}, y_j]$ is $\Delta y \coloneqq \frac{b_2 - a_2}{J}$. Now, for each $i, j = 1, 2, \ldots, J$, pick an arbitrary point $\boldsymbol{\xi}_{i,j} \in Q_{i,j}$ and define the associated *Riemann sum* to be denoted by S_J, by

$$S_J = \sum_{i=1}^{J} \sum_{j=1}^{J} f(\boldsymbol{\xi}_{i,j}) \operatorname{area}(Q_{i,j}) \tag{6.31}$$

where $\operatorname{area}(Q_{i,j}) \coloneqq \Delta x \Delta y$ is the area of each rectangle $Q_{i,j}$.

Intuitively, if J is "large", the value S_J is a "good approximation" to the desired volume. Furthermore, we typically get better approximations by increasing the number of J. One can prove that the Riemann sums S_J indeed have a limit as $J \to \infty$; this limit can be considered as the definition of the *exact* value of the desired volume.

If the function f is not necessarily positive (but still continuous), one can prove that the Riemann sums S_J still converge. Hence, we can define the Riemann integral of the function f over the rectangle Q as

$$\int_Q f(x, y) \, \mathrm{d}(x, y) \coloneqq \lim_{J \to \infty} S_J = \lim_{J \to \infty} \sum_{i=1}^{J} \sum_{j=1}^{J} f(\boldsymbol{\xi}_{i,j}) \operatorname{area}(Q_{i,j}) \tag{6.32}$$

Although we now have a precise definition of integration over Q at hand, it is clear that it is tedious to apply the described procedure in practice. Fortunately, again the situation is similar to what we encountered in $\mathbb{R}$: we

can calculate the integral using the concept of anti-derivatives. In order to explain this, consider again the rectangle Q in (6.27) and a continuous function $f : Q \to \mathbb{R}$. Fix an arbitrary value for $y \in [a_2, b_2]$; then the function $x \mapsto f(x, y)$ is a continuous function, and we can calculate the integral

$$\int_{a_1}^{b_1} f(x, y)\,\mathrm{d}x, \tag{6.33}$$

e.g., as in Corollary 6.2.2. The expression in (6.33) is now a continuous function of the variable y; we can therefore apply Corollary 6.2.2 again, and form the integral of this function, i.e.,

$$\int_{a_2}^{b_2} \left(\int_{a_1}^{b_1} f(x, y)\,\mathrm{d}x \right) \mathrm{d}y. \tag{6.34}$$

It can be shown that the outcome of this *double integration* is identical with the Riemann integral introduced in (6.32). In fact, we could also have integrated first with respect to y and then with respect to x, and we would have arrived at the same value of the integration. Let us formulate this as a theorem:

Theorem 6.3.1 Integration over a rectangle in $\mathbb{R}^2$

Consider a rectangle Q in $\mathbb{R}^2$ of the form

$$Q = \Big\{ (x, y) \in \mathbb{R}^2 \;\Big|\; a_1 \le x \le b_1 \;\wedge\; a_2 \le y \le b_2 \Big\}, \tag{6.35}$$

and let $f : Q \to \mathbb{R}$ denote a continuous function. Then

$$\int_Q f(x, y)\,\mathrm{d}(x, y) = \int_{a_2}^{b_2} \left(\int_{a_1}^{b_1} f(x, y)\,\mathrm{d}x \right) \mathrm{d}y \tag{6.36}$$

$$= \int_{a_1}^{b_1} \left(\int_{a_2}^{b_2} f(x, y)\,\mathrm{d}y \right) \mathrm{d}x. \tag{6.37}$$

Theorem 6.3.1 says that integration over a rectangle in $\mathbb{R}^2$ can be reduced to calculation of *two* "standard integrals" over bounded intervals in $\mathbb{R}$. Let us illustrate formula (6.36) by an example.

Example 6.3.1

Let

$$Q = [2, 3] \times [0, 1] = \Big\{ (x, y) \in \mathbb{R}^2 \;\Big|\; x \in [2, 3] \;\wedge\; y \in [0, 1] \Big\},$$

and consider the function $f(x, y) = x^2 y$. In order to calculate

$\int_Q f(x,y)\,\mathrm{d}(x,y)$, we apply (6.36), which yields that

$$\int_Q f(x,y)\,\mathrm{d}(x,y) = \int_0^1 \left(\int_2^3 x^2 y\,\mathrm{d}x\right)\mathrm{d}y. \tag{6.38}$$

According to the definition of the double integral, we first consider the variable y as a constant and calculate the *inner integral* with respect to x:

$$\int_2^3 x^2 y\,\mathrm{d}x = \left[\frac{1}{3}x^3 y\right]_{x=2}^{x=3} = \frac{19}{3}y.$$

We now insert this result in (6.38) and perform the *outer integration* with respect to y:

$$\int_Q f(x,y)\,\mathrm{d}(x,y) = \int_0^1 \frac{19}{3}y\,\mathrm{d}y = \left[\frac{19}{3}\frac{1}{2}y^2\right]_0^1 = \frac{19}{6}.$$

The two formulas (6.36) and (6.37) in Theorem 6.3.1 show that we can *switch the order of integration*. It is a very useful technique that can be used in case calculation of an integral simplifies by changing the order of integration.

We are now ready to address the technically much more complicated question of integration over a general subset B in $\mathbb{R}^2$. We will assume that

(I) B is bounded, i.e.,
$$B \subset [a_1,b_1]\times[a_2,b_2] \tag{6.39}$$
for some $a_1,a_2,b_1,b_2 \in \mathbb{R}$.

(II) The boundary ∂B of B is formed by a finite number of continuously differentiable curves.

Recall that a curve is just (the image of) a vector function of one variable. We have already discussed boundary curves in Examples 2.2.4 and 2.2.5. The technical condition Item (II) guarantees that the boundary ∂B of B has an area that is zero (more precisely, ∂B can be covered with a set of arbitrarily small area).

The subset B might have a geometrically complicated shape, so in general we can not split it into small rectangles as we did for the set Q in (6.27). However, we can perform such a splitting on the "bigger" rectangle $[a_1,b_1]\times[a_2,b_2]$ in (6.39) as we did in the beginning of this section. That is, we again fix an integer $J \in \mathbb{N}$, split the intervals $[a_1,b_1]$ and $[a_2,b_2]$ into J intervals (not necessarily of equal length), and form the corresponding subrectangles $Q_{i,j}, i,j = 1,2,\ldots,J$; see Equations (6.28) to (6.30).

Among the rectangles $Q_{i,j}$, some of them will be *outside* the subset B; some of them will be *inside* B, and some of them will have parts inside B and parts outside B. We will exclusively consider the rectangles $Q_{i,j}$ that are completely contained in B, i.e., $Q_{i,j} \subset B$. For these sets $Q_{i,j}$, we again pick arbitrary points $\boldsymbol{\xi}_{i,j} \in Q_{i,j}$, and form the corresponding *Riemann sum*, to be denoted by S_J, defined by

$$S_J = \sum_{\{(i,j)|Q_{i,j}\subset B\}} f(\boldsymbol{\xi}_{i,j}) \operatorname{area}(Q_{i,j}), \tag{6.40}$$

where $\operatorname{area}(Q_{i,j}) \coloneqq \Delta x_i \Delta y_j$, and Δx_i and Δy_j are the side lengths of the rectangle $Q_{i,j}$ and $1 \leq i, j \leq J$.

In words: the Riemann sum is formed precisely like we did in (6.31), except that we only consider the rectangles $Q_{i,j}$ that are completely contained in the given subset B. With this setup, one can again show that if the function f is continuous, the Riemann sum S_J tends to a limit as $J \to \infty$. The limit is denoted by

$$\int_B f(x,y)\,\mathrm{d}(x,y), \quad \int_B f(x,y)\,\mathrm{d}\boldsymbol{x}, \tag{6.41}$$

or by

$$\int_B f(x,y)\,\mathrm{d}X, \tag{6.42}$$

and it is called the *Riemann integral of f over the subset B*.

For functions that are not continuous in all points in its domain, we need a slightly more technical limiting procedure, just as was the case in Definition 6.1.1 on page 152.

Definition 6.3.1 Riemann integral in $\mathbb{R}^2$

Consider a subset B in $\mathbb{R}^2$, which satisfies the conditions Items (I) and (II) on the preceding page. A function $f : B \to \mathbb{R}$ is said to be *Riemann integrable* if for any $\epsilon > 0$ there exists $\delta > 0$ such that

$$\left| \int_B f(x,y)\,\mathrm{d}(x,y) - S_J \right| < \epsilon$$

for *any* Riemann sum S_J with $\max_i \Delta x_i < \delta$, $\max_j \Delta y_j < \delta$ and any $\boldsymbol{\xi}_{i,j} \in Q_{i,j}$. The scalar $\int_B f(x,y)\,\mathrm{d}(x,y)$ is called the integral of f over B.

As for functions of one variable (see Theorem 6.1.1), continuous functions on $B \subset \mathbb{R}^2$, satisfying the conditions Items (I) and (II) on the previous page,

are guaranteed to be Riemann integrable. In particular, since the function $f(x) = 1$ for all $x \in B$ is continuous, it is Riemann integrable. This leads to an exact definition of the *area* of a subset B.

Definition 6.3.2 Area

Consider a subset B in $\mathbb{R}^2$, which satisfies the conditions Items (I) and (II) on page 163. Then the *area* of B is defined as

$$\text{area}(B) = \int_B 1 \, \mathrm{d}X. \tag{6.43}$$

In general it is complicated to calculate the integrals in Definition 6.3.1 and (6.43) via the described procedure, except if B has a simple geometric structure. We will now consider a particular case, where it can be done in a similar way as to what we did in Theorem 6.3.1.

Theorem 6.3.2 Integration over axis-parallel regions

(i) Consider two continuously differentiable functions

$$\alpha_1 : [a, b] \to \mathbb{R}, \qquad \alpha_2 : [a, b] \to \mathbb{R},$$

and assume that $\alpha_1(x) \leq \alpha_2(x)$ for all $x \in [a, b]$. Consider the subset B defined as

$$B = \Big\{(x, y) \in \mathbb{R}^2 \;\Big|\; a \leq x \leq b \wedge \alpha_1(x) \leq y \leq \alpha_2(x)\Big\}. \tag{6.44}$$

Then, for any continuous function $f : B \to \mathbb{R}$,

$$\int_B f(x, y) \, \mathrm{d}(x, y) = \int_a^b \left(\int_{\alpha_1(x)}^{\alpha_2(x)} f(x, y) \, \mathrm{d}y \right) \mathrm{d}x. \tag{6.45}$$

(ii) Similarly, if the subset B can be described as

$$B = \Big\{(x, y) \in \mathbb{R}^2 \;\Big|\; c \leq y \leq d \wedge \beta_1(y) \leq x \leq \beta_2(y)\Big\} \tag{6.46}$$

for some continuously differentiable functions $\beta_1, \beta_2 : [c, d] \to \mathbb{R}$ with $\beta_1(y) \leq \beta_2(y)$ for all $y \in [c, d]$, then for any continuous function $f : B \to \mathbb{R}$,

$$\int_B f(x, y) \, \mathrm{d}(x, y) = \int_c^d \left(\int_{\beta_1(y)}^{\beta_2(y)} f(x, y) \, \mathrm{d}x \right) \mathrm{d}y. \tag{6.47}$$

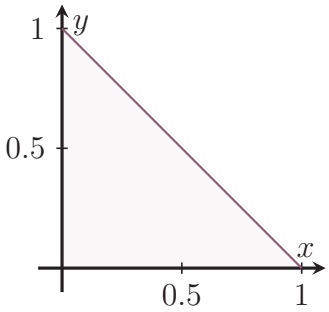


Figure 6.2: The domain B in (6.48).

In words, Theorem 6.3.2 says that integration over a subset in $\mathbb{R}^2$ of the special types (6.44) and (6.46) can be reduced to calculation of *two* "standard integrals" in $\mathbb{R}$. In particular, if both descriptions (6.44) and (6.46) are available, we are free to choose the order of integration in the way that is most convenient; this observation generalizes what we saw in Theorem 6.3.1. Let us illustrate this by an example.

Example 6.3.2

Let

$$B = \Big\{(x,y) \in \mathbb{R}^2 \;\Big|\; 0 \le x \le 1 \;\wedge\; 0 \le y \le 1 - x\Big\}, \tag{6.48}$$

and consider the function $f(x,y) = 2xy$. Then, according to Theorem 6.3.2(i),

$$\begin{aligned}
\int_B f(x,y)\,\mathrm{d}(x,y) &= \int_0^1 \left(\int_0^{1-x} 2xy\,\mathrm{d}y\right)\mathrm{d}x \\
&= \int_0^1 \Big[xy^2\Big]_{y=0}^{y=1-x}\,\mathrm{d}x \\
&= \int_0^1 x(1-x)^2\,\mathrm{d}x \\
&= \int_0^1 (x^3 - 2x^2 + x)\,\mathrm{d}x = \frac{1}{12}.
\end{aligned}$$

The set B in (6.48) is illustrated on Figure 6.2. It also can be described as

$$B = \Big\{(x,y) \in \mathbb{R}^2 \;\Big|\; 0 \le y \le 1 \;\wedge\; 0 \le x \le 1 - y\Big\}.$$

Thus, we can also apply Theorem 6.3.2(ii), which yield that

$$\int_B f(x,y)\,\mathrm{d}(x,y) = \int_0^1 \left(\int_0^{1-y} 2xy\,\mathrm{d}x\right)\mathrm{d}y.$$

From here we can perform the same operations as above, leading to (we skip some of the details)

$$\begin{aligned}\int_B f(x,y)\,\mathrm{d}(x,y) &= \int_0^1 \left[x^2 y\right]_{x=0}^{x=1-y}\,\mathrm{d}y \\ &= \int_0^1 y(1-y)^2\,\mathrm{d}y = \frac{1}{12}.\end{aligned}$$

In this particular example, the two versions of Theorem 6.3.2 lead to the same calculation, but in many cases one of the results is more convenient to apply than the other.

6.4 Change of variables in $\mathbb{R}^2$

Section 6.3 provides us with a very general theory for integrations over subsets in $\mathbb{R}^2$. However, except in the case of integration over rectangles or cases where results of the type in Theorem 6.3.2 apply, concrete calculations of integrals will be tedious. The purpose of this section is to introduce a technique that in certain cases can transform integrals over complicated subsets into integrals over "simpler subsets", e.g., rectangles. We will consider an important concrete case in Section 6.5.

The main tool to transfer integration between "complicated" subsets and rectangles is vector functions of the following form.

Definition 6.4.1 Jacobian determinant

Let A be a subset in $\mathbb{R}^2$. Consider a function $\boldsymbol{r} : A \to \mathbb{R}^2$, i.e., $\boldsymbol{r}(u,v) = (r_1(u,v), r_2(u,v))$ with coordinate functions $r_i : A \to \mathbb{R}$. Suppose $\boldsymbol{r}$ is a C^1 vector function on A°, i.e., the partial derivatives of the coordinate functions exist and are continuous on the interior of A. Let $\boldsymbol{J_r}(u,v) \in \mathsf{M}_{2\times 2}(\mathbb{R})$ denote the Jacobian matrix of $\boldsymbol{r}$ at $(u,v) \in U$. The determinant of $\boldsymbol{J_r}(u,v)$ is called the *Jacobian determinant* $\det\left(\boldsymbol{J_r}(u,v)\right)$ of $\boldsymbol{r}$ at $(u,v) \in U$.

We can be completely explicitly in Definition 6.4.1. The Jacobian matrix of $\boldsymbol{r} = (r_1, r_2)$ at (u,v) is:

$$\boldsymbol{J_r}(u,v) = \begin{bmatrix} \frac{\partial r_1}{\partial u}(u,v) & \frac{\partial r_1}{\partial v}(u,v) \\ \frac{\partial r_2}{\partial u}(u,v) & \frac{\partial r_2}{\partial v}(u,v) \end{bmatrix}, \tag{6.49}$$

and the Jacobian determinant is:

$$\det\left(\boldsymbol{J_r}(u,v)\right) = \begin{vmatrix} \frac{\partial r_1}{\partial u}(u,v) & \frac{\partial r_1}{\partial v}(u,v) \\ \frac{\partial r_2}{\partial u}(u,v) & \frac{\partial r_2}{\partial v}(u,v) \end{vmatrix} \tag{6.50}$$

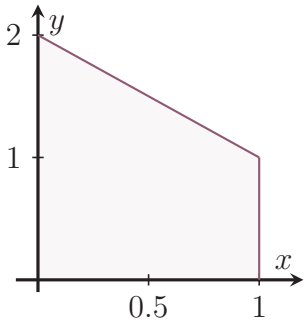


Figure 6.3: The domain B in (6.52).

$$= \frac{\partial r_1}{\partial u}(u,v)\frac{\partial r_2}{\partial v}(u,v) - \frac{\partial r_2}{\partial u}(u,v)\frac{\partial r_1}{\partial v}(u,v). \tag{6.51}$$

Recall that the row vectors in the Jacobian matrix consist precisely of the gradient vectors for the functions r_1 and r_2.

Example 6.4.1

The function

$$\boldsymbol{r}(u,v) = \begin{bmatrix} u^2 v \\ u + v^2 \end{bmatrix}, \quad (u,v) \in \mathbb{R}^2,$$

is a continuously differentiable map from $\mathbb{R}^2$ to $\mathbb{R}^2$. Its Jacobian matrix is

$$\boldsymbol{J_r}(u,v) = \begin{bmatrix} \frac{\partial r_1}{\partial u}(u,v) & \frac{\partial r_1}{\partial v}(u,v) \\ \frac{\partial r_2}{\partial u}(u,v) & \frac{\partial r_2}{\partial v}(u,v) \end{bmatrix} = \begin{bmatrix} 2uv & u^2 \\ 1 & 2v \end{bmatrix},$$

and its Jacobian determinant is

$$\det\left(\boldsymbol{J_r}(u,v)\right) = \det \begin{bmatrix} 2uv & u^2 \\ 1 & 2v \end{bmatrix} = 4uv^2 - u^2.$$

Let us now turn to the question of integrating a function $f : B \to \mathbb{R}$ over a potentially complicated subset $B \subset \mathbb{R}^2$. The key assumption in the subsequent Theorem 6.4.1 is that we can find a injective and surjective map or function $\boldsymbol{r} : \Gamma \to B$ from a "simple subset" Γ in $\mathbb{R}^2$ onto B and then turn the "complicated" integral over B into a "simple integral" over Γ. Before stating Theorem 6.4.1, let us consider an example of such sets B, Γ, and a corresponding map $\boldsymbol{r}$. Before reading this example, it is a good idea to remind yourself how we parametrized a line between two points in Example 3.1.7 on page 87.

Example 6.4.2 Parametrization of a subset

Consider the subset

$$B = \Big\{(x,y) \in \mathbb{R}^2 \;\Big|\; 0 \le x \le 1 \;\wedge\; y \ge 0 \;\wedge\; x + y \le 2\Big\}, \tag{6.52}$$

which is illustrated on Figure 6.3.

We will now argue that the map

$$\boldsymbol{r} : \mathbb{R}^2 \to \mathbb{R}^2, \quad \boldsymbol{r}(u,v) = \begin{bmatrix} u \\ v(2-u) \end{bmatrix} \tag{6.53}$$

is mapping the rectangle

$$\Gamma = [0,1]^2 = \Big\{(u,v) \in \mathbb{R}^2 \;\Big|\; 0 \le u \le 1 \;\wedge\; 0 \le v \le 1\Big\}$$

bijectively onto the set B.

In order to do so, consider first a fixed point $(x,y) \in B$ that is located on the "upper" boundary, that is, $x \in [0,1]$ and $x + y = 2$. Then, for $u = x$

$$\boldsymbol{r}(x,1) = (x, 2-x) = (x,y),$$

i.e., the given point (x,y) is indeed the image of the point $(u,1) \in \Gamma$. Also, if we consider a point $(x,y) \in B$ that is located on the "lower" boundary, then $x \in [0,1]$ and $y = 0$; such a point is also in the image of the set Γ under the map $\boldsymbol{r}$ because

$$\boldsymbol{r}(u,0) = (u,0).$$

Fixing now any $u \in [0,1]$, and letting the parameter $v \in [0,1]$ traverse from $v = 0$ to $v = 1$, the point $\boldsymbol{r}(u,v)$ will initiate in $\boldsymbol{r}(u,0) = (u,0)$ and traverse through a horizontal line to its terminating point $\boldsymbol{r}(u,1) = (u, 2-u)$. That is, for any fixed $u \in [0,1]$ the function $\boldsymbol{r}(u,v)$ traverse from the point $(u,0)$ on the "lower" boundary to the point $(u,2-u)$ on the "upper" boundary. When we now let $u \in [0,1]$ vary, we thus parametrize all horizontal lines connecting the "lower" and "upper" boundary. In other words: each point (x,y) has a representation as $(x,y) = \boldsymbol{r}(u,v)$ for some $(u,v) \in \Gamma$. This shows that $\boldsymbol{r}$ indeed maps the set Γ onto B.

