Mathematics for Advanced Materials Science and Engineering Mathematics

Marc Technau

Marc Technau, Institut für Analysis und Zahlentheorie, Technische Universität Graz, Kopernikusgasse 24, 8010 Graz
Email address: mtechnau@math.tugraz.at
URL: https://www.math.tugraz.at/~mtechnau

2024-02-01 @ 10:30.

Preface

The present notes were originally composed for a mathematics course for students of the *Advanced Materials Science* Dipl. Ing. programme at Graz University of Technology. They have subsequently been adapted to also accomodate a similar course for the *Chemical and Pharmaceutical Engineering* Dipl. Ing. programme. Each of the two courses covers only a strict subset of these notes (cf. Table 1 below).

The intended audience for the courses at hand is supposed to have a background in chemistry and should have had exposure to mathematics at the level of the courses offered by Christian Elsholtz [2, 3]. The material in [2, 3] is neither a subset nor a superset of the material covered here. We do repeat *some* of the material contained in the above references, but our focus and pace here are different.

Moreover, the aim of a course such as the ones derived from these notes is primarily to arm the reader with the mathematics necessary to follow the other courses in the study programme. For this, a rigorous development of the underlying mathematics is obviously not required and would also be too time consuming. We do, however, try (at times) to give *some* insight into the wealth of the material presented. In particular, one can find the occasional 'proof' throughout the text. The uninitiated reader should rest assured that following the proofs is generally optional (albeit advisable!). On the other hand, any potential readers with a stronger mathematical background are kindly asked not to judge the 'proofs' too harshly with respect to rigour. (In fact, we often have to cheat quite a bit in the proofs and gloss over many non-trivial technical issues.)

Readers wishing to consult other sources may like the classic book by Kreyszig [7]. A very nice—yet much more advanced—exposition can be found in [9], which is written by physicists for physicists and never strays far away from actual applications to physics. Readers who are sufficiently fluent in German may also like the amazing book series by Jänich [5, 6, 4]. Jänich uses many figures and lucid explanations to give the reader an honest feel for the inner workings of the developed machinery without delving too deep into dirty technicalities. In fact, the present course attempts to take some inspiration from this approach.

In arranging the material for this course, I have relied on advice from colleagues from the physics department, who have kindly responded to my calls and emails. I would like to thank (in alphabetical order:) Enrico Arrigoni, Wolfgang Sprengel, and Egbert Zojer. Moreover, I have relied on lecture notes handed down to me by Herbert

Wallner and Kurt Tomantschger. In carrying out the later adaptations, I was kindly provided helpful information by Wolfgang Bauer, Günter Brenn, Stefan Radl, and Tim Zeiner.

—Marc Technau

Chapter	§	MAMS	EM
Chapter 0	§ 0.1	∼	∼
	§ 0.2	∼	∼
	§ 0.3	∼	∼
	§ 0.4	∼	∼
	§ 0.5		☑
	§ 0.6		☑
	§ 0.7	∼	☑
Chapter 1	§ 1.1	☑	
	§ 1.2	☑	
	§ 1.3	☑	
	§ 1.4	☑	
	§ 1.5	☑	
Chapter 2	§ 2.1	☑	
	§ 2.2	☑	
	§ 2.3	☑	
	§ 2.4	∼	
	§ 2.5	∼	
Chapter 3	§ 3.1	☑	☑
	§ 3.2	☑	☑
	§ 3.3	☑	☑
	§ 3.4	☑	☑
	§ 3.5	☑	☑
	§ 3.6	∼	∼
Chapter 4	§ 4.1	☑	
	§ 4.2	☑	
	§ 4.3	☑	
	§ 4.4	☑	
Chapter 5	§ 5.1	☑	☑
	§ 5.2	☑	☑
	§ 5.3	☑	☑
Chapter 6	§ 6.1	☑	☑
	§ 6.2		☑
	§ 6.3		☑
	§ 6.4	☑	☑
Chapter 7	§ 7.1	☑	☑
	§ 7.2	☑	☑
	§ 7.3	☑	☑
	§ 7.4	☑	☑
	§ 7.5	☑	☑
	§ 7.6	∼	∼
Chapter 8	§ 8.1		☑
	§ 8.2	☑	☑

Table 1. Overview of which parts of the notes concern which course. (MAMS = *Mathematics for Advanced Materials Science*; EM = *Engineering Mathematics*.)

Contents

Nomenclature

$*$	Convolution of functions, see Proposition 2.4 or Proposition 4.6 and note that both definitions differ.
$:=$	Symbol used to emphasise that the left hand side is *defined* as being equal to the right hand side (e.g., $e := \exp(1)$). This emphasis may be missing even when an equation is to be read as a definition, but usually the situation will be clear from the context. (The symbol is also used in the reverse form: '=:'.)
$\overset{!}{=}$	Symbol used to emphasise equations that ought to be *solved* in contrast to stating that a certain equation holds. This distinction can be subjective at times, so readers should not bother too much with this symbol and read it simply as '=' when in doubt.
$\vec{0}$	$= (0, \dots, 0) \in \mathbb{R}^n$, the zero vector in $\mathbb{R}^n$, where the choice of n depends on the context.
$\mathbf{1}_n$	The $n \times n$ unit matrix, see § 3.1.6.
$A \to B$	An unnamed map between two sets A and B, see § 0.2.
$\begin{pmatrix} \vert & \vert \\ \vec{a} & \vec{b} \\ \vert & \vert \end{pmatrix}$	A matrix with two columns, the columns being given by the vectors $\vec{a}$ and $\vec{b}$ respectively.
$\begin{pmatrix} a_{11} & \dots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \dots & a_{mn} \end{pmatrix}$	An $m \times n$-matrix, see § 3.1.
$\begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix}$	$= (a_1, \dots, a_n)$, a vector, usually in $\mathbb{R}^n$.
$\lvert z \rvert$	Absolute value of a complex number. $\lvert a + \mathrm{i}b \rvert = \sqrt{a^2 + b^2}$ for $a, b \in \mathbb{R}$. See Chapter 1.
arctan	The arcus tangent function, i.e., the inverse function of tan: $(-\frac{\pi}{2}, \frac{\pi}{2}) \to \mathbb{R}$.
$\Longrightarrow$	Arrow used to express that the statement on the left implies the one on the right. ('$A \Longrightarrow B$' means 'if statement A is true, then so is statement B'.)

$\Longleftarrow$	Arrow used to express that the statement on the right implies the one on the left. ('$A \Longrightarrow B$' means 'if statement B is true, then so is statement A'.)
$\Longleftrightarrow$	Arrow used to express that two statements are equivalent. '$A \Longleftrightarrow B$' means '$A \Longrightarrow B$' and '$A \Longleftarrow B$'.
$\mathbb{C}$	$= \{ a+ib : a, b \in \mathbb{R} \}$. The set of complex numbers, see Chapter 1
$\delta_{k\ell}$	The Kronecker delta symbol. It equals 1 if $k = \ell$ and equals 0 otherwise.
δ	Dirac delta distribution, see § 2.4
$\det A$	Determinant of a matrix, see § 3.2.
$\det f$	Determinant of a linear map $f : \mathbb{R}^n \to \mathbb{R}^n$, see § 3.2.
$\frac{\mathrm{d}f}{\mathrm{d}x}$	The derivative of a function f, see § 0.6 or § 1.4.
$\operatorname{rot} \vec{F}$	Divergence of a vector field $\vec{F}$, see § 6.1.
$\mathrm{d}A$	Area (surface) element, see § 7.2 or § 7.3 for a vector version.
$\mathrm{d}s$	Line element, see § 7.2 or § 7.3 for a vector version.
$\mathrm{d}V$	Volume element, see § 7.2 or § 7.3 for a vector version.
$\mathrm{d}x$	Differential of x. Used as a symbol in differentiation (see § 0.6 or § 1.4), in integration (see § 0.7 or Chapter 7), as notation for the total differential of a function (see § 5.1), or in the context of Cartan's calculus of differental forms (see § 7.6).
$\vec{e}_j$	The j-th standard unit vector, see § 3.1.1.
e^z	$= \exp(z)$.
$\exp$	The exponential function, see § 0.5.1 or § 1.5.
$f : A \to B$	A map called f between two sets A and B, see § 0.2.
f'	The derivative of a function f, see § 0.6 or § 1.4.
$\hat{f}(k)$	The k-th Fourier coefficient of the function $f : \mathbb{R} \to \mathbb{C}$, see § 4.2.
$f_{\to c}$	Right shift of the function f by c, see Proposition 2.4 or Proposition 4.6.
$f(A)$	The value of the map f evaluated at A, or possibly the image of A under f. The meaning should be obvious from the context, see § 0.2.
$f(x)$	The value of the map f evaluated at x, see § 0.2.
$f^{-1}(A_0)$	The pre-image of A_0 under the map f, see § 0.2.
$f^{-1}(x)$	The value of the inverse map of f at x, see § 0.2.
$g \circ f$	$= (x \mapsto g(f(x)))$. Composition of the maps g and f, see § 0.2.
$\operatorname{grad} f(\vec{x}_0)$	Gradient of f at $\vec{x}_0$, see § 5.2.
i	Imaginary unit. Satisfies the equation $\mathrm{i}^2 = -1$ (which is also satisfied by $-\mathrm{i}$). See Chapter 1.
$\operatorname{Im} z$	Imaginary part part of a complex number Chapter 1.
$\iint_{\vec{\Phi}(K)} f \, \mathrm{d}A$	Surface integral, see § 7.2.
$\iint_S \vec{K} \, \mathrm{d}\vec{A}$	Surface integral of a surface S over a vector field $\vec{K}$, see § 7.3.