When we consider $\boldsymbol{r}$ as a map on Γ, i.e., $\boldsymbol{r} : \Gamma \to B$, then $\boldsymbol{r}$ is also injective. In order to see this, assume that $\boldsymbol{r}(u_1,v_1) = \boldsymbol{r}(u_2,v_2)$ for some $(u_1,v_1),(u_2,v_2) \in \Gamma$. Then, by writing out how $\boldsymbol{r}$ acts on the vectors,

$$u_1 = u_2 \text{ and } v_1(2-u_1) = v_2(2-u_2). \tag{6.54}$$

Using now that $(u_1,v_1),(u_2,v_2) \in \Gamma$, we know that $0 \le u_1 \le 1$, and therefore $2 - u_1 = 2 - u_2 \neq 0$. Thus the second equality in (6.54) implies that $v_1 = v_2$,

and therefore $(u_1, v_1) = (u_2, v_2)$. This proves that $\boldsymbol{r}$ indeed is injective on the set Γ.

Altogether we have now proved that $\boldsymbol{r}$ maps Γ *bijectively* onto B. We call $(u, v) \in \Gamma$ parameters, and we say that the map $\boldsymbol{r}$ in (6.53) provides a *parametrization* of the set B.

We are now able to state the important *change of variables theorem*. The result is quite involved, and we will explain how to apply it after having stated it.

Theorem 6.4.1 Change of variables in $\mathbb{R}^2$

Let Γ be a bounded subset in $\mathbb{R}^2$ with boundary formed by a finite number of continuously differentiable curves. Suppose that there exists a C^1 vector function $\boldsymbol{r} : \Gamma \to \mathbb{R}^2$ that is injective on the interior Γ° and such that

$$\det\left(\boldsymbol{J}_{\boldsymbol{r}}(u, v)\right) \neq 0 \text{ for all } (u, v) \in \Gamma^\circ. \tag{6.55}$$

Then, for any continuous function $f : \boldsymbol{r}(\Gamma) \to \mathbb{R}$,

$$\int_{\boldsymbol{r}(\Gamma)} f(x, y)\, \mathrm{d}\boldsymbol{x} = \int_{\Gamma} f(r_1(u, v), r_2(u, v)) |\det\left(\boldsymbol{J}_{\boldsymbol{r}}(u, v)\right)|\, \mathrm{d}\boldsymbol{u}, \tag{6.56}$$

where $\boldsymbol{x} = (x, y)$ and $\boldsymbol{u} = (u, v)$.

Note that (6.56) involves the *absolute value* of the Jacobian determinant of the map $\boldsymbol{r}$. The factor $|\det\left(\boldsymbol{J}_{\boldsymbol{r}}(u, v)\right)|$ describes the local *change in area* of a small region around $(u, v) \in \Gamma^\circ$ *under* the mapping $\boldsymbol{r}$. We will make this statement precise and outline the proof of Theorem 6.4.1 when we consider the change of variables for functions of n variables in Theorem 6.6.2 on page 178. Let us here just consider a very simple example that shows that an "area correcting" factor is needed.

Example 6.4.3

Let $B = [0, 2] \times [-3, 0]$. The area of B is 6. Let us also compute this trivial statement using both (6.43) and (6.56). By Definition 6.3.2 on page 165 we get:

$$\begin{aligned} \text{area}(B) &= \int 1\, \mathrm{d}\boldsymbol{x} = \int_{-3}^{0} \left(\int_{0}^{2} 1\, \mathrm{d}x \right) \mathrm{d}y = \int_{-3}^{0} [x]_0^2\, \mathrm{d}y \\ &= \int_{-3}^{0} 2\, \mathrm{d}y = [2y]_{-3}^{0} = 6. \end{aligned}$$

We can parametrize B using $\boldsymbol{r} : \Gamma \to \mathbb{R}^2$ given by $\boldsymbol{r}(u, v) = (2u, -3v)$

and $\Gamma = [0,1]^2$. The Jacobian matrix is $\boldsymbol{J}_{\boldsymbol{r}}(u,v) = \operatorname{diag}(2,-3)$, and its determinant is -6. Using (6.56) we get

$$\begin{aligned}\operatorname{area}(B) &= \int_{\boldsymbol{r}(\Gamma)} 1\,\mathrm{d}\boldsymbol{x} = \int_{\Gamma} |-6|\,\mathrm{d}\boldsymbol{u} = \int_0^1 \left(\int_0^1 6\,\mathrm{d}u\right)\mathrm{d}v \\ &= \int_0^1 6[u]_0^1\,\mathrm{d}v = \int_0^1 6\,\mathrm{d}v = [6v]_0^1 = 6.\end{aligned}$$

Without the factor $|\det(\boldsymbol{J}_{\boldsymbol{r}}(u,v))| = 6$, the latter integral would have given us 1. The Jacobian factor exactly describes the factor by which we "stretch" the unit square Γ to "fit" onto $B = \boldsymbol{r}(\Gamma)$. Note also that we need the absolute value of the Jacobian determinant as we would have otherwise arrived at a negative area.

For the computations in Example 6.4.3 we certainly did not need the "change of variables" theorem since B was a simple rectangle. Let us return the goal of this section: to obtain a tool to calculate integrals $\int_B f(x,y)\,\mathrm{d}\boldsymbol{x}$, when B is a *complicated* subset. The idea is to *identify* or *construct* a map $\boldsymbol{r} : \mathbb{R}^2 \to \mathbb{R}^2$ and a *simple* domain $\Gamma \subset \mathbb{R}^2$ satisfying the conditions in Theorem 6.4.1 such that $\boldsymbol{r}(\Gamma) = B$; in that case, we obtain that

$$\begin{aligned}\int_B f(x,y)\,\mathrm{d}(x,y) &= \int_{\boldsymbol{r}(\Gamma)} f(x,y)\,\mathrm{d}(x,y) \\ &= \int_{\Gamma} f(r_1(u,v), r_2(u,v))|\det(\boldsymbol{J}_{\boldsymbol{r}}(u,v))|\,\mathrm{d}(u,v). \qquad (6.57)\end{aligned}$$

The hope is now that the simpler structure of the set Γ compared with B makes it possible to calculate the integral (6.57).

In practice, the difficult task is to identity an appropriate map $\boldsymbol{r}$ and the set Γ. Let us consider a concrete case, where we can get through with the technical conditions, based on what we already know from Example 6.4.2.

Example 6.4.4

Let us consider the subset B from Example 6.4.2 on page 169. Our aim is to calculate the integral

$$\int_B 2xy\,\mathrm{d}(x,y). \qquad (6.58)$$

The function $\boldsymbol{r}$ in (6.53) maps the rectangle $\Gamma = [0,1]^2$ injectively and surjectively onto B. Hence, $\boldsymbol{r}(\Gamma) = B$. In order to apply Theorem 6.4.1, we therefore calculate the Jacobian determinant

$$\boldsymbol{J}_{\boldsymbol{r}}(u,v) = \begin{vmatrix} 1 & -v \\ 0 & 2-u \end{vmatrix} = 2-u,$$

and note that $\boldsymbol{J_r}(u,v) = 2-u > 0$ on $(u,v) \in \Gamma$. Parametrizations for which the Jacobian determinant is nonzero on Γ are sometimes called *regular*, cf. regular parametrizations Definition 3.1.4 on page 87.

Thus, applying (6.57) with $f(x,y) = 2xy$,

$$\begin{aligned}\int_B 2xy \,\mathrm{d}(x,y) &= \int_\Gamma f(r_1(u,v), r_2(u,v))|\det(\boldsymbol{J_r}(u,v)|)\,\mathrm{d}(u,v)\\ &= \int_{[0,1]^2} 2uv(2-u)|2-u|\,\mathrm{d}(u,v).\end{aligned}$$

Since the Jacobian determinant $2-u$ is positive on the considered interval $u \in [0,1]$, we can remove the absolute sign on the Jacobian determinant, and we obtain that

$$\begin{aligned}\int_B 2xy \,\mathrm{d}(x,y) &= \int_0^1 \left(\int_0^1 2uv(2-u)(2-u)\,\mathrm{d}u\right)\mathrm{d}v\\ &= \int_0^1 \int_0^1 2v(4u-4u^2+u^3)\,\mathrm{d}u\,\mathrm{d}v\\ &= \int_0^1 2v\left[2u^2 - \frac{4}{3}u^3 + \frac{1}{4}u^4\right]_{u=0}^{u=1}\mathrm{d}v\\ &= \frac{11}{12}\int_0^1 2v\,\mathrm{d}v = \frac{11}{12}\left[v^2\right]_0^1 = \frac{11}{12}.\end{aligned}$$

where we used Theorem 6.3.1 on page 162 to compute the integral over $\Gamma = [0,1]^2$.

Note that, in this particular case, we could have obtained the result in Example 6.4.4 in a much simpler way using the same procedure as in Example 6.3.2. In Section 6.5 we will see an application, where Theorem 6.4.1 is of substantial importance.

6.5 Polar coordinates

In this section we will discuss an important application of Theorem 6.4.1. For subsets B in $\mathbb{R}^2$ having a convenient representation in polar coordinates, it leads to a useful and explicit formula for integrals of the type $\int_B f(x,y)\,\mathrm{d}x\,\mathrm{d}y$.

Recall that a point in $\mathbb{R}^2$ can be represented either by its *rectangular coordinates* (also called *Cartesian coordinates*)

$$(x,y), \quad x,y \in \mathbb{R},$$

or its *polar coordinates*

$$(r, \theta), \quad r \in [0, \infty[, \theta \in]-\pi, \pi].$$

The relation between these two types of coordinates is that

$$x = r\cos(\theta), \quad y = r\sin(\theta).$$

trigonometric functions are 2π-periodic in the angle θ hence we can use any interval of length 2π, e.g., $\theta \in]-\pi, \pi]$ as above, or $\theta \in [0, 2\pi[$.

Example 6.5.1 Parametrization of a subset

Consider the subset

$$B = \Big\{(x, y) \in \mathbb{R}^2 \;\Big|\; x \geq 0 \wedge y \geq x \wedge x^2 + y^2 \leq 2\Big\}. \tag{6.59}$$

The region B is an eighth of a circular disc with radius $\sqrt{2}$, exactly, the piece above the line $y = x$ and to the right of $x = 0$ (including both lines). This region is easy to describe using polar coordinates as all pairs (r, θ) with $r \in [0, \sqrt{2}]$ and $\theta \in [\pi/4, \pi/2]$. This description leads to the following parametrization of B with $\Gamma = [0, \sqrt{2}] \times [\pi/4, \pi/2]$:

$$\boldsymbol{p} : \Gamma \to \mathbb{R}^2, \quad \boldsymbol{p}(r, \theta) = (r\cos(\theta), r\sin(\theta)).$$

The function $\boldsymbol{p}$ is a transformation from polar coordinates (r, θ) to Cartesian coordinates (x, y). The function is surjective. It is also injective, but only on the interior of Γ°. It fails to be injective on Γ since $\boldsymbol{p}(0, \theta) = (0, 0)$ for all values of θ.

In Section 6.4 we used parameters (u, v) and denoted the vector function $\boldsymbol{r}$. Here, we use $\boldsymbol{p}$ instead, as it is standard to work with (r, θ) as parameters in polar coordinates. Using $\boldsymbol{r}$ in this context could lead to confusion due to its association with the radial parameter r. This is, of course, just a matter of notation. Moreover, we do not need to use $[0, \sqrt{2}] \times [\pi/4, \pi/2]$ as domain for our parameters. We are always able to re-scale the parameter domains of the form $[a, b] \times [c, d]$ to $[0, 1]^2$, e.g., the function

$$\boldsymbol{r} : [0, 1]^2 \to \mathbb{R}^2, \;\; \boldsymbol{r}(u, v) = \Big(\sqrt{2}u\cos(\tfrac{\pi}{4}(1 + v)), \sqrt{2}u\sin(\tfrac{\pi}{4}(1 + v))\Big)$$

is an alternative parametrization of B. In particular, parametrizations are never unique, but, and this is the important fact, the value of the integral $\int_B f(x, y)\,\mathrm{d}x\,\mathrm{d}y$ is *independent* of the chosen parametrization.

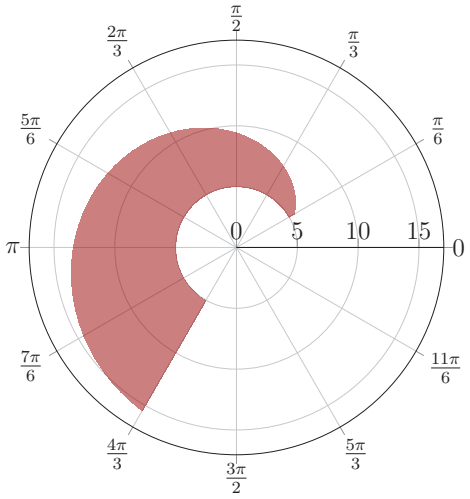


Figure 6.4: The subset $B \subset \mathbb{R}^2$ given by Equation (6.62) with $\alpha = \frac{\pi}{6}$, $\beta = \frac{4\pi}{3}$, $\varphi_1(\theta) = 5$ and $\varphi_2(\theta) = \sqrt{\frac{180\theta}{\pi}}$.

Assume now that we have given $\alpha, \beta \in \mathbb{R}$ such that $0 \leq \alpha < \beta \leq 2\pi$, and two functions

$$\varphi_1 : [\alpha, \beta] \to \mathbb{R}, \quad \varphi_2 : [\alpha, \beta] \to \mathbb{R} \tag{6.60}$$

such that

$$0 \leq \varphi_1(\theta) \leq \varphi_2(\theta), \text{for all } \theta \in [\alpha, \beta]. \tag{6.61}$$

Associated with such functions, we consider the subset B in $\mathbb{R}^2$ which is given by

$$B = \Big\{(x, y) = (r\cos(\theta), r\sin(\theta)) \;\Big|\; \alpha \leq \theta \leq \beta \wedge \varphi_1(\theta) \leq r \leq \varphi_2(\theta)\Big\}. \tag{6.62}$$

An example of such a set is shown in Figure 6.4. The subset B considered Example 6.5.1 is also a set of this form.

In general, a subset B of this form will have a complicated representation in terms of rectangular (x, y)-coordinates which makes it cumbersome to calculate an integral of the form $\int_B f(x, y)\,\mathrm{d}x\,\mathrm{d}y$ directly. This is the point, where Theorem 6.4.1 turns out to be helpful. Indeed, consider the map

$$\boldsymbol{p} : \mathbb{R}^2 \to \mathbb{R}^2, \quad \boldsymbol{p}(r, \theta) = \begin{bmatrix} r\cos(\theta) \\ r\sin(\theta) \end{bmatrix}. \tag{6.63}$$

Then $\boldsymbol{p}$ is mapping the subset

$$\Gamma = \Big\{(r, \theta) \in \mathbb{R}^2 \;\Big|\; \alpha \leq \theta \leq \beta \;\wedge\; \varphi_1(\theta) \leq r \leq \varphi_2(\theta)\Big\} \tag{6.64}$$

bijectively onto the subset B; that is, $\boldsymbol{p}$ is injective on Γ and $\boldsymbol{p}(\Gamma) = B$. The Jacobian matrix of $\boldsymbol{p}$ is

$$\boldsymbol{J}_{\boldsymbol{p}}(r,\theta) = \begin{bmatrix} \frac{\partial p_1}{\partial r}(r,\theta) & \frac{\partial p_1}{\partial \theta}(r,\theta) \\ \frac{\partial p_2}{\partial r}(r,\theta) & \frac{\partial p_2}{\partial \theta}(r,\theta) \end{bmatrix} = \begin{bmatrix} \cos(\theta) & -r\sin(\theta) \\ \sin(\theta) & r\cos(\theta) \end{bmatrix},$$

so its Jacobian determinant is

$$\det(\boldsymbol{J}_{\boldsymbol{p}}(r,\theta)) = \begin{vmatrix} \cos(\theta) & -r\sin(\theta) \\ \sin(\theta) & r\cos(\theta) \end{vmatrix} = r\cos^2(\theta) + r\sin^2(\theta) = r.$$

Note that $\det(\boldsymbol{J}_{\boldsymbol{p}}(r,\theta)) > 0$ for $(r,\theta) \in \Gamma^\circ$. Hence, the technical condition (6.55) is satisfied.

Hence, by Theorem 6.4.1 (see the version stated in (6.57)), for any continuous function $f : B \to \mathbb{R}$,

$$\begin{aligned} \int_B f(x,y)\,\mathrm{d}(x,y) &= \int_\Gamma f(p_1(r,\theta),p_2(r,\theta))\left|\det\left(\boldsymbol{J}_{\boldsymbol{p}}(r,\theta)\right)\right|\mathrm{d}(r,\theta) \\ &= \int_\Gamma f(r\cos(\theta),r\sin(\theta))r\,\mathrm{d}(r,\theta) \\ &= \int_\alpha^\beta \int_{\varphi_1(\theta)}^{\varphi_2(\theta)} f(r\cos(\theta),r\sin(\theta))r\,\mathrm{d}r\,\mathrm{d}\theta. \end{aligned}$$

Let us formulate the obtained result as a theorem.

Theorem 6.5.1 Integration in polar coordinates

Consider a subset B in $\mathbb{R}^2$ of the form (6.62). Then, for any continuous function $f : B \to \mathbb{R}$,

$$\int_B f(x,y)\,\mathrm{d}(x,y) = \int_\alpha^\beta \int_{\varphi_1(\theta)}^{\varphi_2(\theta)} f(r\cos(\theta),r\sin(\theta))r\,\mathrm{d}r\,\mathrm{d}\theta. \tag{6.65}$$

Example 6.5.2

Let us continue Example 6.5.1 with B given as Equation (6.59). We want to calculate the integral $\int_B f(x,y)\,\mathrm{d}x\,\mathrm{d}y$ for the function $f : B \to \mathbb{R}$ given by $f(x,y) = x^2 + y^2$ using Theorem 6.5.1. Hence, we first need to express f is polar coordinates:

$$f(r\cos(\theta),r\sin(\theta)) = (r\cos(\theta))^2 + (r\sin(\theta)))^2 = r^2.$$

Then we need to express B in the form (6.62). We see that $\varphi_1(\theta) = 0$ and $\varphi_2(\theta) = \sqrt{2}$ for $\theta \in [\alpha,\beta] = [\pi/4,\pi/2]$. By Theorem 6.5.1, we finally

compute

$$\int_B f(x,y)\,\mathrm{d}(x,y) = \int_{\frac{\pi}{4}}^{\frac{\pi}{2}} \int_0^{\sqrt{2}} r^2\, r\,\mathrm{d}r\,\mathrm{d}\theta = \left(\frac{\pi}{2} - \frac{\pi}{4}\right) \int_0^{\sqrt{2}} r^3\,\mathrm{d}r$$
$$= \frac{\pi}{4}\left[\frac{1}{4}r^4\right]_0^{\sqrt{2}} = \frac{\pi}{4}.$$

6.6 The Riemann integral of functions of several variables

Integration in $\mathbb{R}^n$ is completely analogue to the procedure we described in Section 6.3 for $\mathbb{R}^2$ so we will not repeat all the details. We will follow the steps in Section 6.3 closely and first consider integration over rectangles.

A *rectangle in* $\mathbb{R}^n$ is a set of the form

$$Q = \Big\{(x_1, x_2, \ldots, x_n) \in \mathbb{R}^n \;\Big|\; a_i \le x_i \le b_i, \forall i = 1, 2, \ldots, n\Big\}. \tag{6.66}$$

In order to define the integral of a continuous function $f : Q \to \mathbb{R}$, we will first split the rectangle Q into subrectangles. Precisely as in Section 6.3, this is done by splitting each of the intervals $[a_i, b_i]$ into J intervals of equal length $\Delta x_i = \frac{b_i - a_i}{J}$ and forming the corresponding subrectangles. Recall that in the case of $\mathbb{R}^2$ the procedure gave us subrectangles $Q_{i,j}, i, j = 1, \ldots, J$, that is, we were dealing with J^2 subrectangles. In the case of $\mathbb{R}^n$, we obtain J^n subrectangles; in order to avoid a cumbersome multi-indexing, we will simply denote them by $Q_i, i = 1, 2, \ldots, J^n$. The n-dimensional volume of each of these rectangles $Q_i, i = 1, 2, \ldots, J^n$, is $\Delta x_1 \Delta x_2 \cdots \Delta x_n$. For each of these subrectangles $Q_i \subset \mathbb{R}^n$, we *define* its n-dimensional volume as the product of its side lengths

$$\mathrm{vol}_n(Q_i) \coloneqq \Delta x_1 \Delta x_2 \cdots \Delta x_n,$$

and we pick a point $\boldsymbol{\xi}_i \in Q_i$. Similarly to what we did in (6.31), the corresponding *Riemann sum* S_J is then defined by

$$S_J = \sum_{i=1}^{J^n} f(\boldsymbol{\xi}_i)\,\mathrm{vol}_n(Q_i) \tag{6.67}$$

as $J \to \infty$, the Riemann sum S_J tends to a limit, which we denote by

$$\int_Q f(x_1, x_2, \ldots, x_n)\,\mathrm{d}(x_1, x_2 \ldots, x_n) \tag{6.68}$$

Similar to Theorem 6.3.1, integration over a rectangle in $\mathbb{R}^n$ can be reduced to n consecutive integrations over bounded intervals in $\mathbb{R}$.