$\int_{\vec{\Phi}(K)} f \, \mathrm{d}s$	Line integral, see § 7.2.
$\int_C \vec{K} \, \mathrm{d}\vec{s}$	Line integral of a curve C against a vector field $\vec{K}$, see § 7.3.
$\int \cdots \int_{\vec{\Phi}(K)} f \, \mathrm{d}V$	Higher-dimensional integral, see § 7.2.
$\int_Q f(\vec{x}) \, \mathrm{d}^n\vec{x}$	Higher-dimensional integral, see § 7.1.
$\int$	Integral, see § 0.7 for the basics and Chapter 7 for more.
$J_f(\vec{x}_0)$	Jacobian matrix of f at $\vec{x}_0$, see § 5.1.
$\vec{K}$	Spherical coordinates, see § 6.4.
$\mathscr{L}\{f\}$	Laplace transform of a function $f\colon [0,\infty) \to \mathbb{R}$, see Chapter 2.
$\lim_{x\to x_0}$	Limit, see § 0.4.
$\log$	The logarithm function (to base $e = \exp(1)$), see § 0.5.3.
$\mapsto$	Symbol used to declare how a map maps elements, see § 0.2.
$\mathbb{N}_0$	$= \{0, 1, 2, 3, \ldots\}$. The set of non-negative integers.
$\mathbb{N}$	$= \{1, 2, 3, 4, \ldots\}$. The set of natural numbers (positive integers).
∇	Nabla operator, see § 6.1.
$\|\vec{v}\|$	$= \sqrt{v_1^2 + \ldots + v_n^2}$. The length (or norm) of a vector $\vec{v} \in \mathbb{R}^n$, see § 3.3.
ω	A differential form, see § 7.6.
$\vec{P}$	Polar coordinates, see § 6.4.
$\frac{\partial f}{\partial \vec{v}}(\vec{x}_0)$	Directional derivative of f at $\vec{x}_0$ with respect to the direction $\vec{v}$, see § 5.1.
$\partial_k f(\vec{x}_0)$	k-th partial derivative of f at $\vec{x}_0$, i.e., the directional derivative of f at $\vec{x}_0$ with respect to the direction $\vec{e}_k$ (k-th standard unit vector), see § 5.1.
$\partial_{\vec{v}} f(\vec{x}_0)$	Directional derivative of f at $\vec{x}_0$ with respect to the direction $\vec{v}$, see § 5.1.
$\prod$	Symbol for a product, see § 0.3.
$\mathbb{Q}$	The set of rational numbers (fractions a/q with $a, b \in \mathbb{Z}$ and $b \neq 0$).
$\square$	Symbol usually marking the end of a proof.
$\mathbb{R}_{\geq 0}$	The set of non-negative real numbers.
$\mathbb{R}_+$	The set of positive real numbers.
$\mathbb{R}$	The set of real numbers (containing the rational numbers $\mathbb{Q}$ as a subset, and more elements like $-\sqrt{2}$, π, Euler's number e, etc.).
$\mathrm{Re}\, z$	Real part of a complex number, see Chapter 1.
$\mathrm{rot}\, \vec{F}$	Rotation (or curl) of a vector field $\vec{F}$, see § 6.1.
$\{\ldots\}$	A set, see § 0.1.
$\#\{\ldots\}$	Number of elements in a set.
$\setminus$	Complement of one set in another, see § 0.1.
$\in, \notin$	Notation for membership (or lack thereof) to a set, see § 0.1.
$\emptyset$	$= \{\}$. The set with no elements, called the empty set.