Theorem 6.6.1 Integration over a rectangle in $\mathbb{R}^n$
Consider a rectangle Q in $\mathbb{R}^n$ of the form (6.66), and let $f : Q \to \mathbb{R}$ denote a continuous function. Then

$$\begin{aligned}&\int_Q f(x_1, x_2, \dots, x_n)\, \mathrm{d}(x_1, x_2, \dots, x_n)\\ &= \int_{a_n}^{b_n} \cdots \left(\int_{a_2}^{b_2} \left(\int_{a_1}^{b_1} f(x_1, x_2, \dots x_n)\, \mathrm{d}x_1 \right) \mathrm{d}x_2 \right) \cdots \mathrm{d}x_n. \end{aligned} \tag{6.69}$$

Calculations using (6.69) are always done starting from *inside*. We first calculate the integral $\int_{a_1}^{b_1} f(x_1, x_2, \dots x_n)\, \mathrm{d}x_1$, then $\int_{a_2}^{b_2} \left(\int_{a_1}^{b_1} f(x_1, x_2, \dots x_n)\, \mathrm{d}x_1 \right) \mathrm{d}x_2$, and so on.

The Riemann integral is extended to more general subsets B in $\mathbb{R}^n$ precisely as we did for $\mathbb{R}^2$ on page 163. To be more precise, we will assume that

(I) B is bounded, i.e.,

$$B \subset [a_1, b_1] \times [a_2, b_2] \times \cdots \times [a_n, b_n] \tag{6.70}$$

for some $a_i, b_i, \in \mathbb{R}, i = 1, \dots, n$.

(II) The boundary ∂B of B is *smooth*, i.e., it is formed as the union of the graphs for a finite number of continuously differentiable functions from $\mathbb{R}^{n-1}$ to $\mathbb{R}$.

Fix again an integer $J \in \mathbb{N}$, split each of the intervals $[a_1, b_1], [a_2, b_2], \dots, [a_n, b_n]$ into J intervals of equal length $\Delta x_i = \frac{b_i - a_i}{J}$, and form the corresponding subrectangles $Q_i \subset \mathbb{R}^n, i = 1, \dots, \dots, J^n$. Among the rectangles Q_i, we will *exclusively consider* the ones that are completely contained in B. For these sets Q_i, we pick arbitrary points $\boldsymbol{\xi}_i \in Q_i$, and form the corresponding *Riemann sum*

$$S_J = \sum_{\{i | Q_i \subset B\}} f(\boldsymbol{\xi}_i)\, \mathrm{vol}_n(Q_i). \tag{6.71}$$

where the notation $\sum_{\{i|Q_i \subset B\}}$ means that we only include the ith term in the sum if Q_i is contained in B.

One can again show that if the function f is continuous, the Riemann sums S_J tend to a limit as $J \to \infty$. The limit is denoted by

$$\int_B f(x_1, x_2, \dots, x_n)\, \mathrm{d}(x_1, x_2 \dots, x_n), \quad \int_B f(x_1, x_2, \dots, x_n)\, \mathrm{d}\boldsymbol{x}, \tag{6.72}$$

or by

$$\int_B f(x_1, x_2, \ldots, x_n)\, \mathrm{d}X, \tag{6.73}$$

and it is called the *Riemann integral of f over the subset B*. Note that the Riemann integral of f over the subset B is a *real scalar*, and we will sometimes use the compact notation $\int_B f(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x}$.

Similar to the definition of *area* in $\mathbb{R}^n$, we define the *n-dimensional volume* of $B \subset \mathbb{R}^n$ as

$$\mathrm{vol}_n(B) = \int_B 1\, \mathrm{d}X.$$

Exercise 6.6.1

A parallelotope P in $\mathbb{R}^n$ "spanned" by n linearly independent vectors $\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_n \in \mathbb{R}^n$ is defined by:

$$P = \{\boldsymbol{y} \in \mathbb{R}^n \mid \boldsymbol{y} = A\boldsymbol{x}, \quad \text{where } x_i \in [0,1] \text{ for } i = 1, 2, \ldots, n\},$$

where A is the $n \times n$ matrix whose ith column is $\boldsymbol{a}_i$. The set of points P can also be written as $P = A([0,1]^n)$. The n-dimensional volume of P is given by the formula:

$$\mathrm{vol}_n(P) = |\det(A)|. \tag{6.74}$$

We want to prove this fact *without* using tools from integration theory.

(a) Give a proof of (6.74) in the case $n = 2$.

(b) Give a proof of (6.74) in the case $n = 3$.

(c) Give a proof of (6.74) in the general case $n \in \mathbb{N}$ (Note: this is a difficult exercise).

Finally, the definition of a (not necessarily continuous) function $f : B \to \mathbb{R}$, $B \subset \mathbb{R}^n$, being *Riemann integrable* is similar to how we did in Definition 6.3.1 on page 164, and we leave the details to the reader.

The technique for changing variables is a straightforward generalization of the result we saw for $\mathbb{R}^2$ in Theorem 6.4.1.

Theorem 6.6.2 Change of variables in $\mathbb{R}^n$

Let Γ be a bounded subset in $\mathbb{R}^n$ having a smooth boundary. Suppose that there exists a C^1 vector function $\boldsymbol{r} : \Gamma \to \mathbb{R}^n$ which is injective on the interior Γ° and such that

$$\det\left(\boldsymbol{J}_{\boldsymbol{r}}(\boldsymbol{u})\right) \neq 0 \text{ for all } \boldsymbol{u} = (u_1, \ldots, u_n) \in \Gamma^\circ. \tag{6.75}$$

Then, for any continuous function $f : \boldsymbol{r}(\Gamma) \to \mathbb{R}$,

$$\int_{\boldsymbol{r}(\Gamma)} f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \int_{\Gamma} f(\boldsymbol{r}(\boldsymbol{u}))|\det\left(\boldsymbol{J}_{\boldsymbol{r}}(\boldsymbol{u})\right)|\,\mathrm{d}\boldsymbol{u}, \tag{6.76}$$

that is,

$$\int_{\boldsymbol{r}(\Gamma)} f(x_1, x_2, \ldots, x_n)\,\mathrm{d}(x_1, x_2, \ldots, x_n) = \int_{\Gamma} f(\boldsymbol{r}(u_1, \ldots, u_n))|\det\left(\boldsymbol{J}_{\boldsymbol{r}}(u_1, \ldots, u_n)\right)|\,\mathrm{d}(u_1, u_2, \ldots u_n).$$

Proof. A full proof of Theorem 6.6.2 is beyond the scope of this book, but it is useful to do a heuristic argument in order to see where the factor $|\det\left(\boldsymbol{J}_{\boldsymbol{r}}(\boldsymbol{u})\right)|$ comes from. It is clear from Example 6.4.3 on page 170 that some "Jacobian factor" is needed.

Let $r_1, r_2, \ldots, r_k : \Gamma \to \mathbb{R}$ be the coordinate functions, so that $\boldsymbol{r}(\boldsymbol{u}) = (r_1(\boldsymbol{u}), \ldots, r_k(\boldsymbol{u}))$. Let $\boldsymbol{u} \in \Gamma$ be fixed. Since $\boldsymbol{r}$ is continuously differentiable, it follows, e.g., using Taylor's formula, that

$$\boldsymbol{r}(\boldsymbol{u} + \boldsymbol{h}) = \boldsymbol{r}(\boldsymbol{u}) + \boldsymbol{J}_{\boldsymbol{r}}(\boldsymbol{u})\,\boldsymbol{h} + \boldsymbol{\varepsilon}(\boldsymbol{h})\|\boldsymbol{h}\|,$$

for $\|\boldsymbol{h}\| < \delta$, where δ is chosen so small that $\boldsymbol{u} + \boldsymbol{h} \in \Gamma^{\circ}$. For brevity we write the Jacobian matrix at $\boldsymbol{u}$ as $\boldsymbol{J} = \boldsymbol{J}_{\boldsymbol{r}}(\boldsymbol{u})$.

Let $\epsilon_i > 0, i = 1, 2, \ldots, n$, be given (small) real scalars. The rectangle, depending on $\boldsymbol{u}$ and $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \ldots, \epsilon_n)$,

$$\begin{aligned} Q &= [u_1, u_1 + \epsilon_1] \times [u_2, u_2 + \epsilon_2] \times \cdots [u_n, u_n + \epsilon_n] \\ &= \{\boldsymbol{h} \mid u_i \le h_i \le u_i + \epsilon_i \text{ for } i = 1, 2, \ldots, n\} \end{aligned}$$

is under $\boldsymbol{r}$ mapped to $\boldsymbol{r}(Q)$. By Taylor's formula this rectangle is approximately mapped to

$$\{\boldsymbol{r}(\boldsymbol{u}) + \boldsymbol{J}\,\boldsymbol{h} \mid 0 \le h_1 \le \epsilon_1, \ldots, 0 \le h_n \le \epsilon_n\},$$

which can also be written as $\boldsymbol{r}(\boldsymbol{u}) + \boldsymbol{J}([0, \epsilon_1] \times \cdots \times [0, \epsilon_n])$. However, $\boldsymbol{J}([0, \epsilon_1] \times \cdots \times [0, \epsilon_n])$ is a parallelotope given by $A([0, 1]^n)$, where A is the matrix whose ith column is ϵ_i times the ith column of $\boldsymbol{J} = \boldsymbol{J}_{\boldsymbol{r}}(\boldsymbol{u})$.

The n-dimensional volume of the parallelotope $\boldsymbol{J}([0, \epsilon_1] \times \cdots \times [0, \epsilon_n])$ is

$$\epsilon_1 \cdot \epsilon_2 \cdots \epsilon_n \cdot |\det\left(\boldsymbol{J}_{\boldsymbol{r}}(\boldsymbol{u})\right)|$$

while the n-dimensional volume of the rectangle Q is

$$\mathrm{vol}_n(Q) = \epsilon_1 \cdot \epsilon_2 \cdots \epsilon_n$$

Intuitively, we thus have that

$$\frac{\mathrm{vol}_n\left(\boldsymbol{J}([0,\epsilon_1]\times\cdots\times[0,\epsilon_n])\right)}{\mathrm{vol}_n(Q)} \to \left|\det\left(\boldsymbol{J}_{\boldsymbol{r}}(\boldsymbol{u})\right)\right| \quad \text{as } \boldsymbol{\epsilon}\to\mathbf{0},$$

i.e., $\left|\det\left(\boldsymbol{J}_{\boldsymbol{r}}(\boldsymbol{u})\right)\right|$ is the local expansion factor at the point $\boldsymbol{u}$ under the transformation $\boldsymbol{u}\mapsto\boldsymbol{x}=\boldsymbol{r}(\boldsymbol{u})$.

In other words, the "infinitesimal" rectangle $[u_1,u_1+\Delta u_1]\times\cdots\times[u_n,u_n+\Delta u_n]$ is a rectangle with side lengths $\Delta u_1,\ldots,\Delta u_n$, and the image of this "infinitesimal" rectangle under $\boldsymbol{r}$ is a parallelotope with volume $\left|\det\left(\boldsymbol{J}_{\boldsymbol{r}}(\boldsymbol{u})\right)\right|\Delta u_1\cdots\Delta u_n$. Hence, if we consider a Riemann sum (6.71) for $f\circ\boldsymbol{r}$ over $\boldsymbol{u}\in\Gamma$, we need to replace $f(\boldsymbol{\xi}_i)\,\mathrm{vol}_n(Q_i)$ with the corrected volume

$$f(\boldsymbol{r}(\boldsymbol{\gamma}_i))\,\mathrm{vol}_n(Q_i)\left|\det\left(\boldsymbol{J}_{\boldsymbol{r}}(\boldsymbol{u}_i)\right)\right|,$$

where $\boldsymbol{\xi}_i=\boldsymbol{r}(\boldsymbol{\gamma}_i)$ and $\boldsymbol{u}_i\in Q_i$.

■

Remark 6.6.1 Jacobian

The "Jacobian factor" $\left|\det\left(\boldsymbol{J}_{\boldsymbol{r}}(\boldsymbol{u})\right)\right|$ appearing in Equation (6.76) is often simply called *the Jacobian* of the change of variables. When using this terminology, one has to be careful not to confuse the Jacobian with the Jacobian matrix and the Jacobian determinant.

Exercise 6.6.2

Consider again the parallelotope from Exercise 6.6.1. Show that the formula $\mathrm{vol}_n(P)=|\det(A)|$ follows immediately from Theorem 6.6.2.

Let us end this section with a few examples of the important special case of $\mathbb{R}^3$.

Example 6.6.3 Coordinate systems in three-dimensional space

There are three commonly used coordinate systems in $\mathbb{R}^3$: the well-known Cartesian coordinates (x,y,z), the cylindrical coordinates (ρ,ϕ,z), and the spherical coordinates (r,θ,ϕ). See Figure 6.5 on the following page for an illustration.

In cylindrical coordinates, the position of a point in three-dimensional space is described by the parameters (ρ,ϕ,z), where ρ is the radial distance from the z-axis, ϕ is the so-called azimuthal angle measured in the xy-plane from the positive x-axis (similar to θ in polar coordinates), and z is the height above the xy-plane. The parametrization of a point in space using

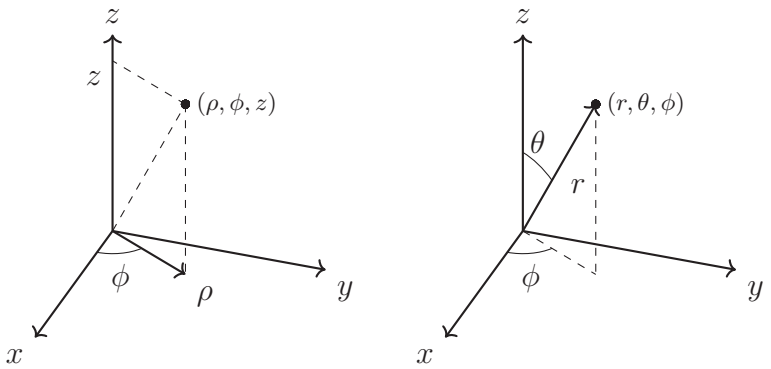


Figure 6.5: Illustration of cylindrical and spherical coordinates of a point in $\mathbb{R}^3$.

cylindrical coordinates can be expressed as the vector function

$$\boldsymbol{c}(\rho, \phi, z) = \begin{bmatrix} \rho\cos(\phi) \\ \rho\sin(\phi) \\ z \end{bmatrix}, \tag{6.77}$$

where $\rho \geq 0$, $\phi \in [0, 2\pi[$ and $z \in \mathbb{R}$. The Jacobian determinant of the transformation from cylindrical to Cartesian coordinates is given by the formula $\boldsymbol{J_c}(\rho, \phi, z) = \rho$.

In spherical coordinates, the position of a point is given by three parameters (r, θ, ϕ), where r is the radial distance from the origin, θ is the polar angle measured from the positive z-axis, and ϕ is the azimuthal angle in the xy-plane from the positive x-axis. The parametrization of a point in space using spherical coordinates can be expressed as the vector function

$$\boldsymbol{s}(r, \theta, \phi) = \begin{bmatrix} r\sin(\theta)\cos(\phi) \\ r\sin(\theta)\sin(\phi) \\ r\cos(\theta) \end{bmatrix}, \tag{6.78}$$

where $r \geq 0$, $\theta \in [0, \pi]$ and $\phi \in [0, 2\pi[$. The Jacobian determinant for the transformation from spherical to Cartesian coordinates is $\boldsymbol{J_s}(r, \theta, \phi) = r^2\sin(\theta)$.

Example 6.6.4 Volume of a ball

Let us compute the volume of a closed[a] ball B in $\mathbb{R}^3$ of radius $R > 0$. We

need to compute

$$\mathrm{vol}_3(B) = \int_B 1 \,\mathrm{d}(x, y, z).$$

A point belongs to B if, in spherical coordinates (r, θ, ϕ), the radial distance r is smaller than or equal to R, or, in other words, if $r \in [0, R]$, $\theta \in [0, \pi]$ and $\phi \in [0, 2\pi[$. Thus, we let $\Gamma = [0, R] \times [0, \pi] \times [0, 2\pi[$ so that $B = \boldsymbol{s}(\Gamma)$, where $\boldsymbol{s}$ is given by (6.78). On Γ°, the function $\boldsymbol{s}$ is injective and its Jacobian determinant is positive. Hence, by Theorem 6.6.2 on page 178, it follows

$$\begin{aligned}
\mathrm{vol}_n(B) &= \int_{\boldsymbol{s}(\Gamma)} 1 \,\mathrm{d}(x, y, z) = \int_\Gamma \det\left(\boldsymbol{J}_{\boldsymbol{s}}(r, \theta, \phi)\right) \mathrm{d}(r, \theta, \phi) \\
&= \int_0^{2\pi} \int_0^{\pi} \int_0^{R} r^2 \sin(\theta) \,\mathrm{d}r \,\mathrm{d}\theta \,\mathrm{d}\phi = \int_0^{2\pi} \int_0^{\pi} \left[\tfrac{1}{3} r^3\right]_0^R \sin(\theta) \,\mathrm{d}\theta \,\mathrm{d}\phi \\
&= \frac{1}{3} R^3 \int_0^{2\pi} \left[-\cos(\theta)\right]_0^{\pi} \mathrm{d}\phi = \frac{1}{3} R^3 \left(-1 - (-1)\right) \int_0^{2\pi} \mathrm{d}\phi \\
&= \frac{1}{3} R^3 \cdot 2 \cdot 2\pi = \frac{4\pi}{3} R^3.
\end{aligned}$$

[a] Whether the ball is open or closed is not essential for the computation and the result is the same.

6.7 The Riemann integral of vector functions

We now turn to the case of integrating vector functions of several variables. The situation is simple, in the sense that, we will just integrate each coordinate of the vector function separately using the techniques we have just developed in Section 6.6 on page 176.

Let us consider a vector function $\boldsymbol{f} = (f_1, \ldots, f_k) : \mathrm{dom}(\boldsymbol{f}) \to \mathbb{R}^k$. Given a subset $B \subseteq \mathrm{dom}(\boldsymbol{f})$ in $\mathbb{R}^n$ that satisfies our standard assumptions Items (I) and (II) on page 177, we define the Riemann integral of $\boldsymbol{f}$ over B

$$\int_B \boldsymbol{f}(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x} \tag{6.79}$$

as a vector in $\mathbb{R}^k$ whose ith coordinate is $\int_B f_i(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x}$ for $i = 1, \ldots, k$.

Example 6.7.1 Center of mass

Let $f : B \to \mathbb{R}$ be a continuous function on a bounded, connected set $B \subset \mathbb{R}^n$ describing the density ($\mathrm{kg\,m}^{-n}$). In other words, we consider a "solid" $B \subset \mathbb{R}^n$, whose mass distribution is continuous with density $f(\boldsymbol{x})$ for $\boldsymbol{x} \in B$. Typically $n = 3$, but we can easily consider the general case.

The total mass M [kg] of B is given by

$$M = \int_B f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}. \tag{6.80}$$

Let us argue that the unit of M is indeed kg. The integral is a limit of Riemann sums S_J, and S_J is a sum of terms of the form $f(\xi_i)\,\mathrm{vol}_n(Q_i)$. Since $f(\xi_i)$ and $\mathrm{vol}_n(Q_i)$ have units $\mathrm{kg\,m^{-n}}$ and m^n, respectively, the Riemann sum and, therefore also, the integral in (6.80) have unit kg. The integral as a limiting process of Riemann sums also explains *why* the mass M is given by (6.80).

By similar considerations, the center of mass $\boldsymbol{x}^{\mathrm{CM}} \in \mathbb{R}^n$ is the integral of the vector-valued function $\boldsymbol{x} \mapsto \boldsymbol{x} f(\boldsymbol{x})$, $B \to \mathbb{R}^n$:

$$\boldsymbol{x}^{\mathrm{CM}} = \frac{1}{M} \int_B \boldsymbol{x} f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}, \tag{6.81}$$

i.e.,

$$x_i^{\mathrm{CM}} = \frac{1}{M} \int_B x_i f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$$

for each $i = 1, \ldots, n$. The unit of each coordinate x_i^{CM} is m.

CHAPTER 7

Vector Fields

Let U be an open subset of $\mathbb{R}^n$. A *vector* field is a vector function $\boldsymbol{f} : U \to \mathbb{R}^k$, where $k = n$, i.e., where the dimension of the domain and codomain are equal. We have already seen vector fields several times, cf. Example 1.4.3 on page 28. Vector fields are often denoted $\boldsymbol{V}$ in place of $\boldsymbol{f}$, and we will follow this tradition. When dealing with vector fields, we often think of $\boldsymbol{V}$ as a *field* of vectors on U, i.e., we attach the vector $\boldsymbol{V}(\boldsymbol{x})$ to the point $\boldsymbol{x} \in U$, see Figure 1.5 on page 29.

The vectors could represent the velocity of a fluid, the inner forces or strains in a beam, or the electric field around an antenna. Since a vector field is a special case of a vector function, we can integrate them over subsets of $\mathbb{R}^n$ as outlined in Section 6.7. However, in applications, we are often interested in another type of integration. Let us give a two examples of such integrals.

If $\boldsymbol{V} : U \to \mathbb{R}^3$ is a force field, we may want to calculate the work done by the force. Recall that work is done when a force acts on something that undergoes a displacement from one position to another. This could be a displacement along a space curve $\boldsymbol{r} : [a, b] \to \mathbb{R}^3$ connecting the starting point $\boldsymbol{r}(a) \in \mathbb{R}^3$ to the end-point $\boldsymbol{r}(b) \in \mathbb{R}^3$. The curve could be, e.g., the helix considered in (1.10) on Page 21. To calculate the work done, we need to integrate the "length" of the force vectors in the direction of the tangent vector of the curve over the entire curve from the start-point $\boldsymbol{r}(a) \in \mathbb{R}^3$ to the end-point $\boldsymbol{r}(b) \in \mathbb{R}^3$.

If $\boldsymbol{V} : U \to \mathbb{R}^3$ represents the velocity of a fluid, we may want to calculate the flow rate, i.e., the flowing per unit of time through a surface $\mathcal{F}$ in $\mathbb{R}^3$. In this case, the function we are integrating is the length of the velocity vectors in the direction of the *normal* vector to the surface, i.e., in the direction of the orthogonal complement of the tangent plane. A surface could be the upper half of the unit sphere; cf. the parametrization considered in (1.11)

on Page 21 with $u \in [0, 2\pi[$ and $v \in [0, \pi/2]$.

In both these cases we need to integrate scalar quantities $f : U \to \mathbb{R}$, $U \subseteq \mathbb{R}^3$, over "thin" subsets of U such as curves or surfaces in $\mathbb{R}^3$. However, if we use the integration techniques from the previous chapter, we do not get a useful outcome. E.g., let $f(x, y, z) = 1$ be the constant function for $(x, y, z) \in \mathbb{R}^3$. The Riemann integral of f over the upper half of the unit sphere $\mathcal{F}$ in $\mathbb{R}^3$ is

$$\int_{\mathcal{F}} f(x, y, z)\, \mathrm{d}X = \int_{\mathcal{F}} 1\, \mathrm{d}X = \mathrm{vol}_3(\mathcal{F}) = 0$$

since the 3-dimensional volume of the sphere is zero. A sphere of radius $r > 0$ has *surface area* $4\pi r^2$, and the desired outcome of the integration of $f = 1$ over $\mathcal{F}$, where $r = 1$, is indeed $\frac{1}{2}4\pi = 2\pi$.

However, no matter which continuous function we are trying to integrate, the Riemann integral in $\mathbb{R}^n$ will give 0 if we integrate over "thin" subsets such as the surface considered above. The aim of this chapter is to develop a useful integration over "thin" subsets, and to develop the notion of curve (or line, as we will call it) and surface integrals of vector fields.

7.1 Parametric curves and surfaces

Our starting point when considering curves and surfaces in $\mathbb{R}^n$ will always be a parametrization. So throughout this chapter we let $\boldsymbol{r} : \Gamma \to \mathbb{R}^n$ be a continuous vector function, where Γ is a connected subset of $\mathbb{R}^m$ ($m \leq n$) that satisfies Items (I) and (II) on page 177. Usually, we will take Γ to be a rectangle of the form

$$\begin{aligned}\Gamma &= [a_1, b_1] \times [a_2, b_2] \times \cdots \times [a_m, b_m] \\ &= \{\boldsymbol{u} \in \mathbb{R}^m \mid a_j \leq u_j \leq b_j \ \forall j = 1, 2, \ldots, m\},\end{aligned}$$

where $a_j < b_j$ for $j = 1, \ldots, m$.

The number of parameters in the parametrization $\boldsymbol{r}$ is given by m. For curves we have $m = 1$ (and $n \geq 2$), and we then often take Γ to be a bounded and closed interval, i.e., $\Gamma = [a, b]$. For surfaces we have $m = 2$ (and $n \geq 3$), and we then usually take Γ to be a bounded and closed rectangle, i.e., $\Gamma = [a_1, b_1] \times [a_2, b_2]$. A *closed curve* is a curve $\boldsymbol{r}([a, b])$ with no "endpoints", i.e., $\boldsymbol{r}(a) = \boldsymbol{r}(b)$. It is called a simple closed curve, if the curve does not intersect itself, i.e., if $\boldsymbol{r}$ is injective on the interior of Γ, e.g., on $]a, b[$. A closed surface is a surface with no "boundary points". It is beyond the scope of this text to precisely define what it means for a surface to have no

"boundary points", but the intuitive meaning should be clear: the boundary $\delta\mathcal{F}$ of a surface $\mathcal{F}$ is where the surface stops. Let us consider some examples.

Example 7.1.1 Closed surfaces

The sphere with center $\boldsymbol{a} \in \mathbb{R}^3$ and radius $r \geq 0$:

$$\mathcal{F}_1 = \{\boldsymbol{x} \in \mathbb{R}^3 \mid \|\boldsymbol{x} - \boldsymbol{a}\| = r\}$$

is a closed surface with no "boundary points", i.e., $\delta\mathcal{F}_1 = \emptyset$. The same holds for the torus. Note the potential confusion of terminology here. According to Definition 2.2.4 on page 51 the sphere is equal to its boundary, i.e., $\mathcal{F}_1 = \partial\mathcal{F}_1$. We use $\delta\mathcal{F}_1$ to denote the boundary of the surface, while $\partial\mathcal{F}_1$ denotes the boundary of $\mathcal{F}_1$ as a subset of $\mathbb{R}^n$.

The half sphere given by

$$\mathcal{F}_2 = \{\boldsymbol{x} \in \mathbb{R}^3 \mid \|\boldsymbol{x}\| = r \ \wedge \ x_3 \geq 0\}$$

has "surface boundary" given by

$$\delta\mathcal{F}_2 = \{\boldsymbol{x} \in \mathbb{R}^3 \mid x_1^2 + x_2^2 = r \ \wedge \ x_3 = 0\}.$$

Since $\mathcal{F}_2$ has a boundary $\delta\mathcal{F}_2 \neq \emptyset$, it is not a closed surface.

Let $h_1, h_2 \in \mathbb{R}$. A hollow cylinder surface without top and buttom given by

$$\mathcal{F}_3 = \{\boldsymbol{x} \in \mathbb{R}^3 \mid x_1^2 + x_2^2 = r \ \wedge x_3 \in [h_1, h_2]\}$$

has boundary

$$\delta\mathcal{F}_3 = \{\boldsymbol{x} \in \mathbb{R}^3 \mid x_1^2 + x_2^2 = r \ \wedge \ (x_3 = h_1 \ \vee \ x_3 = h_2)\}.$$

Definition 7.1.1 Regular parametrization

Let $m \leq n$, and let $\boldsymbol{r} : \Gamma \to \mathbb{R}^n$, $\Gamma \subset \mathbb{R}^m$, be a C^1 vector function on an open set containing Γ, i.e., all partial derivatives of the coordinate functions $\boldsymbol{r} = (r_1, r_2, \ldots, r_m)$ exist and are continuous. The *Jacobian* (also called the *geometric tensor*) of $\boldsymbol{r}$ is defined to be the function given by:

$$\Gamma \to \mathbb{R}, \ \boldsymbol{u} \mapsto \sqrt{\det(\boldsymbol{J}^T\boldsymbol{J})}, \tag{7.1}$$

where $\boldsymbol{J} \coloneqq \boldsymbol{J}_{\boldsymbol{r}}(\boldsymbol{u}) \in \mathsf{M}_{n\times m}(\mathbb{R})$ is the Jacobian matrix of $\boldsymbol{r}$ at $\boldsymbol{u} \in \Gamma$. If

$$\det\Big((\boldsymbol{J}_{\boldsymbol{r}}(\boldsymbol{u}))^T\boldsymbol{J}_{\boldsymbol{r}}(\boldsymbol{u})\Big) \neq 0 \tag{7.2}$$

for all $\boldsymbol{u} \in \Gamma^\circ$, the parametrization $\boldsymbol{r}$ is said to be *regular*, and $\boldsymbol{r}(\Gamma)$ is said to be a *smooth* m-fold (a 1-fold is a curve and a 2-fold is a surface) .

Let us study the matrix product $\boldsymbol{G} \coloneqq \boldsymbol{J}^T\boldsymbol{J}$ in (7.1) in more detail. Recall that the ith row, $i = 1, 2, \ldots, n$, of the Jacobian matrix $\boldsymbol{J}_{\boldsymbol{r}}(\boldsymbol{u}) \in \mathbb{R}^{n\times m}$ is the (transpose of the) gradient vector of the jth coordinate function, i.e., $\Big(\nabla r_j(\boldsymbol{u})\Big)^T \in \mathbb{R}^m$. On the other hand, we can also consider the columns of the Jacobian matrix. The jth column, $j = 1, \ldots, m$, of $\boldsymbol{J}_{\boldsymbol{r}}(\boldsymbol{u})$ at $\boldsymbol{u} = (u_1, u_2, \ldots, u_m) \in \Gamma$ is the *tangent vector* $\frac{\partial \boldsymbol{r}}{\partial u_j}(\boldsymbol{u})$, also denoted by

$$\boldsymbol{r}'_{u_j}(\boldsymbol{u}) \coloneqq \left(\frac{\partial r_1}{\partial u_j}(\boldsymbol{u}), \frac{\partial r_2}{\partial u_j}(\boldsymbol{u}), \ldots, \frac{\partial r_n}{\partial u_j}(\boldsymbol{u})\right) \in \mathbb{R}^n.$$

Hence, the (i, j) entry of $\boldsymbol{G} = \boldsymbol{J}^T\boldsymbol{J} \in \mathbb{R}^{m\times m}$ is the inner product[1] $\langle \boldsymbol{r}'_{u_i}, \boldsymbol{r}'_{u_j}\rangle = (\boldsymbol{r}'_{u_j})^T\boldsymbol{r}'_{u_i}$. Such matrices are often called a Gram matrix or the *Gramian* of the tangent vectors. One can show that the condition (7.2) is equivalent with $\boldsymbol{J}_{\boldsymbol{r}}(\boldsymbol{u})$ being of rank m for every $\boldsymbol{u} \in \Gamma^\circ$, i.e., the tangent vectors $\boldsymbol{r}'_{u_j}$, $j = 1, 2, \ldots, m$, being linearly independent and, hence, the tangent vectors span the tangent plane of $\boldsymbol{r}$ at $\boldsymbol{u}$.

From Example 2.9.1 on page 76 we have that the $\boldsymbol{G}$ is a positive semi-definite matrix, and therefore $\det(\boldsymbol{G}) = \lambda_1 \cdots \lambda_n \geq 0$. Hence, the condition in (7.2) simply says that $\boldsymbol{G}$ is positive definite, i.e., the eigenvalues of $\boldsymbol{G}$ cannot become zero.

We introduced the Jacobian in Remark 6.6.1 on page 180 in Chapter 6 for parametrizations where $n = m$. We will show in Lemma 7.1.1 that this form of the Jacobian is identical to the Jacobian introduced in Definition 7.1.1, that is, there is *one* expression of the Jacobian "factor", namely (7.1), that can be used in all cases. However, it is often cumbersome to compute the Jacobian by the formula $\Big(\det((\boldsymbol{J}_{\boldsymbol{r}}(\boldsymbol{u}))^T\boldsymbol{J}_{\boldsymbol{r}}(\boldsymbol{u}))\Big)^{1/2}$. Hence, in Lemma 7.1.1, we will also give explicit formulas in the most common situations. For this formulation we need to introduce the *cross product* of two vectors $\boldsymbol{x}, \boldsymbol{y}$ in $\mathbb{R}^3$:

$$\begin{aligned}\boldsymbol{x} \times \boldsymbol{y} &= \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \times \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} \\ &= \begin{bmatrix} x_2y_3 - x_3y_2 \\ x_3y_1 - x_1y_3 \\ x_1y_2 - x_2y_1 \end{bmatrix}. \end{aligned} \tag{7.3}$$

[1] In case of complex-valued vector functions $\boldsymbol{r}$, one needs to replace the transpose $\boldsymbol{J}^T$ by the adjoint $\boldsymbol{J}^*$ in the definition $\boldsymbol{G} \coloneqq \boldsymbol{J}^T\boldsymbol{J}$.

The cross product is not commutative as it satisfies

$$\boldsymbol{x} \times \boldsymbol{y} = -\boldsymbol{y} \times \boldsymbol{x}.$$

The cross product is a mapping from $\mathbb{R}^3 \times \mathbb{R}^3$ to $\mathbb{R}^3$ for which $\boldsymbol{x} \times \boldsymbol{y}$ is orthogonal to $\boldsymbol{x}$ and $\boldsymbol{y}$, that is $\boldsymbol{x} \times \boldsymbol{y} \in \{\boldsymbol{x}, \boldsymbol{y}\}^{\perp}$. The length of the cross product can be computed by Lagrange's identity:

$$\|\boldsymbol{x} \times \boldsymbol{y}\|^2 = \|\boldsymbol{x}\|^2 \|\boldsymbol{y}\|^2 - \langle \boldsymbol{x}, \boldsymbol{y} \rangle^2 \tag{7.4}$$

which, in turn, follows from the Cauchy-Binet formula.

Lemma 7.1.1 Jacobian in special cases

Consider regular parametrizations as in Definition 7.1.1 on page 186.

(i) (curves in $\mathbb{R}^n$). Let $m = 1$ and let $I = \Gamma \subset \mathbb{R}$ be an interval. The tangent vector is $\boldsymbol{r}'(u) = (r_1'(u), r_2'(u), \ldots, r_m'(u))$, and the Jacobian (7.1) is the norm of the tangent vector:

$$\boldsymbol{u} \mapsto \|\boldsymbol{r}'(u)\|. \tag{7.5}$$

(ii) (surfaces in $\mathbb{R}^3$). Let $m = 2$ and let $n = 3$. For any $\boldsymbol{u} = (u_1, u_2) \in \Gamma^{\circ}$ the tangent plane is spanned by the two tangent vectors $\boldsymbol{r}'_{u_1}(\boldsymbol{u})$ and $\boldsymbol{r}'_{u_2}(\boldsymbol{u})$, and the Jacobian (7.1) is the norm of the cross product of the two tangent vectors:

$$\boldsymbol{u} \mapsto \|\boldsymbol{r}'_{u_1}(\boldsymbol{u}) \times \boldsymbol{r}'_{u_2}(\boldsymbol{u})\|. \tag{7.6}$$

(iii) ("Thick" subsets of $\mathbb{R}^n$). Let $m = n$. The Jacobian in (7.1) is identical to the Jacobian introduced in Remark 6.6.1 on page 180, that is, the Jacobian satisfies

$$\sqrt{\det\Big((\boldsymbol{J}_{\boldsymbol{r}}(\boldsymbol{u}))^T \boldsymbol{J}_{\boldsymbol{r}}(\boldsymbol{u})\Big)} = |\det(\boldsymbol{J}_{\boldsymbol{r}}(\boldsymbol{u}))|.$$

Proof. (i): The matrix $\boldsymbol{G} = \boldsymbol{J}^T \boldsymbol{J} \in \mathbb{R}^{1\times 1}$ is a 1×1. By the discussion above, the $(1,1)$ entry of $\boldsymbol{G}$ is the inner product $\langle \boldsymbol{r}'(u), \boldsymbol{r}'(u) \rangle = \|\boldsymbol{r}'(u)\|^2$. The determinant of a 1×1 matrix is just the entry at $(1,1)$ itself. Taking the square root of $\|\boldsymbol{r}'(u)\|^2$ we end up with

$$\sqrt{\det\Big((\boldsymbol{J}_{\boldsymbol{r}}(u))^T \boldsymbol{J}_{\boldsymbol{r}}(u)\Big)} = \|\boldsymbol{r}'(u)\|.$$

(ii): The matrix $\boldsymbol{G} = \boldsymbol{J}^T\boldsymbol{J} \in \mathbb{R}^{2\times 2}$ is in the form of a Gram matrix given by:

$$\boldsymbol{G} = \begin{bmatrix} \langle \boldsymbol{r}'_{u_1}, \boldsymbol{r}'_{u_1}\rangle & \langle \boldsymbol{r}'_{u_1}, \boldsymbol{r}'_{u_2}\rangle \\ \langle \boldsymbol{r}'_{u_2}, \boldsymbol{r}'_{u_1}\rangle & \langle \boldsymbol{r}'_{u_2}, \boldsymbol{r}'_{u_2}\rangle \end{bmatrix},$$

where each tangent vector should be evaluated at $\boldsymbol{u} = (u_1, u_2)$. The determinant of this two-by-two matrix is easily found as:

$$\det(\boldsymbol{G}) = \langle \boldsymbol{r}'_{u_1}, \boldsymbol{r}'_{u_1}\rangle\langle \boldsymbol{r}'_{u_2}, \boldsymbol{r}'_{u_2}\rangle - \langle \boldsymbol{r}'_{u_1}, \boldsymbol{r}'_{u_2}\rangle^2.$$

By Lagrange's identity (7.4) we conclude that

$$\det(\boldsymbol{G}) = \|\boldsymbol{r}'_{u_1}\|^2\|\boldsymbol{r}'_{u_2}\|^2 - \langle \boldsymbol{r}'_{u_1}, \boldsymbol{r}'_{u_2}\rangle^2 = \|\boldsymbol{r}'_{u_1} \times \boldsymbol{r}'_{u_2}\|^2.$$

(iii): Recall that $\det(AB) = \det(A)\det(B)$ and $\det(A) = \det(A^T)$ for square matrices A and B. For brevity, denote $\boldsymbol{J} \coloneqq \boldsymbol{J}_{\boldsymbol{r}}(\boldsymbol{u})$. Then

$$\sqrt{\det\left(\boldsymbol{J}^T\boldsymbol{J}\right)} = \sqrt{\det(\boldsymbol{J})\det(\boldsymbol{J})} = |\det(\boldsymbol{J})|$$

which was what we wanted to show.

■

Note that in this chapter we are interested in the situation of "thin" subsets where $m < n$, hence we will use Items (i) and (ii) from Lemma 7.1.1 in the next sections, while Item (iii) relates to integration as in Chapter 6.

For surfaces as in Lemma 7.1.1, the vector $\boldsymbol{n}_{\mathcal{F}}(\boldsymbol{u}) \coloneqq \boldsymbol{r}'_{u_1}(\boldsymbol{u}) \times \boldsymbol{r}'_{u_2}(\boldsymbol{u})$ is called the *normal vector* of the surface $\mathcal{F} = \boldsymbol{r}(\Gamma)$ at the point $\boldsymbol{r}(\boldsymbol{u}) \in \mathbb{R}^3$. The tangent plane at $\boldsymbol{r}(\boldsymbol{u}) \in \mathbb{R}^3$ can be expressed as

$$\{\boldsymbol{x} \in \mathbb{R}^3 \mid \langle \boldsymbol{x} - \boldsymbol{r}(\boldsymbol{u}), \boldsymbol{n}_{\mathcal{F}}(\boldsymbol{u})\rangle = 0\},$$

i.e., the set of all vectors "from $\boldsymbol{r}(\boldsymbol{u})$ to $\boldsymbol{x}$" orthogonal to the normal vector.

In the remainder of this chapter, we will assume that parametrizations $\boldsymbol{r}$ are always *regular* and *injective* on Γ°. However, we will allow $\boldsymbol{r}$ to be continuous and only *piecewise* C^1. When $m = 1$, this simply means that we can subdivide $[a, b]$ in k pieces $a = u_0 < u_1 \ldots u_k = b$ such that $\boldsymbol{r}$ is C^1 on each interval $[u_{i-1}, u_i]$. Such a curve $\boldsymbol{r}([a, b])$ can be considered as a joining of k curves, each being C^1. For piecewise C^1 curves, the tangent vector is not well-defined at the points u_i. For piecewise C^1 surfaces, the normal vector is not well-defined along curves where two surfaces meet, but this will not be an issue when it comes to defining integration since these problematic subsets are "small".

7.2 Line and surface integrals

It is possible to introduce integration over curves and surface directly from Riemann sums in much the same manner as we did in Chapter 6. However, we will here simply *define* integration of continuous functions over thin sets (m-folds) using the change of variables technique.

Definition 7.2.1 Integral over m-folds

Let $\boldsymbol{r} : \Gamma \to \mathbb{R}^n$, $\Gamma \subset \mathbb{R}^m$, be a regular parametrization that is injective on Γ°. Suppose $f : A \to \mathbb{R}$, $A \subseteq \mathbb{R}^n$ is a scalar, continuous function on the piecewise smooth m-fold $\mathcal{D} \coloneqq \boldsymbol{r}(\Gamma)$. Then the Riemann integral of f over $\mathcal{D}$ is defined by

$$\int_{\mathcal{D}} f(\boldsymbol{x})\,\mathrm{d}S \coloneqq \int_{\Gamma} f(\boldsymbol{r}(\boldsymbol{u}))\sqrt{\det\Big((\boldsymbol{J}_{\boldsymbol{r}}(\boldsymbol{u}))^T\boldsymbol{J}_{\boldsymbol{r}}(\boldsymbol{u})\Big)}\,\mathrm{d}\boldsymbol{u}, \tag{7.7}$$

where the integral on the right-hand-side of (7.7) was introduced in Section 6.6

Note that the Jacobian (7.1) appearing in (7.7) is again an m-dimensional volume correcting factor. The argument for why this factor is correct will not be given here, but we mention that the argument is similar to the proof sketch of Theorem 6.6.2 on page 178, e.g., in case of the surface integral, $m = 2$, one needs to show that the *area* of the parallelogram spanned by two linearly independent vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ in $\mathbb{R}^n$ *on the two-dimensional subspace* $\operatorname{span}\{\boldsymbol{x}, \boldsymbol{y}\}$ of $\mathbb{R}^n$ is given by $\sqrt{\det(J^TJ)}$, where $J = [\boldsymbol{x}\ \boldsymbol{y}] \in \mathbb{R}^{n\times 2}$.