$\cap$, $\bigcap$	Intersection of two sets, see § 0.1.
$\subset$, $\subseteq$, $\subsetneq$, $\nsubseteq$	Notation for subsets, see § 0.1.
$\cup$, $\bigcup$	Union of two sets, see § 0.1.
sin, cos	The sine and cosine functions, see § 0.5.2 or § 1.5.
$\sum$	Symbol for a sum (or possibly an infinite series), see § 0.3.
tan	The tangent function, $\tan(z) = \sin(z)/\cos(z)$.
$\vec{v} \times \vec{w}$	The cross product of two vectors, see § 3.4.
$\vec{v} \cdot \vec{w}$	The dot product of two vectors, see § 3.3.
$\vec{v}$	A vector, see § 3.1.1. Its entries are usually denoted by $v_1, \ldots, v_n$ (if $\vec{v} \in \mathbb{R}^n$).
$\wedge$	Wedge product of differential forms, see § 7.6.
$\dot{x}$, $\ddot{x}$, $\dddot{x}$	Alternative notation for derivatives of a function, $\dot{x} = x'$, $\ddot{x} = x''$, and so on.
$x^{(k)}$	k-th derivative of a function x, $x^{(0)} = x$, $x^{(1)} = x'$, $x^{(2)} = x''$, and so on.
$\overline{z}$	Complex-conjugate of a complex number. $\overline{(a + \mathrm{i}b)} = a - \mathrm{i}b$ for $a, b \in \mathbb{R}$. See Chapter 1.
$\mathbb{Z}$	$= \{0, -1, 1, -2, 2, -3, 3, \ldots\}$. The set of integers.
$\vec{Z}$	Cylindrical coordinates, see § 6.4.

Greek alphabet. Mathematicians frequently use Greek letters, because it is convenient to have more resources to name objects at one's disposal. All letters have a lowercase and an uppercase variant. Some of them also have additional variants (e.g., ϕ and φ are both symbols for a lowercase phi).

Letters	Name	Letters	Name	Letters	Name
αA	Alpha	ιI	Iota	$\rho \varrho P$	Rho
βB	Beta	κK	Kappa	$\sigma \Sigma$	Sigma
$\gamma \Gamma$	Gamma	$\lambda \Lambda$	Lambda	τT	Tau
$\delta \Delta$	Delta	μM	Mu	$\upsilon \Upsilon$	Upsilon
$\epsilon \varepsilon E$	Epsilon	νN	Nu	$\phi \varphi \Phi$	Phi
ζZ	Zeta	$\xi \Xi$	Xi	χX	Chi
ηH	Eta	oO	Omicron	$\psi \Psi$	Psi
$\theta \vartheta \Theta$	Theta	$\pi \Pi$	Pi	$\omega \Omega$	Omega

CHAPTER 0

Basics

This chapter is meant to refresh some material that should already be well-known to the reader. (If not, please ask in the lecture.) A specific item may be here for one of two reasons:

(1) some familiarity with it will be assumed later during the course, or
(2) it was requested to be included by a professor from your study programme, but did not fit well enough into the arrangement of the remaining material as to be placed elsewhere.