Remark 7.2.1 Re-parametrization

The integral introduced in Definition 7.2.1 is independent of the chosen parametrization of the m-fold. That is, the value of the integral does not change under re-parametrization of the m-fold. A re-parametrization of $\boldsymbol{r} : \Gamma_1 \to \mathbb{R}^n$ is a mapping of the form $\boldsymbol{r} \circ \boldsymbol{p}$, where $\boldsymbol{p}$ is a *bijective* C^1 vector function $\boldsymbol{p} : \Gamma_2 \to \Gamma_1$, whose Jacobian determinant is never zero. Here, as usual, Γ_1 and Γ_2 are connected subsets of $\mathbb{R}^m$ satisfying Items (I) and (II) on page 177. The domain of $\boldsymbol{r} \circ \boldsymbol{p}$ is Γ_2, hence we have changed parameter domain from Γ_1 to Γ_2, while not changing the m-fold itself $\boldsymbol{r}(\Gamma_1) = \boldsymbol{r} \circ \boldsymbol{p}(\Gamma_2)$.

Let us explicitly write out the definition in (7.7) for the cases $m = 1$ and $m = 2$. Thus, we assume the setup used in Definition 7.2.1.

The line integral. Let $m = 1$ and $\Gamma = [a, b]$. The "infinitesimal element" $\mathrm{d}S$ for curves is written $\mathrm{d}s$. Let us denote the curve by $\mathcal{C} \coloneqq \boldsymbol{r}(\Gamma)$. The

integral is called a *line integral* (also called path or curve integral), and the definition in (7.7) becomes

$$\int_{\mathcal{C}} f(\boldsymbol{x})\,\mathrm{d}s = \int_a^b f(\boldsymbol{r}(u))\|\boldsymbol{r}'(u)\|\,\mathrm{d}u. \tag{7.8}$$

The length of the curve $\mathcal{C}$ is found by setting $f(\boldsymbol{x}) = 1$ for all $\boldsymbol{x} \in \mathrm{dom}(f)$:

$$\mathrm{length}(\mathcal{C}) = \int_{\mathcal{C}} 1\,\mathrm{d}s = \int_a^b \|\boldsymbol{r}'(u)\|\,\mathrm{d}u. \tag{7.9}$$

Example 7.2.1 Arc of semi-circle

Let us return to the curve

$$\boldsymbol{r} : [0,\pi] \to \mathbb{R}^2, \quad \boldsymbol{r}(u) = \begin{bmatrix}\cos(u)\\ \sin(u)\end{bmatrix},$$

considered in Example 3.1.5. According to (7.9) the length of the curve $\mathcal{C} = \boldsymbol{r}([0,\pi])$ is

$$\mathrm{length}(\mathcal{C}) = \int_0^\pi \|\boldsymbol{r}'(u)\|\,\mathrm{d}u = \int_0^\pi \left\| \begin{bmatrix}-\sin(u)\\ \cos(u)\end{bmatrix} \right\| \mathrm{d}u = \int_0^\pi 1\,\mathrm{d}u = \pi$$

which is as expected half the circumference of a unit circle.

Example 7.2.2 Length of a helix

Let us return to the curve

$$\boldsymbol{r}(u) = \begin{bmatrix}\cos(u)\\ \sin(u)\\ u\end{bmatrix} \quad \text{for } u \in [0, 2\pi] \tag{7.10}$$

introduced in Equation (1.10) on page 21. The tangent vector is

$$\boldsymbol{r}'(u) = \begin{bmatrix}-\sin(u)\\ \cos(u)\\ 1\end{bmatrix}$$

Hence, the length of the helix $\mathcal{C} = \boldsymbol{r}([0, 2\pi])$ is

$$\begin{aligned}\mathrm{length}(\mathcal{C}) &= \int_0^{2\pi} \|\boldsymbol{r}'(u)\|\,\mathrm{d}u = \int_0^{2\pi} \left((-\sin(u))^2 + (\cos(u))^2 + 1\right)^{1/2} \mathrm{d}u \\ &= \int_0^{2\pi} (1+1)^{1/2}\,\mathrm{d}u = 2\sqrt{2}\pi\end{aligned}$$

which is a factor $\sqrt{2}$ longer than the unit circle in the plane.

The surface integral. Let $m = 2$, $\Gamma \subset \mathbb{R}^2$, and let us denote the surface by $\mathcal{F} := \boldsymbol{r}(\Gamma)$, and denote the vectors in Γ by $\boldsymbol{u} = (u_1, u_2)$. The *surface integral* defined in (7.7) reads explicitly, using (7.6),

$$\int_{\mathcal{F}} f(\boldsymbol{x})\,\mathrm{d}S = \int_{\Gamma} f(\boldsymbol{r}(\boldsymbol{u}))\|\boldsymbol{r}'_{u_1}(\boldsymbol{u}) \times \boldsymbol{r}'_{u_2}(\boldsymbol{u})\|\,\mathrm{d}(u_1, u_2). \tag{7.11}$$

The area of the surface $\mathcal{F}$ is found by setting $f(\boldsymbol{x}) = 1$ for all $\boldsymbol{x} \in \mathrm{dom}(f)$:

$$\mathrm{area}(\mathcal{F}) = \int_{\mathcal{F}} 1\,\mathrm{d}S = \int_{\Gamma} \|\boldsymbol{r}'_{u_1}(\boldsymbol{u}) \times \boldsymbol{r}'_{u_2}(\boldsymbol{u})\|\,\mathrm{d}(u_1, u_2). \tag{7.12}$$

Example 7.2.3

A "parameter" subset $\Gamma \subset \mathbb{R}^2$ can itself be identified with a surface $\mathcal{F}$ in $\mathbb{R}^3$, simply by taking

$$\boldsymbol{r}(u, v) = (u, v, 0), \quad (u, v) \in \Gamma.$$

Then, the tangent vectors are

$$\boldsymbol{r}'_u(u, v) = \frac{\partial \boldsymbol{r}}{\partial u}(u, v) = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \quad \boldsymbol{r}'_v(u, v) = \frac{\partial \boldsymbol{r}}{\partial v}(u, v) = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix},$$

so

$$\frac{\partial \boldsymbol{r}}{\partial u}(u, v) \times \frac{\partial \boldsymbol{r}}{\partial v}(u, v) = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \times \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}.$$

Thus (7.12) yields

$$\mathrm{area}(\mathcal{F}) = \int_{\mathcal{F}} 1 dS = \int_{\Gamma} \left\|\boldsymbol{r}'_u(u, v) \times \boldsymbol{r}'_v(u, v)\right\| \mathrm{d}(u, v) = \int_B 1\,\mathrm{d}(u, v)$$

which coincides with the definition of the area of $\Gamma \subset \mathbb{R}^2$ in Definition 6.3.2 on page 165.

Example 7.2.4 Area of cylinder

With an appropriate positioning of the coordinate system, a hollow cylinder $\mathcal{Z}$ with height h and diameter $2r$ can be described as

$$\mathcal{Z} = \{(x, y, z) \in \mathbb{R}^3 \mid x^2 + y^2 = r^2 \ \wedge \ z \in [0, h]\}. \tag{7.13}$$

The surface $\mathcal{Z}$ can easily be parametrized using cylindrical coordinates,

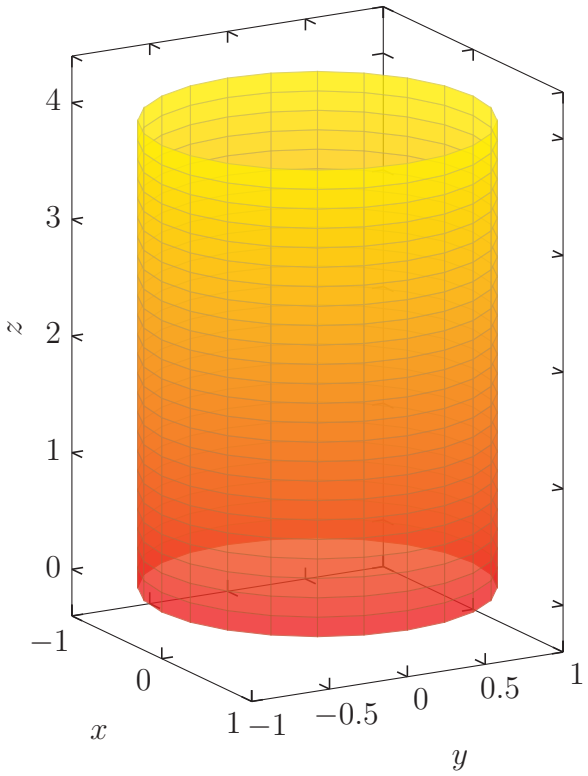


Figure 7.1: The cylinder as in Equation (7.13) with $r = 1$ and $h = 4$.

see Equation (6.77) on page 181, as

$$\boldsymbol{r}(u,v) = \begin{bmatrix} r\cos(u) \\ r\sin(u) \\ v \end{bmatrix}, \quad (u,v) \in \Gamma = [0, 2\pi[\times[0,h]. \tag{7.14}$$

Then

$$\boldsymbol{r}'_u(u,v) = \frac{\partial \boldsymbol{r}}{\partial u}(u,v) = \begin{bmatrix} -r\sin(u) \\ r\cos(u) \\ 0 \end{bmatrix} \quad \text{and} \quad \boldsymbol{r}'_v(u,v) = \frac{\partial \boldsymbol{r}}{\partial v}(u,v) = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix},$$

so

$$\frac{\partial \boldsymbol{r}}{\partial u}(u,v) \times \frac{\partial \boldsymbol{r}}{\partial v}(u,v) = \begin{bmatrix} -r\sin(u) \\ r\cos(u) \\ 0 \end{bmatrix} \times \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} = \begin{bmatrix} r\cos(u) \\ r\sin(u) \\ 0 \end{bmatrix}.$$

Note that the vector $\boldsymbol{n}_{\mathcal{Z}}(u,v) = [r\cos(u), r\sin(u), 0]^T$ is indeed a normal vector to the surface $\mathcal{Z}$. The Jacobian of the parametrization of $\mathcal{Z}$ is the

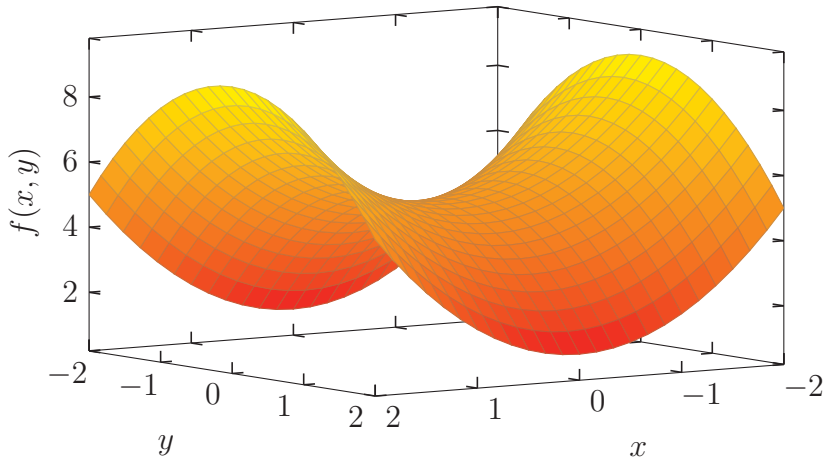


Figure 7.2: The graph surface associated with the function h in (7.17).

norm of the normal vector:

$$\|\boldsymbol{n}_{\mathcal{Z}}(u,v)\| = \|\boldsymbol{r}'_u(u,v) \times \boldsymbol{r}'_v(u,v)\| = \sqrt{(r\cos(u))^2 + (r\sin(u))^2} = r,$$

hence, the area-correcting factor does not depend on the parameters u, v, but is constant over the entire surface.

More importantly, the Jacobian $\|\boldsymbol{n}_{\mathcal{Z}}(u)\|$ is positive, so (7.12) finally yields

$$\begin{aligned}\operatorname{area}(\mathcal{Z}) = \int_{\mathcal{Z}} 1\,\mathrm{d}S &= \int_{\Gamma} \|\boldsymbol{r}'_u(u,v) \times \boldsymbol{r}'_v(u,v)\|\,\mathrm{d}(u,v)\\ &= \int_{\Gamma} r\,\mathrm{d}(u,v)\\ &= \int_0^h \int_0^{2\pi} r\,\mathrm{d}u\,\mathrm{d}v = 2\pi r h.\end{aligned}$$

This coincides with the standard definition of the surface area of a cylinder.

An important class of surfaces occur by considering the graph of a continuous function of two variables.

Definition 7.2.2 Graph surface

Let B be a connected subset of $\mathbb{R}^2$ which satisfies the conditions Items (I) and (II) on page 163. Given a continuous function

$$h : B \to \mathbb{R}, \tag{7.15}$$

a surface of the form

$$\mathcal{S} = \{(u, v, h(u, v)) \in \mathbb{R}^3 \mid (u, v) \in B\} \tag{7.16}$$

is called a *graph surface*.

Figure 7.2 illustrates the concept of a graph surface for the function

$$h : [-2, 2] \times [-2, 2] \to \mathbb{R}, \quad h(x, y) = x^2 - y^2 + 5. \tag{7.17}$$

To compute surface integrals over graph surfaces, we need to specify a Jacobian, cf. Definition 7.1.1 on page 186. So, let us consider a graph surface of $h : B \to \mathbb{R}$ as in Definition 7.2.2. Let $\Gamma = B$ and define the parametrization

$$\boldsymbol{r}(u, v) = [u, v, h(u, v)]^T \quad (u, v) \in \Gamma. \tag{7.18}$$

The tangent vectors are

$$\boldsymbol{r}'_u(u, v) = \begin{bmatrix} 1 \\ 0 \\ \frac{\partial h}{\partial u}(u, v) \end{bmatrix} \quad \text{and} \quad \boldsymbol{r}'_v(u, v) = \begin{bmatrix} 0 \\ 1 \\ \frac{\partial h}{\partial v}(u, v) \end{bmatrix}.$$

The normal vector $\boldsymbol{n}_{\mathcal{S}}(u, v)$ is given as the cross product of the two tangent vectors:

$$\boldsymbol{n}_{\mathcal{S}}(u, v) = \boldsymbol{r}'_u(u, v) \times \boldsymbol{r}'_v(u, v) = \begin{bmatrix} -\frac{\partial h}{\partial u}(u, v) \\ -\frac{\partial h}{\partial v}(u, v) \\ 1 \end{bmatrix}$$

Hence, the Jacobian (7.6) is given by

$$(u, v) \mapsto \|\boldsymbol{n}_{\mathcal{S}}(u, v)\| = \sqrt{\left(\frac{\partial h}{\partial u}(u, v)\right)^2 + \left(\frac{\partial h}{\partial v}(u, v)\right)^2 + 1}.$$

7.3 Vector fields and gradient fields

We already have discussed the notion of vector fields several times before. Recall that this is just vector functions of the form $\boldsymbol{V} : U \to \mathbb{R}^n$, where the domain U is an open subset of $\mathbb{R}^n$. The vector field expressed in terms of its coordinate functions is:

$$\boldsymbol{V}(x_1, x_2, \ldots, x_n) = \begin{bmatrix} V_1(x_1, x_2, \ldots, x_n) \\ V_2(x_1, x_2, \ldots, x_n) \\ \vdots \\ V_n(x_1, x_2, \ldots, x_n) \end{bmatrix}, \quad (x_1, x_2, \ldots, x_n) \in U,$$

where each coordinate function $V_i : U \to \mathbb{R}$, $i = 1, \ldots, n$, is a scalar function of n variables.

Let us give some examples of vector fields.

Example 7.3.1

The function

$$\boldsymbol{V}(x, y) = \begin{bmatrix} x^2 y \\ \cos(x + y) + e^x \end{bmatrix}$$

defines a vector field $\boldsymbol{V} : U \to \mathbb{R}^2$ where the domain is $U = \mathbb{R}^2$. Similarly, the function

$$\boldsymbol{V}(x_1, x_2, x_3) = \begin{bmatrix} x_1 x_3 \\ \sin(x_1 x_2) \\ (x_1 + x_2 + x_3)^3 \end{bmatrix}$$

defines a vector field $\boldsymbol{V} : U \to \mathbb{R}^3$, where $U = \mathbb{R}^3$.

Also, any $n \times n$ matrix with real entries in a natural way corresponds to a vector field $\boldsymbol{V} : \mathbb{R}^n \to \mathbb{R}^n$.

Example 7.3.2

Let A denote an arbitrary $n \times n$ matrix with real entries. Then

$$\boldsymbol{V} : \mathbb{R}^n \to \mathbb{R}^n, \quad \boldsymbol{V}(\boldsymbol{x}) := A\boldsymbol{x} \tag{7.19}$$

defines a so-called *linear vector field*. As a concrete example,

$$\boldsymbol{V} : \mathbb{R}^2 \to \mathbb{R}^2, \quad \boldsymbol{V}(x, y) = \begin{bmatrix} 2x + 4y \\ 3x + 6y \end{bmatrix} \tag{7.20}$$

defines a vector field, and it corresponds to (7.19) with the choice

$$A = \begin{bmatrix} 2 & 4 \\ 3 & 6 \end{bmatrix}.$$

Example 7.3.3

In Example 3.3.4 on page 96 we considered the function

$$f : \mathbb{R}^2 \to \mathbb{R}, \quad f(x, y) = xy^2 + x^2 + y^3,$$

and we computed the gradient vector as

$$\nabla f(x, y) = (y^2 + 2x, 2xy + 3y^2), \quad (x, y) \in \mathbb{R}^2. \tag{7.21}$$

Thus, the gradient ∇f is actually vector field as a function from $\mathbb{R}^2$ to $\mathbb{R}^2$, i.e., $\nabla f : \mathbb{R}^2 \to \mathbb{R}^2$.

Similarly, in Example 3.3.5, we saw that for the function

$$f : \mathbb{R}^3 \to \mathbb{R}, \quad f(x_1, x_2, x_3) = x_1 x_2 + x_2^2 x_3^3,$$

the gradient vector at the point $(x_1, x_2, x_3) \in \mathbb{R}^2$ is

$$\nabla f(x_1, x_2, x_3) = (x_2, x_1 + 2x_2 x_3^3, 3x_2^2 x_3^2);$$

that is, the gradient is a vector field $\nabla f : \mathbb{R}^3 \to \mathbb{R}^3$.

The vector fields in the examples above are examples of very nice vector fields in the sense that they are C^∞ vector functions, that is, vector functions whose coordinate functions are infinitely often differentiable in the entire domain of the vector function.

We have already seen that the gradient of a scalar C^1 function $f : U \to \mathbb{R}$, $U \subseteq \mathbb{R}^n$, gives rise to a vector field $\nabla f : U \to \mathbb{R}^n$. In Theorem 7.4.2 we will see that vector fields arising from a differentiable function f in this way, have certain particularly pleasant properties. This motivates the following definition.

Definition 7.3.1 Gradient field and anti-derivative

A continuous vector field $\boldsymbol{V} : U \to \mathbb{R}^n$, where $U \subseteq \mathbb{R}^n$ is an open set, is called a *gradient field* if there exists a C^1 function $f : U \to \mathbb{R}$ such that $\boldsymbol{V} = \nabla f$, i.e.,

$$\boldsymbol{V}(\boldsymbol{x}) = (V_1(\boldsymbol{x}), \ldots, V_n(\boldsymbol{x})) = \left(\frac{\partial f}{\partial x_1}(\boldsymbol{x}), \ldots, \frac{\partial f}{\partial x_n}(\boldsymbol{x}) \right) \tag{7.22}$$

for all $\boldsymbol{x} \in U$. The function f is called the *anti-derivative* of the vector field $\boldsymbol{V}$.

In physics, gradient fields are also known under the name *conservative fields*, and the anti-derivative is known as the (scalar) *potential*. Note that, in the definition of potentials f of vector fields $\boldsymbol{V}$, it is not uncommon to use the convention $\boldsymbol{V} = -\nabla f$ (the anti-derivative is multiplied by -1).