0.1. Sets

0.1.1. Definitions. A ***set*** is the mathematician's one-size-fit-all bag. We shy away from trying to make this all too precise, but the basic idea is this: a set is a collection of ***elements*** much like a shopping cart may contain groceries. Sets are usually denoted by curly braces { and } encasing the elements contained in that set. For instance, $\{1,2,3\}$ is the set containing the elements 1, 2, and 3 (and no others). For any object x and any set M, either of the propositions 'x is an ***element of*** M' ('x is ***contained in*** M') or 'x is not an element of M' hold true. In the former case we may write '$x \in M$' and '$x \notin M$' in the latter case. One also often uses reversed versions of these symbols, i.e., '$M \ni x$' means '$x \in M$'.

Example. $1 \in \{1,2,3\}$, but $4 \notin \{1,2,3\}$.

Example. Sets *themselves* can be *elements*, e.g., $\{1,2\} \in \{\{\}, \{1\}, \{1,2\}, \{1,2,3\}\}$.

We often employ an ellipsis '…' in the definition of sets (and elsewhere). When such a device is encountered, the reader is expected to 'fill in the rest' in 'the obvious way'.

Example. $\{1,2,\ldots,5\}$ means $\{1,2,3,4,5\}$ and certainly not $\{1,2,42,5\}$.

For two sets A and B we say that A is a ***subset*** of B if for every element $x \in A$ one has $x \in B$. One writes this succinctly as '$A \subseteq B$'. If A is not a subset of B (i.e., there exists some element of A which is not contained in B), then we may write '$A \nsubseteq B$'. If $A \subseteq B$, but B contains an element which is not also contained in A (i.e., B contains strictly more elements than A), then we write '$A \subset B$' or '$A \subsetneq B$'.

Examples.

- $\{1,2\} \subseteq \{1,2,3\}$,

- $\{1,2,3\} \supseteq \{1,2\}$,
- $\{1,2\} \subset \{1,2,3\}$,
- $\{1,2\} \subsetneq \{1,2,3\}$,
- $\{1,2,3\} \subseteq \{1,2,3\}$,
- $\{1,2,3\} \not\subseteq \{1,2\}$.

The set containing no element at all is called the ***empty set***. (One may also say that a certain set is ***empty***.) It is denoted by $\emptyset$ or $\{\}$. Clearly the empty set is a subset of any set M: $\emptyset \subseteq M$.

Two sets A and B are said to be ***equal*** if and only if $A \subseteq B$ and $B \subseteq A$. Equivalently, this means that, for any x, x is an element of A if and only if x is an element of B.

Example. Note that $\{1,1\} \subseteq \{1\}$ and $\{1\} \subseteq \{1,1\}$, so $\{1,1\}$ and $\{1\}$ are the same set. There is no notion of 'multiplicity' of an element in a set.

0.1.2. Constructing subsets. Given a set A and some parametric statement $P(\cdot)$ which makes sense whenever the parameter is specialised to be an element of A, we use the notation $\{\, x \in A : P(x) \,\}$ to denote the subset of A which contains precisely those elements x of A for which $P(x)$ is a true statement.

Example. Let $A = \{1,2,3,4,5,6\}$. Then $\{\, x \in A : x > 4 \,\} = \{5,6\}$.

We also often employ similar constructions without referring to an encompassing set A. For instance, we write

$$\mathbb{C} = \{\, a + \mathrm{i}b : a, b \in \mathbb{R} \,\}$$

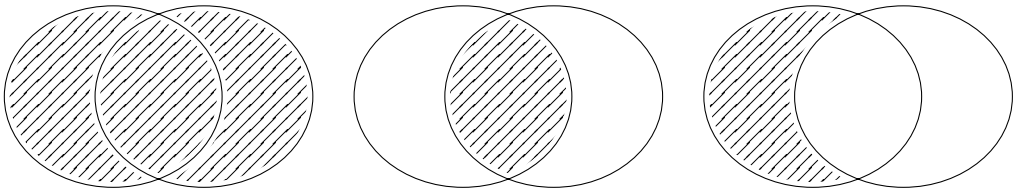
to define the complex numbers (see Chapter 1 and § 0.1.4), without specifying a set containing $a + \mathrm{i}b$ inside the notation $\{\ldots : \ldots\}$. Such abuse of notation can cause problems (see § 0.1.3 below), but we leave these issues to the mathematicians and hope that, in practice, the reader will have no trouble in deciphering the meaning of expressions such as the above.

0.1.3. A warning about naïve use of sets. The following example may hint at the fact that naïve manipulations with sets may yield contradictions. (Here, we resolve this issue by plainly ignoring it.)

Example (Russel's antinomy)**.** Let M denote the set of all sets not containing themselves, $M = \{A : A \notin A\}$. Is it true then, that $M \in M$? Certainly not, for this would contradict the definition of M. Hence, $M \notin M$. But this is contradictory as well, for if $M \notin M$, then $M \in M$ by another appeal to the definition of M. The generally accepted way for avoiding these sort of contradictions is by restricting the operations that one allows for defining sets. We do not discuss the details of this here.

0.1.4. Commonly used sets. There are various interesting sets of numbers. The most common ones are

- The ***positive integers***, denoted by $\mathbb{N} = \{1,2,3,4,5,6,\ldots\}$. These are also often called ***natural numbers***.



(a) $A \cup B$ **(b)** $A \cap B$ **(c)** $A \setminus B$

Figure 1. Illustration of various set operations. In each case, the two ellipses depict A and B respectively.

- The ***non-negative integers*** denoted by $\mathbb{N}_0 = \{0, 1, 2, 3, 4, 5, 6, \ldots\}$.
- The ***integers*** denoted by $\mathbb{Z} = \{0, -1, 2, -2, 2, -3, 3, \ldots\}$.
- The ***rational numbers***, denoted by $\mathbb{Q}$, which are given by all fractions a/q with $a, q \in \mathbb{Z}$ and $q \neq 0$.
- The ***real numbers***, denoted by $\mathbb{R}$. The elements of the reals are, roughly speaking, all decimal expansions which may be infinite to the right, for instance,

  $$3.14159265359\ldots \quad \text{(continuing somehow)}.$$

  The specifics concerning how the reals are constructed and, for instance, how to make sense of what it means 'to add' two reals[1] are left to those who are to open a book on the subject of *analysis*.
- $\mathbb{R}_{\geq 0}$ and $\mathbb{R}_+$ denote the sets of non-negative and positive real numbers respectively.
- The ***complex numbers***, denoted by $\mathbb{C}$ (see Chapter 1).

The order relation $\leq$ on the real numbers allows yields certain subsets of the real numbers known as intervals. These turn up all over the place in the study of analysis. For real numbers $a, b \in \mathbb{R}$ one defines

- the ***closed interval*** $[a, b] = \{\, x \in \mathbb{R} : a \leq x \leq b \,\}$,
- the ***open interval*** $(a, b) = \{\, x \in \mathbb{R} : a < x < b \,\}$,
- the ***(left-)half-open interval*** $(a, b] = \{\, x \in \mathbb{R} : a < x \leq b \,\}$,
- the ***(right-)half-open interval*** $[a, b) = \{\, x \in \mathbb{R} : a \leq x < b \,\}$.

0.1.5. Operations with sets. There are various other ways of constructing new sets from old ones. We list some common ones (see § 0.1.5). Let A and B be sets. Then ones defines their ***intersection*** $A \cap B$ by

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

[1] Observe that the well-known algorithm for adding integers by starting with the right-most digits and working with carrys is of limited use here, as any particular real number may not admit a 'right-most digit' to begin with.

their ***union***

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

and the ***complement*** of B in A (or 'A ***without*** B')

$$A \setminus B := \{\, x \in A : x \notin B \,\}.$$

Examples.

- $\{1,2\} \cap \{2,8\} = \{2\}$,
- $\mathbb{N} \cap \mathbb{N}_0 = \mathbb{N}$,
- $\{1,2\} \cup \{2,8\} = \{1,2,8\}$,
- $\mathbb{N}_0 \setminus \{0\} = \mathbb{N}$,
- $\mathbb{Z} \setminus \{-1,-2,-3,\ldots\} = \mathbb{N}$.

For taking the intersection or union of many sets one often employs larger symbols. For instance,

$$\bigcap_{n\in\mathbb{N}} \{1,n\} = \{1\}, \quad \text{and} \quad \bigcup_{n\in\mathbb{N}} \{n\} = \mathbb{N}.$$

One also allows for the formation of pairs. For two elements $a \in A$ and $b \in B$ one considers the pair (a,b). Two such pairs (a,b) and (a',b') are considered to be equal if and only if both $a = a'$ and $b = b'$. The set containing all such pairs is denoted by $A \times B$ and is called the ***Cartesian product*** of the sets A and B:

$$A \times B = \{\, (a,b) : a \in A \text{ and } b \in B \,\}.$$

Example. $\{1,2\} \times \{3,4,5\} = \{(1,3), (1,4), (1,5), (2,3), (2,4), (2,5)\}$.