Let $U \subseteq \mathbb{R}^n$ be an open set, and let $\boldsymbol{V} : U \to \mathbb{R}^n$ be a C^1 vector function. If $\boldsymbol{V}$ is a gradient field, i.e., $\boldsymbol{V} = \nabla f$ for some C^2 function $f : U \to \mathbb{R}$. Note that we assume $\boldsymbol{V}$ is C^1, not only C^0 as in Definition 7.3.1, so we can compute the derivatives of $\boldsymbol{V}$. Since $V_i = \frac{\partial f}{\partial x_i}$ the Jacobian matrix of $\boldsymbol{V}$ is

$$\boldsymbol{J}_{\boldsymbol{V}}(\boldsymbol{x}) = \begin{bmatrix} \frac{\partial V_1}{\partial x_1}(\boldsymbol{x}) & \frac{\partial V_1}{\partial x_2}(\boldsymbol{x}) & \cdots & \frac{\partial V_1}{\partial x_n}(\boldsymbol{x}) \\ \frac{\partial V_2}{\partial x_1}(\boldsymbol{x}) & \frac{\partial V_2}{\partial x_2}(\boldsymbol{x}) & \cdots & \frac{\partial V_2}{\partial x_n}(\boldsymbol{x}) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial V_n}{\partial x_1}(\boldsymbol{x}) & \frac{\partial V_n}{\partial x_2}(\boldsymbol{x}) & \cdots & \frac{\partial V_n}{\partial x_n}(\boldsymbol{x}) \end{bmatrix}$$

$$= \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2}(\boldsymbol{x}) & \frac{\partial^2 f}{\partial x_1 \partial x_2}(\boldsymbol{x}) & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n}(\boldsymbol{x}) \\ \frac{\partial^2 f}{\partial x_2 \partial x_1}(\boldsymbol{x}) & \frac{\partial^2 f}{\partial x_2^2}(\boldsymbol{x}) & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n}(\boldsymbol{x}) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1}(\boldsymbol{x}) & \frac{\partial^2 f}{\partial x_n \partial x_2}(\boldsymbol{x}) & \cdots & \frac{\partial^2 f}{\partial x_n^2}(\boldsymbol{x}) \end{bmatrix} = \boldsymbol{H}_f(\boldsymbol{x}).$$

Hence, for a C^1 gradient field $\boldsymbol{V} = \nabla f$, its Jacobian matrix is the Hessian matrix of the anti-derivative f, i.e., $\boldsymbol{J}_{\boldsymbol{V}}(\boldsymbol{x}) = \boldsymbol{H}_f(\boldsymbol{x})$. Since the Hessian matrix is symmetric for C^2 functions, we conclude that the Jacobian matrix of a gradient field is symmetric. Note that a symmetric Jacobian matrix is a rather restrictive property that even the simplest vector fields need not have.

Exercise 7.3.4

Show that the Jacobian matrix of the linear vector field $\boldsymbol{V}$ defined in (7.20) is not symmetric. Conclude that the vector field is *not* a gradient field, i.e., conclude that $\boldsymbol{V}$ has no anti-derivative.

Let us formulate our findings as a lemma:

Lemma 7.3.1

Let $U \subseteq \mathbb{R}^n$ be an open subset, and let $\boldsymbol{V} : U \to \mathbb{R}^n$ be a C^1 vector field. If $\boldsymbol{V}$ is a gradient field, then $\boldsymbol{J}_{\boldsymbol{V}}(\boldsymbol{x})$ is symmetric for all $\boldsymbol{x} \in U$, i.e.,

$$\frac{\partial V_i}{\partial x_j}(\boldsymbol{x}) = \frac{\partial V_j}{\partial x_i}(\boldsymbol{x}) \quad \text{for all } i, j = 1, 2, \ldots, n. \tag{7.23}$$

Example 7.3.5

For the vector field

$$\boldsymbol{V}(x, y) = \begin{bmatrix} x^2 y \\ \cos(x + y) + e^x \end{bmatrix}$$

in Example 7.3.1 we have

$$\frac{\partial V_1}{\partial y}(x, y) = x^2$$

and

$$\frac{\partial V_2}{\partial x}(x, y) = -\sin(x + y) + e^x.$$

Thus, the condition (7.23) is not satisfied, i.e., $\boldsymbol{V}$ is not a gradient vector field.

We will see in Theorem 7.4.2 on page 204 that the converse of Lemma 7.3.1 is also true if we assume the domain U is simple connected. A subset U is *simple connected* if it is a connected set (recall Definition 5.2.1 on page 144) and if U is "without holes" that goes completely through the set. Intuitively, "without holes" means that any closed curve can be continuously transformed into a single point in U without leaving the subset U. Hence, the sphere in $\mathbb{R}^3$ is simply connected since any closed curve, i.e., a loop, on the sphere can be shrunk to a point. On the other hand, the "sphere" in $\mathbb{R}^2$ and the torus in $\mathbb{R}^3$ are not simply connected since a loop can be "trapped" around a hole.

Definition 7.3.2 Simply connected set

Let $U \subseteq \mathbb{R}^n$ be a connected set. Let $\boldsymbol{x}$ and $\boldsymbol{y}$ be any two points in U, and $\boldsymbol{r}_i : [0, 1] \to U$, $i = 1, 2$, be two continuous curves from $\boldsymbol{x}$ to $\boldsymbol{y}$ in U, that is, $\boldsymbol{r}_i([0, 1]) \subset U$, $i = 1, 2$, $\boldsymbol{x} = \boldsymbol{r}_1(0) = \boldsymbol{r}_2(0)$ and $\boldsymbol{y} = \boldsymbol{r}_1(1) = \boldsymbol{r}_2(1)$. If for any two such curves, $\boldsymbol{r}_1$ can be continuously deformed into $\boldsymbol{r}_2$ without leaving U, the set U is *simply connected*. (It is beyond the scope of this text to explain precisely what is meant by "continuously deformed".)

Let us introduce a large class of subsets, the so-called star-shaped sets, that are straightforward to define and that are simply connected.

Definition 7.3.3 Star-shaped set

A subset $S \in \mathbb{R}^n$ is called a *star-shaped* set if the exists a point (a "center" of S) $\boldsymbol{x} \in S$ such that the line segment between $\boldsymbol{x}$ any other point in S is contained in S.

Example 7.3.6

Let $n \in \mathbb{N}$, and let us consider subsets of $\mathbb{R}^n$.

(a) The sets $\mathbb{R}^n$ itself is star-shaped and therefore also simply connected. More generally, any convex set is star-shaped and therefore also simply connected. Open (and closed) balls $B(\boldsymbol{x}, r)$ and rectangles in $\mathbb{R}^n$ are examples of convex sets. Convex sets are, in fact, star-shaped where any point in the set can be chosen as the center $\boldsymbol{x}$ according to Definition 7.3.3.

(b) Let $r_2 > r_1$. The annulus

$$\{\boldsymbol{x} \in \mathbb{R}^n \mid r_1 \leq \|\boldsymbol{x}\| \leq r_2\} = \overline{B(\mathbf{0}, r_2)} \setminus B(\mathbf{0}, r_2)$$

is a ring-shaped figure between two concentric balls. It is not star-shaped for any value of $n \in \mathbb{N}$ as it has no "center". If $n = 1$, the set, being a union of two disjoint intervals $[-r_2, -r_1] \cup [r_1, r_2]$, is not even connected so it cannot be simply connected. If $n = 2$, the annulus is connected, but it is not simply connected as any closed curve looping around the inner circle cannot be contracted to a point. However, if $n \geq 3$, the annulus is simply connected as any closed curve can be continuously transformed into a single point.

(c) Let $\boldsymbol{x} \in \mathbb{R}^n$ be any point; often we take $\boldsymbol{x} = \mathbf{0}$. The set $\mathbb{R}^n \setminus \{\boldsymbol{x}\}$ has the same properties as the annulus when it comes to being star-shaped and simply connected (or not), e.g., for $n = 2$ the set is not simply connected, while for $n \geq 3$ the set is simply connected. The arguments are the same as for the annulus (e.g., by considering $r_1 \to 0$ and $r_2 \to \infty$).

We can now state a converse of Lemma 7.3.1. It is beyond the scope of this text to give a proof of the result.

Lemma 7.3.2

Let $U \subseteq \mathbb{R}^n$ be an open and *simply connected* subset, and let $\boldsymbol{V} : U \to \mathbb{R}^n$ be a C^1 vector field. If $\boldsymbol{J}_{\boldsymbol{V}}(\boldsymbol{x})$ is symmetric for all $\boldsymbol{x} \in U$, i.e., if (7.23) holds, then $\boldsymbol{V}$ is a gradient field.

Example 7.3.7

For the vector field

$$\boldsymbol{V}(x, y) = \begin{bmatrix} y^2 + 2x \\ 2xy + 3y^2 \end{bmatrix}, \ (x, y) \in \mathbb{R}^2$$

we have

$$\frac{\partial V_1}{\partial y}(x, y) = 2y$$

and

$$\frac{\partial V_2}{\partial x}(x, y) = 2y.$$

Since $\mathbb{R}^2$ is a star-shaped set, and therefore simply connected, and since Equation (7.23) on page 198 is satisfied, it follows from Lemma 7.3.2 that

$\boldsymbol{V}$ is a gradient vector field. This is, of course, what we already knew from Example 7.3.3.

7.4 Line integrals of vector fields and computing anti-derivatives

In (7.8) we defined the line integral of scalar functions of n variables. It is straightforward to extend this integral to a vector-valued integral of vector functions using the same definition as in Equation (6.79) on page 182. However, for vector *fields* $\boldsymbol{V} : U \to \mathbb{R}^n$, $U \subseteq \mathbb{R}^n$, it is often more useful to integrate the projection of the vector fields onto the tangent vector of the curve. Let $\boldsymbol{r} : [a, b] \to \mathbb{R}^n$ be a regular C^1 parametrization that is injective on $]a, b[$. Recall that the orthogonal projection of the vector $\boldsymbol{V}(\boldsymbol{r}(u))$ onto $\boldsymbol{r}'(u)$ is given by

$$\operatorname{proj}_{\boldsymbol{r}'(u)}(\boldsymbol{V}(\boldsymbol{r}(u)) = \left\langle \boldsymbol{V}(\boldsymbol{r}(u)), \frac{\boldsymbol{r}'(u)}{\|\boldsymbol{r}'(u)\|} \right\rangle \frac{\boldsymbol{r}'(u)}{\|\boldsymbol{r}'(u)\|}$$

where $\frac{\boldsymbol{r}'(u)}{\|\boldsymbol{r}'(u)\|}$ is a unit norm tangent vector. The norm of the projection is the absolute value of the inner product of $\boldsymbol{V}(\boldsymbol{r}(u)$ and the unit norm tangent vector. If we think of the vector field as a force field, and the curve $\mathcal{C} = \boldsymbol{r}([a, b])$ the trajectory along which we move, e.g., a particle, the work done by traversing the particle along the curve needs to take sign of the inner product of $\boldsymbol{V}(\boldsymbol{r}(u)$ and the unit norm tangent vector into account. Hence, the combined work is calculated as line integral along the curve $\mathcal{C}$ of the *scalar-valued* function

$$\boldsymbol{u} \mapsto \left\langle \boldsymbol{V}(\boldsymbol{r}(u)), \frac{\boldsymbol{r}'(u)}{\|\boldsymbol{r}'(u)\|} \right\rangle \tag{7.24}$$

Hence, by Equation (7.8), the integral is:

$$\begin{aligned} \int_{\mathcal{C}} \left\langle \boldsymbol{V}(\boldsymbol{r}(u)), \frac{\boldsymbol{r}'(u)}{\|\boldsymbol{r}'(u)\|} \right\rangle \mathrm{d}s &= \int_a^b \left\langle \boldsymbol{V}(\boldsymbol{r}(u)), \frac{\boldsymbol{r}'(u)}{\|\boldsymbol{r}'(u)\|} \right\rangle \|\boldsymbol{r}'(u)\| \,\mathrm{d}u \\ &= \int_a^b \langle \boldsymbol{V}(\boldsymbol{r}(u)), \boldsymbol{r}'(u) \rangle \,\mathrm{d}u \end{aligned}$$

Let us turn these ideas into a definition.

Definition 7.4.1 Line integral of a vector field

Let $U \subseteq \mathbb{R}^n$ be an open set, and let $\boldsymbol{V} : U \to \mathbb{R}^n$ be a continuous vector field. Let $\boldsymbol{r} : [a, b] \to \mathbb{R}^n$ denote a regular, piecewise C^1 parametrization of the curve $\mathcal{C} \coloneqq \boldsymbol{r}([a, b]) \subset U$. The *line integral of the vector field* is defined as

$$\int_{\mathcal{C}} \boldsymbol{V} \cdot \mathrm{d}\boldsymbol{s} = \int_a^b \langle \boldsymbol{V}(\boldsymbol{r}(u)), \boldsymbol{r}'(u) \rangle \, \mathrm{d}u \tag{7.25}$$

Note that $\int_{\mathcal{C}} \boldsymbol{V} \cdot \mathrm{d}\boldsymbol{s}$ is just our notation for the line integral of $\boldsymbol{V}$ along $\mathcal{C} = \boldsymbol{r}([a, b])$. In particular, we cannot take the dot product of $\boldsymbol{V}$ and $d\boldsymbol{r}$. The notation is used since (7.25) can be written as

$$\int_{\mathcal{C}} \boldsymbol{V} \cdot \mathrm{d}\boldsymbol{s} = \int_a^b \boldsymbol{V}(\boldsymbol{r}(u)) \cdot \boldsymbol{r}'(u) \, \mathrm{d}u.$$

Example 7.4.1

Consider the vector field

$$\boldsymbol{V}(x, y) = \begin{bmatrix} x + y \\ xy \end{bmatrix}, \quad (x, y) \in \mathbb{R}^2.$$

(a) The line integral of $\boldsymbol{V}$ over the curve $\mathcal{C}_1$ parametrized as

$$\boldsymbol{r}_1(u) = \begin{bmatrix} u \\ u \end{bmatrix}, \quad u \in [0, 1],$$

is

$$\begin{aligned} \int_{\mathcal{C}_1} \boldsymbol{V} \cdot \mathrm{d}\boldsymbol{s} &= \int_a^b \boldsymbol{V}(\boldsymbol{r}_1(u)) \cdot \boldsymbol{r}_1'(u) \, \mathrm{d}u \\ &= \int_0^1 \begin{bmatrix} 2u \\ u^2 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ 1 \end{bmatrix} \mathrm{d}u \\ &= \int_0^1 (2u + u^2) \, \mathrm{d}u = \frac{4}{3}. \end{aligned}$$

(b) The line integral of $\boldsymbol{V}$ over the curve $\mathcal{C}_2$ parametrized as

$$\boldsymbol{r}_2(u) = \begin{bmatrix} u \\ u^2 \end{bmatrix}, \quad u \in [0, 1],$$

is

$$\int_{\mathcal{C}_2} \boldsymbol{V} \cdot \mathrm{d}\boldsymbol{s} = \int_a^b \boldsymbol{V}(\boldsymbol{r}_2(u)) \cdot \boldsymbol{r}_2'(u) \, \mathrm{d}u$$

$$= \int_0^1 \begin{bmatrix} u+u^2 \\ u^3 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ 2u \end{bmatrix} \mathrm{d}u$$
$$= \int_0^1 (u + u^2 + 2u^4)\ \mathrm{d}u = \frac{37}{30}.$$

Note that both the curves $\mathcal{C}_1$ and $\mathcal{C}_2$ start at the point $(0,0)$ and end at the point $(1,1)$. In particular, we see that, in general, different paths that connect two points also lead to different values of the line integrals.

Given a vector field $\boldsymbol{V}$ and a parametrization of a curve $\mathcal{C}$, we can directly calculate the line integral (7.25). However, the calculation can, of course, be cumbersome if the parametrization $\boldsymbol{r}(t)$ is given by a complicated expression. A significant simplification occurs if the vector field $\boldsymbol{V}$ is a *gradient field*. Then, it turns out, the line integral of $\boldsymbol{V}$ over a curve $\mathcal{C}$ only depends on the *starting point* and the *endpoint* of the curve, but it is *independent* of the curve itself. This is the content of the following result.

Lemma 7.4.1

Let $U \subseteq \mathbb{R}^n$ be an open set, and let $\boldsymbol{V} : U \to \mathbb{R}^n$ be a continuous vector field. Let $\boldsymbol{r} : [a,b] \to \mathbb{R}^n$ denote a regular, piecewise C^1 parametrization of the curve $\mathcal{C} \coloneqq \boldsymbol{r}([a,b]) \subset U$. Suppose $\boldsymbol{V}$ has an anti-derivative $f : U \to \mathbb{R}$, i.e., suppose $\boldsymbol{V}$ is a gradient field $\boldsymbol{V} = \nabla f$. Then

$$f(\boldsymbol{r}(b)) - f(\boldsymbol{r}(a)) = \int_{\mathcal{C}} \boldsymbol{V} \cdot \mathrm{d}\boldsymbol{s}. \tag{7.26}$$

Proof. Consider a continuously differentiable parametrization $\boldsymbol{r} : [a,b] \to \mathbb{R}^n$ of the curve $\mathcal{C}$. Using $\boldsymbol{V} = \nabla f$, we can rewrite the integrand of the line integral as

$$\boldsymbol{V}(\boldsymbol{r}(u)) \cdot \boldsymbol{r}'(u) = \nabla f(\boldsymbol{r}(u)) \cdot \boldsymbol{r}'(u).$$

Define $h = f \circ \boldsymbol{r}$, i.e., $h(u) = f(\boldsymbol{r}(u))$ for $u \in [a,b]$. By Corollary 3.7.2 on page 109, the function h is differentiable, and the derivative of h is given by $h'(u) = \langle \boldsymbol{r}'(u), \nabla f(\boldsymbol{r}(u)) \rangle = \nabla f(\boldsymbol{r}(u)) \cdot \boldsymbol{r}'(u)$. Therefore, by inserting in the definition of the line integral, we get

$$\int_{\mathcal{C}} \boldsymbol{V} \cdot \mathrm{d}\boldsymbol{s} = \int_a^b \nabla f(\boldsymbol{r}(u)) \cdot \boldsymbol{r}'(u)\ \mathrm{d}u = \int_a^b h'(u)\,\mathrm{d}u$$
$$= [h(u)]_a^b = h(b) - h(a) = f(\boldsymbol{r}(b)) - f(\boldsymbol{r}(a)$$

as claimed.

In case the continuous curve $\boldsymbol{r} : [a, b] \to \mathbb{R}^n$ is *piecewise* C^1, we use the above argument on each C^1 piece $u \in [u_{i-1}, u_i]$. We then piece together each of these finitely many pieces using the insertion rule of integration stated in Equation (6.8) on page 155. This completes the proof.

■

Lemma 7.4.1 explains why gradient fields are also called conservative vector fields. They represent forces of physical systems in which energy is conserved, where the amount of work done as one moves through a path in configuration space is determined solely by the path's starting and endpoints. This allows for the definition of potential energy (given by f) that does not depend on the specific route traveled.

We now turn to the question of *computing* the anti-derivatives (i.e., the potential) of a vector field that is *already known* to be a gradient field. The formula for the anti-derivatives is given as a line integral of the vector field:

Theorem 7.4.2 Anti-derivative of a vector field

Let $U \subseteq \mathbb{R}^n$ be an open and *connected set*, and let $\boldsymbol{x}_0 \in U$. Let $\boldsymbol{V} : U \to \mathbb{R}^n$ be a continuous vector field. Suppose $\boldsymbol{V}$ has an anti-derivative, i.e., suppose $\boldsymbol{V}$ is a gradient field. Then the anti-derivative $f : U \to \mathbb{R}$ that satisfies $f(\boldsymbol{x}_0) = 0$ is given by

$$f(\boldsymbol{x}) = \int_{\mathcal{C}} \boldsymbol{V} \cdot \mathrm{d}\boldsymbol{s}, \quad \boldsymbol{x} \in U, \tag{7.27}$$

where $\mathcal{C} \subset U$ is *any* piecewise C^1 curve from $\boldsymbol{x}_0$ to $\boldsymbol{x}$, that is, there is a piecewise C^1 parametrization $\boldsymbol{r} : [a, b] \to \mathbb{R}^n$ such that $\boldsymbol{r}$ is regular, $\boldsymbol{x}_0 = \boldsymbol{r}(a)$, $\boldsymbol{x} = \boldsymbol{r}(b)$, and $\mathcal{C} = \boldsymbol{r}([a, b]) \subset U$.

Moreover, all anti-derivatives are of the form $f(\boldsymbol{x}) + c$ for some constant $c \in \mathbb{R}$.

The choice $f(\boldsymbol{x}_0) = 0$ in Theorem 7.4.2 is arbitrary in the sense that if $f : U \to \mathbb{R}$ is any anti-derivative of $\boldsymbol{V} : U \to \mathbb{R}^n$ and if $\mathcal{C} \subset U$ is *any* piecewise C^1 curve from $\boldsymbol{x}_0 \in U$ to $\boldsymbol{x} \in U$ that runs entirely inside U, then f satisfies

$$f(\boldsymbol{x}) - f(\boldsymbol{x}_0) = \int_{\mathcal{C}} \boldsymbol{V} \cdot \mathrm{d}\boldsymbol{s}, \quad \boldsymbol{x} \in U. \tag{7.28}$$

The results in (7.27) and in (7.28) are just reformulations of the result in Lemma 7.4.1 on the preceding page. As a consequence, the integral in (7.27) is (also) independent of the chosen curve as long as the curve is continuous, starts in $\boldsymbol{x}_0$, and ends in $\boldsymbol{x}$.

Note that U is assumed to be connected, hence any two points can, by definition, be connected by a continuous curve. However, Theorem 7.4.2 uses a *piecewise* C^1 continuous curve. It can be shown that, using U is also assumed to be open, that one can always modify the continuous curve so that it becomes piecewise C^1. It is beyond the scope of this text to include the proof. However, we do prove the final piece of new information in Theorem 7.4.2 the "Moreover"-part: if $f : U \to \mathbb{R}$ is an anti-derivative of $\boldsymbol{V}$, then clearly so is $f + c$, where $c \in \mathbb{R}$ is a constant since

$$\nabla(f + c) = \nabla f + \nabla c = \boldsymbol{V} + \boldsymbol{0} = \boldsymbol{V}.$$

On the other hand, if f_1 and f_2 are two anti-derivatives, then for $g \coloneqq f_1 - f_2$

$$\nabla g = \nabla f_1 - \nabla f_2 = \boldsymbol{V} - \boldsymbol{V} = \boldsymbol{0}.$$

Now, (7.28) shows that g is constant since

$$g(\boldsymbol{x}) - g(\boldsymbol{x}_0) = \int_{\mathcal{C}} \nabla g \cdot \mathrm{d}\boldsymbol{s} = \int_{\mathcal{C}} \boldsymbol{0} \cdot \mathrm{d}\boldsymbol{s} = 0$$

for all values of $\boldsymbol{x}, \boldsymbol{x_0} \in U$. Thus, $f_1 = f_2 + c$ for some constant $c \in \mathbb{R}$.

In order to apply Theorem 7.4.2, it is an advantage to choose the curve $\mathcal{C}$ as simple as possible. One possibility is to take $\boldsymbol{x}_0 = \boldsymbol{0}$ and simply let $\mathcal{C}$ be the straight line from $\boldsymbol{0}$ to $\boldsymbol{x}$. Let us see how this works in practice.

Example 7.4.2

Consider the vector field

$$\boldsymbol{V}(x, y) = \begin{bmatrix} y^2 + 2x \\ 2xy + 3y^2 \end{bmatrix}, \quad (x, y) \in \mathbb{R}^2.$$

In order to decide whether $\boldsymbol{V}$ is a gradient vector field, we will now apply Equation (7.27). The simplest curve connecting the starting point $\boldsymbol{x}_0 = \boldsymbol{0}$ and the endpoint $\boldsymbol{x} = (x, y)$ is the straight line parametrized by

$$\boldsymbol{r} : [0, 1] \to \mathbb{R}^2, \quad \boldsymbol{r}(u) = u\boldsymbol{x} = u \begin{bmatrix} x \\ y \end{bmatrix}.$$

With this choice (7.27) yields the function

$$\begin{aligned} f(x, y) = \int_{\mathcal{C}} \boldsymbol{V} \cdot \mathrm{d}\boldsymbol{s} &= \int_0^1 \boldsymbol{V}(\boldsymbol{r}(u)) \cdot \boldsymbol{r}'(u)\, \mathrm{d}u \\ &= \int_0^1 \begin{bmatrix} u^2y^2 + 2xu \\ 2u^2xy + 3u^2y^2 \end{bmatrix} \cdot \begin{bmatrix} x \\ y \end{bmatrix} \mathrm{d}u \end{aligned}$$

$$= \int_0^1 (u^2y^2x + 2x^2u + 2u^2xy^2 + 3u^2y^3)\,\mathrm{d}u$$
$$= xy^2 + x^2 + y^3.$$

Thus, the function $f(x,y) = xy^2 + x^2 + y^3$ is our *candidate* for an anti-derivative of $\boldsymbol{V}$. A fast check, as we already did in Example 7.3.3, confirms that $\boldsymbol{V} = \nabla f$ indeed is satisfied, i.e., $\boldsymbol{V}$ is a gradient field, and functions $f(x,y) = xy^2 + x^2 + y^3 + c$, where $c \in \mathbb{R}$, are the anti-derivatives of $\boldsymbol{V}$.

For applications of Theorem 7.4.2, it is important to stress that (7.27) defines a function even for vector fields without anti-derivatives. Hence, we should only use the formula (7.27) *after* we have proved existence of an anti-derivative *or* we should check that the computed function f indeed satisfies $\nabla f = \boldsymbol{V}$. To prove that a vector field has an anti-derivative, that is, to prove the vector field is a gradient field *without* computing the line integral (7.27), we usually apply Lemma 7.3.2 on page 200.

As an alternative, the existence of anti-derivatives is actually characterized in the following result. However, the result is difficult to use in practice as one has to check a condition for infinitely many curves.

Corollary 7.4.3 The circulation theorem

Let $U \subseteq \mathbb{R}^n$ be an open and *connected set*. Suppose $\boldsymbol{V} : U \to \mathbb{R}^n$ is a continuous vector field. Then the following assertions are equivalent:

(i) The vector field $\boldsymbol{V}$ is a gradient field, i.e., it has an anti-derivative.

(ii) For any *closed* and piecewise C^1 curve $\mathcal{C}$ in U it holds

$$\int_{\mathcal{C}} \boldsymbol{V} \cdot \mathrm{d}\boldsymbol{s} = 0. \tag{7.29}$$

Proof. Assume (i) holds. Since $\mathcal{C}$ is assumed to be closed, any parametrization $\boldsymbol{r} : [a,b] \to \mathbb{R}^n$ of $\mathcal{C}$ satisfies $\boldsymbol{r}(b) = \boldsymbol{r}(a)$. Since the anti-derivative f satisfies (7.26), it follows that

$$0 = f(\boldsymbol{r}(b)) - f(\boldsymbol{r}(b)) = f(\boldsymbol{r}(b)) - f(\boldsymbol{r}(a)) = \int_{\mathcal{C}} \boldsymbol{V} \cdot \mathrm{d}\boldsymbol{s},$$

which is the assertion in (ii). We leave the proof of (ii) $\Rightarrow$ (i) to the reader. ■

Example 7.4.3

Consider the curve parametrized as

$$\boldsymbol{r} : [0, 2\pi] \to \mathbb{R}^2,\ \boldsymbol{r}(u) = \begin{bmatrix} \cos u \\ \sin u \end{bmatrix}.$$

Then $\boldsymbol{r}(0) = \boldsymbol{r}(2\pi)$, so the curve is closed. Thus, for any gradient field $\boldsymbol{V}$,

$$\int_{\mathcal{C}} \boldsymbol{V} \cdot \mathrm{d}\boldsymbol{s} = 0.$$

For *closed* curves $\mathcal{C}$ the line integral $\int_{\mathcal{C}} \boldsymbol{V} \cdot \mathrm{d}\boldsymbol{s}$ is called the *circulation* of the vector field, and the integral is sometimes denoted by

$$\oint_{\mathcal{C}} \boldsymbol{V} \cdot \mathrm{d}\boldsymbol{s}$$

to stress the fact that $\mathcal{C}$ is assumed to be closed.

7.5 Surface integrals of vector fields

We introduced the integral of scalar functions over m-folds in $\mathbb{R}^n$ and, in particular, over surfaces in $\mathbb{R}^3$ in Section 7.2. The scalar function of interest in this section is the projection of a vector field onto the (unit) normal vector of a surface. The integral of this scalar function over the surface is called the *flux* of a vector field. For simplicity, we restrict ourselves to vector fields in $\mathbb{R}^3$.

Let $\boldsymbol{V} : U \to \mathbb{R}^3$, $U \subseteq \mathbb{R}^3$ be a vector field, and let $\mathcal{F}$ be a piecewise C^1 surface in $\mathbb{R}^3$ parametrized by $\boldsymbol{r} : \Gamma \to \mathbb{R}^3$, where $\Gamma \subset \mathbb{R}^2$. For $(u_1, u_2) \in \Gamma$ the normal vector is $\boldsymbol{n}_{\mathcal{F}}(u_1, u_2) = \boldsymbol{r}'_{u_1}(u_1, u_2) \times \boldsymbol{r}'_{u_2}(u_1, u_2)$. Similar to (7.24), the integrand, i.e., the scalar function, is

$$(u_1, u_2) \mapsto \left\langle \boldsymbol{V}(\boldsymbol{r}(u_1, u_2)), \frac{\boldsymbol{n}_{\mathcal{F}}(u_1, u_2)}{\|\boldsymbol{n}_{\mathcal{F}}(u_1, u_2)\|} \right\rangle, \tag{7.30}$$

where $\frac{\boldsymbol{n}_{\mathcal{F}}(u_1,u_2)}{\|\boldsymbol{n}_{\mathcal{F}}(u_1,u_2)\|}$ is a unit normal vector. Note that, the direction of the normal vector depends on the chosen parametrization, e.g., $-\boldsymbol{r}'_{u_1}(u_1, u_2) \times \boldsymbol{r}'_{u_2}(u_1, u_2)$ is also a normal vector to the surface at the point $\boldsymbol{r}(u_1, u_2) \in \mathcal{F}$ and the integrand will change sign if we use this definition of the normal vector.

By (7.11), the surface integral of the function (7.30) reads:

$$\int_\Gamma \left\langle \boldsymbol{V}(\boldsymbol{r}(u_1,u_2)), \frac{\boldsymbol{n}_\mathcal{F}(u_1,u_2)}{\|\boldsymbol{n}_\mathcal{F}(u_1,u_2)\|} \right\rangle \|\boldsymbol{n}_\mathcal{F}(u_1,u_2)\| \,\mathrm{d}(u_1,u_2) = \int_\Gamma \left\langle \boldsymbol{V}(\boldsymbol{r}(u_1,u_2)), \boldsymbol{n}_\mathcal{F}(u_1,u_2) \right\rangle \mathrm{d}(u_1,u_2).$$

Let us turn these ideas into a definition.

Definition 7.5.1 Flux: surface integral of a vector field

Let $U \subseteq \mathbb{R}^3$ be an open set, and let $\boldsymbol{V} : U \to \mathbb{R}^3$ be a continuous vector field. Let $\boldsymbol{r} : \Gamma \to \mathbb{R}^3$ denote a regular, piecewise C^1 parametrization of the surface $\mathcal{F} := \boldsymbol{r}(\Gamma)$. The *surface integral of the vector field* is defined as

$$\int_\mathcal{F} \boldsymbol{V} \cdot \mathrm{d}\boldsymbol{S} = \int_\Gamma \left\langle \boldsymbol{V}(\boldsymbol{r}(u_1,u_2)), \boldsymbol{n}_\mathcal{F}(u_1,u_2) \right\rangle \mathrm{d}(u_1,u_2). \tag{7.31}$$

Note that $\int_\mathcal{F} \boldsymbol{V} \cdot \mathrm{d}\boldsymbol{S}$ is just our notation for the integral of the vector field $\boldsymbol{V}$ over the surface $\mathcal{F}$. The value of this integral is also called the *flux* of the vector field through the surface. Note that the sign of the flux changes if we use $-\boldsymbol{n}_\mathcal{F}(u_1,u_2)$ as normal vector.

Example 7.5.1

Consider the cylinder surface $\mathcal{Z}$ from Example 7.2.4 of radius $r > 0$ and height $h > 0$ parametrized by

$$\boldsymbol{r}(u,v) = \begin{bmatrix} r\cos(u) \\ r\sin(u) \\ v \end{bmatrix}, \quad (u,v) \in \Gamma = [0, 2\pi[\times[0,h].$$

Let $\boldsymbol{V} : \mathbb{R}^3 \to \mathbb{R}^3$ be given by

$$\boldsymbol{V}(x,y,z) = \begin{bmatrix} x + 2\sinh(z) \\ y \\ z(h-z) \end{bmatrix}.$$

We want to compute the flux of $\boldsymbol{V}$ out through the cylinder, where we by "out" understand the direction away from the z-axis.

We already computed the normal vector in the previous example:

$$\boldsymbol{n}_\mathcal{Z}(u,v) = \boldsymbol{r}'_u(u,v) \times \boldsymbol{r}'_v(u,v) = \begin{bmatrix} r\cos(u) \\ r\sin(u) \\ 0 \end{bmatrix},$$

and we note that it points away from the z-axis which is the desired direction.

It is not necessary to compute the Jacobian of the parametrization as the integrand in the flux integral simply is:

$$\begin{aligned}\langle \boldsymbol{V}(\boldsymbol{r}(u,v)), \boldsymbol{n}_{\mathcal{Z}}(u,v)\rangle &= \Big(r\cos(u) + 2\sinh(v)\Big) r\cos(u) \\ &\quad + r\sin(u) r\sin(u) + v(h-v)\,0 = r^2 + 2r\sinh(v)\cos(u)\end{aligned}$$

Hence, the flux is

$$\begin{aligned}\int_{\mathcal{Z}} \boldsymbol{V}\cdot \mathrm{d}\boldsymbol{S} &= \int_{\Gamma} \langle \boldsymbol{V}(\boldsymbol{r}(u,v)), \boldsymbol{n}_{\mathcal{Z}}(u,v)\rangle \,\mathrm{d}(u,v) \\ &= \int_0^h \int_0^{2\pi} r^2 + 2r\sinh(v)\cos(u)\,\mathrm{d}u\,\mathrm{d}v\end{aligned}$$

We split the integral in two integrals and compute first

$$\begin{aligned}\int_0^h \int_0^{2\pi} 2r\sinh(v)\cos(u)\,\mathrm{d}u\,\mathrm{d}v &= \int_0^h 2r\sinh(v)[\sin(u)]_0^{2\pi}\,\mathrm{d}v \\ &= \int_0^h 2r\sinh(v)(0-0)\,\mathrm{d}v = 0.\end{aligned}$$

We then compute

$$\int_0^h \int_0^{2\pi} r^2\,\mathrm{d}u\,\mathrm{d}v = 2\pi r^2 h$$

Using linearity of the integral, we conclude that the flux is

$$\int_{\mathcal{Z}} \boldsymbol{V}\cdot \mathrm{d}\boldsymbol{S} = 2\pi r^2 h.$$

Appendices

APPENDIX A

Additional Proofs

A.1 Proof of Taylor's theorem

Before we prove Taylors theorem we state some results which will be used in the proof. We begin with *Rolle's theorem*:

Theorem A.1.1

Assume that the function $f : [a, b] \to \mathbb{R}$ is differentiable and that $f(a) = f(b) = 0$. Then there exists a point $\xi \in]a, b[$, for which $f'(\xi) = 0$.

Proof. The function f has a maximal value as well as a minimal value in the interval $[a, b]$. If the maximal value equals the minimal value, then f is constant, and $f'(x) = 0$ for all $x \in]a, b[$. On the other hand, if the minimal value is different from the maximal value, the assumption $f(a) = f(b)$ implies that there exists $\xi \in]a, b[$ where f assumes an extremal value; but this implies that $f'(\xi) = 0$.

■

Lemma A.1.2

Assume that $f : [a, b] \to \mathbb{R}$ is arbitrarily often differentiable, and let x, x_0 be arbitrary points in $[a, b]$, with $x > x_0$. Assume that for some $N \in \mathbb{N}$,

$$f(x) = 0 \text{ and } f^{(j)}(x_0) = 0 \quad \text{for all} \quad j \in \{0, 1, \ldots, N\}. \tag{A.1}$$

Then there exists $\xi \in]x_0, x[$ such that $f^{(N+1)}(\xi) = 0$.

Proof. The proof consists of a continued application of Rolle's theorem. We first prove the result for $N = 0$, and then move on with $N = 1$, etc. Observe that when we want to prove the theorem for one value of N, the assumption

(A.1) is available for all $j \in \{0, 1, \ldots, N\}$; this will play a key role in the proof.

For $N = 0$, our assumption is that $f(x) = f(x_0) = 0$. We shall prove the existence of a point $\xi_1 \in]x_0, x[$ for which$f'(\xi_1) = 0$; but this is exactly the conclusion in Rolle's theorem.

Now consider $N = 1$; in this case, the assumption (A.1) is available for $j = 0$ and $j = 1$. Via the assumption for $j = 0$ we just proved the existence of $\xi_1 \in]x_0, x[$, for which $f'(\xi_1) = 0$. Now, if we use the assumption for $j = 1$ we also have that $f'(x_0) = 0$. Applying Rolle's theorem again, this time to the function f' and the interval $[x_0, \xi_1]$, we conclude that there exists $\xi_2 \in]x_0, \xi_1[$ such that $f''(\xi_2) = 0$. This is exactly the conclusion we wanted to obtain for $N = 1$.

We now consider the general case, where $N > 1$. Thus, the assumption (A.1) means that

$$f(x) = f(x_0) = f'(x_0) = f''(x_0) = \cdots = f^{(N)}(x_0) = 0.$$

Applying Rolle's theorem to the function f and the interval $[x_0, x]$, we can find $\xi_1 \in]x_0, x[$ for which $f'(\xi_1) = 0$. Applying Rolle's theorem to f' and the interval $[x_0, \xi_1]$ leads to a point $\xi_2 \in]x_0, \xi_1[$ for which $f''(\xi_2) = 0$. Applying this procedure $N + 1$ times finally leads to a point ξ_{N+1} in $]x_0, \xi_N[$ for which $f^{(N+1)}(\xi_{N+1}) = 0$.

■

Now we only need one step before we are ready to prove Taylor's theorem. As before, we let P_N denote the Nth Taylor polynomial for f at x_0.

Lemma A.1.3

Assume that $f : [a, b] \to \mathbb{R}$ is arbitrarily often differentiable, and let $x_0 \in]a, b[$, $N \in \mathbb{N}$. Then, for each $x \in [a, b]$ we can find ξ between x and x_0 such that

$$f(x) = P_N(x_0) + \frac{f^{(N+1)}(\xi)}{(N+1)!}(x - x_0)^{N+1}.$$

Proof. Fix $x \in [a, b]$. The statement certainly holds if $x = x_0$, so we assume that $x_0 < x$ (the proof is similar if $x_0 > x$). Consider the function ϕ defined via

$$\phi(t) = f(t) - P_N(t) - K(t - x_0)^{N+1}; \tag{A.2}$$

here K is a constant, which we choose such that $\phi(x) = 0$, i.e.,

$$K = \frac{f(x) - P_N(x)}{(x - x_0)^{N+1}}.$$

Differentiating leads to

$$\phi'(t) = f'(t) - P_N'(t) - (N+1)K(t-x_0)^N,$$

and more generally, for each integer $j \in \{1, 2, 3, \ldots, N+1\}$,

$$\begin{aligned}\phi^{(j)}(t) = f^{(j)}(t) - P_N^{(j)}(t) \\ - (N+1)N(N-1)\cdots(N-j+2)K(t-x_0)^{N-j+1}.\end{aligned}$$

In particular,

$$\phi^{(N+1)}(t) = f^{(N+1)}(t) - P_N^{(N+1)}(t) - (N+1)!K. \tag{A.3}$$

By definition,

$$\begin{aligned}P_N(t) = f(x_0) + \frac{f'(x_0)}{1!}(t-x_0) + \frac{f''(x_0)}{2!}(t-x_0)^2 \\ + \frac{f^{(3)}(x_0)}{3!}(t-x_0)^3 + \cdots + \frac{f^{(N)}(x_0)}{N!}(t-x_0)^N.\end{aligned}$$

In this expression the first term is a constant, so

$$\begin{aligned}P_N'(t) &= f'(x_0) + \frac{f''(x_0)}{2!}2(t-x_0) \\ &\quad + \frac{f^{(3)}(x_0)}{3!}3(t-x_0)^2 + \cdots + \frac{f^{(N)}(x_0)}{N!}N(t-x_0)^{N-1} \\ &= f'(x_0) + f''(x_0)(t-x_0) \\ &\quad + \frac{f^{(3)}(x_0)}{2!}(t-x_0)^2 + \cdots + \frac{f^{(N)}(x_0)}{(N-1)!}(t-x_0)^{N-1}.\end{aligned}$$

Differentiating once more,

$$\begin{aligned}P_N''(t) &= f''(x_0) + \frac{2}{2!}f^{(3)}(x_0)(t-x_0) \\ &\quad + \cdots + \frac{(N-1)}{(N-1)!}f^{(N)}(x_0)(t-x_0)^{N-2} \\ &= f''(x_0) + f^{(3)}(x_0)(t-x_0) + \cdots + \frac{f^{(N)}(x_0)}{(N-2)!}(t-x_0)^{N-2}.\end{aligned}$$

More generally, for $j \in \{1, 2, 3, \ldots, N\}$,

$$P_N^{(j)}(t) = f^{(j)}(x_0) + f^{(j+1)}(x_0)(t-x_0)$$

$$+\frac{f^{(j+2)}(x_0)}{2!}(t-x_0)^2+\cdots+\frac{f^{(N)}(x_0)}{(N-j)!}(t-x_0)^{N-j}.$$

Note that

$$P_N^{(N+1)}(t)=0,\ \forall t\in\mathbb{R}.$$

For $j\in\{0,1,\ldots,N\}$, we have $P_n^{(j)}(x_0)=f^{(j)}(x_0)$, and therefore $\phi^{(j)}(x_0)=0$. This means that the function ϕ satisfies the conditions in Lemma A.1.2. Thus, there exists $\xi\in]x_0,x[$ for which $\phi^{(N+1)}(\xi)=0$. But according to (A.3),

$$\phi^{(N+1)}(\xi)=f^{(N+1)}(\xi)-(N+1)!K$$

so

$$f^{(N+1)}(\xi)=(N+1)!K \text{ and } K=\frac{f^{(N+1)}(\xi)}{(N+1)!}.$$

Using the expression (A.2) for ϕ and that $\phi(x)=0$, we see that

$$f(x)=P_N(x)+\frac{f^{(N+1)}(\xi)}{(N+1)!}(x-x_0)^{N+1}.$$

■

Here is finally the proof for Taylor's theorem, in the case $I=[a,b]$:

Proof of Theorem 4.3.3 on page 126. Let $N\in\mathbb{N}$ and $x\in]a,b[$ be given. According to Lemma A.1.3 we can find $\xi\in]a,b[$ such that

$$f(x)=\sum_{n=0}^{N}\frac{f^{(n)}(x_0)}{n!}(x-x_0)^n+\frac{f^{(N+1)}(\xi)}{(N+1)!}(x-x_0)^{N+1}.$$

Thus,

$$f(x)-\sum_{n=0}^{N}\frac{f^{(n)}(x_0)}{n!}(x-x_0)^n=\frac{f^{(N+1)}(\xi)}{(N+1)!}(x-x_0)^{N+1}.$$

By assumption we have that $|f^{(N+1)}(x)|\leq C$ for all $x\in]a,b[$, so this implies that

$$\begin{aligned}\left|f(x)-\sum_{n=0}^{N}\frac{f^{(n)}(x_0)}{n!}(x-x_0)^n\right|&=\left|\frac{f^{(N+1)}(\xi)}{(N+1)!}(x-x_0)^{N+1}\right|\\&\leq\frac{C}{(N+1)!}\left|x-x_0\right|^{N+1}.\end{aligned}$$

This finally proves (4.14).

■

APPENDIX B

Symbols and terminology

B.1 Short list of symbols

Symbol	Meaning
$\mathbb{R}$	the real numbers
$\mathbb{C}$	the complex numbers
$\mathbb{Z}$	the integers, i.e., $0, \pm 1, \pm 2, \ldots$
$\mathbb{N}$	the positive integers, i.e., $1, 2, \ldots$
$\mathbb{Q}$	the rational numbers, the fractions
$x \in A$	x is an element in the set A
$A \cup B$	the union of the sets A and B
$A \cap B$	the intersection of the sets A and B
$A \subseteq B$	the set A is a subset of the set B, possible that $A = B$
$A \subset B$	the set A is a subset of the set B and $A \neq B$
$\forall$	for all
$\exists$	there exists
$f : A \to B$	function f with domain A and codomain B
A	a matrix
A^{-1}	inverse of a matrix, defined when $\det(A) \neq 0$
$\boldsymbol{x}$	a vector
$\boldsymbol{f}$	a vector function

B.2 Short mathematical dictionary (English ↔ Danish)

From English to Danish:

English	Danish
absolute value	absolutværdi, modulus (i tilfældet af komplekse tal)
affine map	affin afbildning
anti-derivative	stamfunktion
argument (of a complex number)	argument (af et komplekst tal)
area	areal
augmented matrix	totalmatrix
ball	kugle
bi-implication	biimplikation
bijective, bijection	bijektiv, bijektion
binomial equation, the	den binome ligning
bounded set	begrænset mængde
boundary (of a set)	rand (af en mængde)
the chain rule	kædereglen
closed set	afsluttet eller lukket mængde
closure of a set	afslutning af en mængde
codomain	dispositionsmængde
column (of a matrix)	søjle (*aldrig* kolonne)
column vector	søjlevektor
complement (of a set), the	komplementet (af en mængde)
complex conjugate	kompleks konjugerede
complex number	komplekst tal
complex plane, the	den komplekse talplan
composite, composition	sammensat, sammensætning
connected set	sammenhængende mængde
conservative field	konservativt felt
continuous function	kontinuert funktion
coordinate function	koordinatfunktion
coordinate vector	koordinatvektor
cross product	krydsprodukt
curve	kurve
definite integral	bestemt integral

English	**Danish**
degree	grad
derivative of a function	den afledte af en funktion
diagonal	diagonal
diagonal matrix	diagonalmatrix
diagonalizable	diagonalisérbar
differential	differential
directional derivative	retningsafledet
discrete Fourier transform	diskret Fourier-transformation
disjoint	disjunkt
domain (of a function)	definitionsmængde
dot product	prikprodukt
double integration	dobbeltintegral
(row) echelon form	trappeform
ellipsoid	ellipsoide
endpoint	slutpunkt
equation	ligning
even number	lige tal
extremum	ekstremum
field	legeme
flux	flux
Fourier matrix	Fourier-matrix
fraction	brøk
general solution	fuldstændige løsning
geometric tensor	geometrisk tensor
gradient	gradient
gradient field	gradientfelt
gradient vector	gradientvektor
graph	graf
graph surface	grafflade
Hessian matrix	Hesse-matricen
hyperbola	hyperbel
identity matrix	identitetsmatrix
image (of a function)	billede, værdimængde
implication	implikation
indefinite integral	ubestemt integral
indeterminate	ubekendte
inequality	ulighed
injective	injektiv

English	**Danish**
inner product	indre produkt
integration by substitution	integration ved substitution
integration, partial	partiel integration
integer	heltal
interior of a set	indre af en mængde
intersection of sets	fællesmængde
inverse element	invers element
invertible matrix	invertérbar matrix
Jacobian determinant	Jacobi-determinant
Jacobian matrix	Jacobi-matrix
length of a curve	længden af en kurve
level curves	niveaukurver
level sets	niveaumængder
limit	grænseværdi
line integral	(tangentielt) kurveintegral
linear combination	linearkombination
linear map	lineær afbildning
linear vector field	lineært vektorfelt
linearly (in)dependent	lineært (u)afhængig
local maximum	lokalt maksimum
local minimum	lokalt minimum
lower bound	undertal
lower triangular matrix	nedre trekantsmatrix
matrix, matrices	matrix, matricer (*aldrig* matrice)
matrix of a linear map	afbildningsmatrix
matrix representation, mapping matrix	afbildningsmatrix
modulus	modulus, absolutværdi
natural number	naturligt tal
neural network	neuralt netværk
norm	norm
odd number	ulige tal
orthogonal basis	ortogonal basis
orthonormal basis	ortonormal basis
partial derivative	partiel afledet
partial integration	delvis integration, partiel integration
particular solution	partikulær løsning
polar coordinates	polære koordinater

English	**Danish**
polynomial	polynomium
pre-image	urbillede
principal value of the argument	hovedargument (af et komplekst tal)
product rule	produktreglen
primitive function	stamfunktion
quotient rule	kvotientreglen
range (of a function)	billede, værdimængde
rank	rang (af en matrix)
real line, the	den reelle tallinje
real number	reelt tal
reduced row echelon form	reduceret trappeform (af en matrix)
rectified linear unit (ReLU)	rektificeret lineær enhed
root of a polynomial	rod i et polynomium
row (of a matrix)	række
row echelon form	trappeform
row vector	rækkevektor
scalar	skalar
set	mængde
span	udspænding (af en mængde af vektorer)
sphere	sfæren (i $\mathbb{R}^n$)
square matrix	kvadratisk matrix
square root	kvadratrod
surface	flade
surjective	surjektiv
tangent vector	tangentvektor
transpose	transponerede (matrix)
triangular matrix	trekantsmatrix
union (of sets)	foreningsmængde
unit ball	enhedskuglen
unit circle	enhedscirklen (i $\mathbb{R}^2$ eller $\mathbb{C}$)
unit sphere	enhedssfæren
upper bound	overtal
upper triangular matrix	øvre trekantsmatrix
vector field	vektorfelt
vector function	vektorfunktion
vector space	vektorrum
vertex (vertices)	knude(r), hjørne(r)

English	Danish
zero vector	nulvektor

From Danish to English:

Danish	English
absolutværdi, modulus (i tilfældet af komplekse tal)	absolute value
afbildningsmatrix	matrix of a linear map
afbildningsmatrix	matrix representation, mapping matrix
affin afbildning	affine map
areal	area
argument (af et komplekst tal)	argument (of a complex number)
begrænset mængde	bounded set
billede, værdimængde	image (of a function)
billede, værdimængde	range (of a function)
biimplikation	bi-implication
bijektiv, bijektion	bijective, bijection
brøk	fraction
den afledte af en funktion	derivative of a function
den binome ligning	binomial equation, the
den komplekse talplan	complex plane, the
den reelle tallinje	real line, the
definitionsmængde	domain (of a function)
diskret Fourier-transformation	discrete Fourier transform
disjunkt	disjoint
dispositionsmængde	codomain
dobbeltintegral	double integration
diagonal	diagonal
diagonalmatrix	diagonal matrix
diagonalisérbar	diagonalizable
direkte produkt	dot product
direkte felt	conservative field
dispositionsmængde	codomain
diskret Fourier-transformation	discrete Fourier transform
ellipsoide	ellipsoid
enhedskuglen	unit ball

Danish	**English**
enhedscirklen (i $\mathbb{R}^2$ eller $\mathbb{C}$)	unit circle
enhedssfæren	unit sphere
extremum	extremum
fællesmængde	intersection of sets
flade	surface
flux	flux
foreningsmængde	union (of sets)
Fourier-matrix	Fourier matrix
fuldstændige løsning	general solution
geometrisk tensor	geometric tensor
gradient	gradient
gradientfelt	gradient field
gradientvektor	gradient vector
graf	graph
grafflade	graph surface
grænseværdi	limit
hovedargument (af et komplekst tal)	principal value of the argument
hyperbel	hyperbola
identitetsmatrix	identity matrix
indre produkt	inner product
indre af en mængde	interior of a set
injektiv	injective
integral ved substitution	integration by substitution
invertérbar matrix	invertible matrix
kædereglen	the chain rule
kompleks konjugerede	complex conjugate
komplekst tal	complex number
komplementet (af en mængde)	complement (of a set), the
kugle	ball
kurve	curve
kvadratrod	square root
kvadratisk matrix	square matrix
lige tal	even number
linearkombination	linear combination
lineært (u)afhængig	linearly (in)dependent
lineært vektorfelt	linear vector field
lokalt maksimum	local maximum
lokalt minimum	local minimum

Danish	**English**
ligning	equation
længden af en kurve	length of a curve
mængde	set
modulus, absolutværdi	modulus
nedre trekantsmatrix	lower triangular matrix
neuralt netværk	neural network
niveaukurver	level curves
niveaumængder	level sets
norm	norm
nulvektor	zero vector
polære koordinater	polar coordinates
prikprodukt	dot product
produktreglen	product rule
række	row (of a matrix)
rækkevektor	row vector
rektificeret lineær enhed	rectified linear unit (ReLU)
rod i et polynomium	root of a polynomial
sfæren (i $\mathbb{R}^n$)	sphere
slutpunkt	endpoint
span	span
stamfunktion	anti-derivative
stamfunktion	primitive function
surjektiv	surjective
tangentvektor	tangent vector
totalmatrix	augmented matrix
transponerede (matrix)	transpose
trekantsmatrix	triangular matrix
trappeform	row echelon form
ulige tal	odd number
undertal	lower bound
urbillede	pre-image
vektorfelt	vector field
vektorfunktion	vector function
vektorrum	vector space

B.3 Some well-known formulas

Useful formulas:

$$
\begin{aligned}
\mathrm{e}^{(a+i\omega)t} &= \mathrm{e}^{at}(\cos\omega t + i\sin\omega t),\ a,\omega\in\mathbb{R};\\
\cos\omega t &= \frac{1}{2}(\mathrm{e}^{i\omega t}+\mathrm{e}^{-i\omega t}),\ \omega\in\mathbb{R};\\
\sin\omega t &= \frac{1}{2i}(\mathrm{e}^{i\omega t}-\mathrm{e}^{-i\omega t}),\ \omega\in\mathbb{R};\\
a+ib &= r\,\mathrm{e}^{i\theta},\ r=|a+ib|,\ \theta=\mathrm{Arg}(a+ib),\\
a &= r\cos\theta,\quad b=r\sin\theta;\\
|a+ib| &= \sqrt{a^2+b^2},\ a,b\in\mathbb{R}\ \text{(Pythagoras)};\\
|a+b| &\le |a|+|b|,\ a,b\in\mathbb{C}\ \text{(the triangle inequality)};\\
\det\begin{pmatrix}a & b\\ c & d\end{pmatrix} &= ad-bc\\
\begin{pmatrix}a & b\\ c & d\end{pmatrix}^{-1} &= \frac{1}{ad-bc}\begin{pmatrix}d & -b\\ -c & a\end{pmatrix}\ \text{when } \det(A)=ad-bc\neq 0\\
\det\begin{pmatrix}a_{11} & a_{12} & a_{13}\\ a_{21} & a_{22} & a_{23}\\ a_{31} & a_{32} & a_{33}\end{pmatrix} &= a_{11}\begin{vmatrix}a_{22} & a_{23}\\ a_{32} & a_{33}\end{vmatrix} - a_{21}\begin{vmatrix}a_{12} & a_{13}\\ a_{32} & a_{33}\end{vmatrix} + a_{31}\begin{vmatrix}a_{12} & a_{13}\\ a_{22} & a_{23}\end{vmatrix}\\
\sin(x+y) &= \sin x\cos y+\cos x\sin y\\
\cos(x+y) &= \cos x\cos y-\sin x\sin y
\end{aligned}
$$

Relationship between growth of classical functions:

$$
\begin{aligned}
\frac{x^a}{b^x} &\to 0 \text{ for } x\to\infty \text{ for all } a>0, b>1\\
\frac{\ln x}{x^a} &\to 0 \text{ for } x\to\infty \text{ for all } a>0.
\end{aligned}
$$

Partial integration:

$$
\int f(x)g(x)\,dx = F(x)g(x) - \int F(x)g'(x)\,\mathrm{d}x,\ \text{where } F(x)=\int f(x)\,dx.
$$

Integration by substitution:

$$
\int f(g(x))g'(x)\,\mathrm{d}x = F(g(x)),\ \text{where } F(x)=\int f(x)\,\mathrm{d}x.
$$

Some useful integrals:

$$\int x^n \, \mathrm{d}x = \frac{1}{n+1} x^{n+1}, n \neq -1;$$
$$\int \frac{1}{x} \, \mathrm{d}x = \ln(x);$$
$$\int \cos(x) \, \mathrm{d}x = \sin(x);$$
$$\int \sin(x) \, \mathrm{d}x = -\cos(x);$$
$$\int \mathrm{e}^x \, \mathrm{d}x = \mathrm{e}^x \, .$$

Bibliography

[1] Garcia, S. and Horn, R. *A Second Course in Linear Algebra*. Cambridge Mathematical Textbooks. Cambridge University Press, 2017. ISBN: 9781107103818.

[2] Gravesen, J. *Vector Calculus – Additional Lecture Notes for 01002 Mathematics 1b*. DTU Compute, Jan. 2024.

[3] Karlsson, P. W. and Hansen, V. L. *Matematisk Analyse 2: Funktioner af flere reelle variable*. Dansk. Institut for Matematik, DTU, 1998.

[4] Rudin, W. *Principles of Mathematical Analysis*. McGraw-Hill, 1976.

Index

1-norm, 39
2-norm, 39
C^0 vector function, 91
C^1 vector function, 114
C^m vector function, 114
∞-norm, 39
p-norm, 39

activation functions, 19
affine map, 18
affine map
    differential, 112
anti-derivative, 155
    vector field, 197
approximating polynomial, *see* Taylor polynomial
area, 165
    surface, 192
attains its maximum, 139
attains its minimum, 139

ball
    in V, 49
basis, 55
    orthonormal, 56
bias vector, 22
bijective, 7
boundary
    of a set, 51
bounded, 3
bounded above, 2
bounded from below, 2
bounded set, 53

Cartesian coordinates, 172
chain rule
    generalized, 115
    simple, 109
change of variables
    in $\mathbb{R}^2$, 170
change of variable
    in $\mathbb{R}^n$, 178
circle, 30
circulation, 207
closed set, 51
closure of set, 52
codomain, 6
complement, 50
composition, 21
connected set, 144
conservative field, *see* gradient field
continuity
    vector function, 88
continuous function
    of one variable, 81
continuously differentiable function, 84

coordinate function, 18
coordinate system
    cylindrical coordinates, 180
    polar coordinates, 173
    rectangular coordinates, 172
    spherical coordinates, 181
cross product, 187
curve
    parametrization, 86
    boundary, 52
    connected set, 144
    in $\mathbb{R}^k$, 86
    regular, 87
    tangent vector, 87
curve integral, *see* line integral
cylindrical coordinates, 180

deep neural network, 22
definite integral, 155
derived norm, 40
diagonalizable, 66
differentiability
    vector function of several variables, 112
differentiable function
    of one variable, 84
    of several variables, 106
differential, 112, 114
    scalar function of several variables, 106
directional derivative, 98
discontinuous, 81
Discrete Fourier Transform, 65
domain, 6
dot product, 36
double integration, 162

ellipsoid, 30
endpoint, 203
extremum, 139

feedforward neural network, *see* neural network
flux, 208
Fourier matrix, 65
function, 6
    composition, 21
    continuous, 88
    continuously differentiable, 84
    directional derivative, 98
    gradient, 96
    of one variable, 13, 80
    of several variables, 15
    partial derivative, 95
    rectified linear unit, 19
    ReLU, 19
    scalar functions of several variables, 20
    softmax, 20
    vector field, 20
    vector functions of several variables, 20

generalized chain rule, 115
geometric tensor, 186
gradient ascent method, 100
gradient descent method, 100
gradient field, 197
gradient vector, 96
Gram-Schmidt process, 58
Gramian, 187
graph
    for function of one variable, 27
    of function of several variables, 27
graph surface, 195
greatest lower bound, 2

Hessian matrix, 102
hyperbola, 30

image, 7
indefinite integral, 155

infimum, 2
injective, 7
injective on, 7
inner product, 39
  Frobenius, 42
  on $\mathbb{C}^n$, 38
  on $\mathbb{F}^n$, 41
  on L^2, 42
inner product space, 39
integration
  by substitution, 158
  linearity of, 157
  partial, 158
  rules, one variable, 154, 157
interior of set, 53
invertible matrix, 9

Jacobian, 186
Jacobian determinant, 167
Jacobian matrix, 112

least upper bound, 2
length
  curve, 191
level sets, 29
line integral, 191
  vector field, 202
linear vector field, 196
linear map
  differential, 112
local maximum, 142, 147
local minimum, 142, 147
lower bound, 2

matrix, 8
  diagonalizable, 66
  Fourier, 65
  Hermitian, 10, 67
  Hessian matrix, 102
  idempotant, 10
  inverse, 9
  Jacobian, 112
  normal, 10, 67
  ortogonal projection, 11
  positive definite, 10
  positive semi-definite, 11
  real diagonalizable, 66
  real orthogonal, 10
  similar to, 66
  symmetric, 10
  unitarily diagonalizable, 66
  unitary, 10
maximum, 3, 139
minimum, 3, 139
multi-layer perceptron, *see* neural network

neural network, 22
  definition, 22
  Hessian, 150
  shallow, 23
norm, 36, 38
  derived norm, 40
normal vector, 189
normed space, 38
normed vector space, 38

open set, 50
orthogonal, 43
orthogonal projection, 53, 78
orthogonal projection matrix, 11
orthonormal, 43
orthonormal basis, 56

parametrization
  curve, 21, 52, 86
  regular, 187
  surface, 21
parametrization of a set B, 170
parametrized
  curve, 146
partial integration, 158
partial derivative, 95
polar coordinates, 173

positive definite, 10
positive semi-definite, 11
potential, 197

quadratic form, 16
    definition, 16
quadratic forms
    symmetric matrix, 73
quadratic form
    differential, 108

range, 7
real orthogonally diagonalizable, 66
rectangle
    in $\mathbb{R}^2$, 160
    in $\mathbb{R}^n$, 176
rectangular coordinates, 172
ReLU, 19
remainder term, 124
Riemann integrable, 152, 164
Riemann sum, 152, 161, 164, 177
Rolle's theorem, 211

saddle point, 142, 149
scalar, 1
set
    boundary, 51
    bounded, 53
    closed, 51
    closure, 52
    complement, 50
    connected, 144
    interior, 53
    open, 50
shallow neural network, 22, 23
similar, 66
simply connected, 199
smooth set, 177
softmax, 20
    continuous, 90
    differentiable, 115
special functions, 15
spectral decomposition
    normal matrix, 75
    real, symmetric matrix, 72
Spectral theorem
    normal matrix, 75
    real, symmetric matrix, 70
spectrum, 35
sphere
    in $\mathbb{R}^3$, 31
spherical coordinates, 181
standard inner product, *see* inner product
    on $\mathbb{R}^n$, 36
star-shaped, 199
starting point, 203
stationary point, 145
steepest descent method, 100
stochastic gradient descent, 111
strict local maximum, 147
strict local minimum, 147
supremum, 2
surface integral, 192
    vector field, 208
surjective, 7

tangent
    curve, 87
    scalar function of one variable, 118
    scalar function of several variables, 129
    vector function of one variable, 87
tangent plane, *see* tangent
tangent vector, 87
    surface, 187
Taylor polynomial
    one variable, first degree, 118
Taylor polynomial
    notation, 118

one variable, Kth degree, 121
several variables, first degree, 129
several variables, second degree, 131
vector function of several variables, 136
Taylor's formula
remainder term, 124
scalar function of one variable, 125
scalar function of several variables, 134
Taylor's theorem
scalar function of one variable, 126
the Jacobian, 180
trace, 11

unit norm vector, 43
unit vector, 43
unitarily diagonalizable, 66
upper bound, 2

vector
notation, 16
vector field, 20, 184
vector function
C^0, 91
C^1, 114
C^m, 114
continuous, 88, 90
continuously differentiable, 114
differential, 112
vector function of several variables, 18
volume
in $\mathbb{R}^n$, 178

weight matrix, 22