Given multiple sets $A_1, A_2, \ldots$, one defines their product $A_1 \times A_2 \times \ldots$ similarly. For a given set A and $n \in \mathbb{N}$, one writes $A^n = A \times \ldots \times A$ (n times), e.g., $A^1 = A$ and $A^2 = A \times A$. Prime examples are sets of the type $\mathbb{R}^n$ and subsets thereof. These are generally used as domains of definition of functions of many variables.

If a set A contains only finitely many elements, then we write $\#A$ for the number of (distinct) elements in A.

Examples. $\#\{44,62\} = 2$, $\#\{1,2,3,3,3\} = 3$.

0.2. Maps

0.2.1. Definitions. A ***map*** f consists of three things: two sets A and B, and some rule which assigns to every element x of A an element $f(x)$ of B. The short-hand for such a map is $f : A \to B$. This should be read as referring to a map from A to B, indicated by '$A \to B$', and that map being given the label or name f. In this context, the set A is called the ***domain (of definition)*** of f and B is called the ***target set*** of f (or ***co-domain***).

To specify the rule which does the assignment, one often uses the notation $x \mapsto f(x)$. For instance the map $f : \mathbb{N} \to \mathbb{R}$, $x \mapsto \sqrt{x}$, sends every positive integer x to its square root.

Example. The 'map' $f\colon \mathbb{Z} \to \mathbb{R}$, $x \mapsto \sqrt{x}$, is not well-defined, because we have no idea what the square root of a negative number is supposed to be. (Okay, upon inspection of Chapter 1, we have a good idea of what such a square root could be, but it certainly is not a real number, so the problem remains.) The 'map' $f\colon \mathbb{N} \to \mathbb{Z}$, $x \mapsto \sqrt{x}$, is also not well-defined, because the square root of a positive integer is not an integer in general.

Examples.

(1) $f\colon \mathbb{R} \to \mathbb{R}$, $x \mapsto x^2$.
(2) $g\colon \mathbb{R} \to \mathbb{R}_{\geq 0}$, $x \mapsto x^2$.

Example (Identity map)**.** For any set A, we let id_A be the map $A \to A$, $a \mapsto a$, the so-called ***identity map***.

When the domain A of definition of a map $f\colon A \to B$ happens to be a subset of $\mathbb{R}^n$, and $\vec{a} = (a_1, \ldots, a_n) \in A$ is some element of A, we usually write $f(a_1, \ldots, a_n)$ for $f(\vec{a}) = f((a_1, \ldots, a_n))$, thus omitting the second pair of parentheses.

0.2.2. Image and preimage. Let $f\colon A \to B$ be a map. For every subset $A_0 \subseteq A$ we define the ***image*** of A_0 under f as

$$f(A_0) \coloneqq \{ f(a) : a \in A \} \coloneqq \{ b \in B : \text{there exists some } a \in A_0 \text{ such that } f(a) = b \}.$$

Moreover, for every subset $B_0 \subseteq B$, we define the ***preimage*** of B_0 under f as

$$f^{-1}(B_0) \coloneqq \{ a \in A : f(a) \in B_0 \}.$$

0.2.3. Functions. Maps whose target set consists of 'numbers' are usually also called ***functions***. (Here, ***number*** should mean complex number. Because of $\mathbb{N} \subset \mathbb{N}_0 \subset \mathbb{Z} \subset \mathbb{Q} \subset \mathbb{R} \subset \mathbb{C}$, this also includes real numbers, rational numbers, etc.)

Examples.

(1) The map $f\colon \mathbb{R} \to \mathbb{R}$, $x \mapsto x^2$, is a function.
(2) $h\colon \mathbb{R} \to \mathbb{R}^2$, $x \mapsto (x, x)$, is not a function according to the aforementioned convention.
(3) The map $c\colon \{\text{apple}, \text{orange}, \text{melon}, \text{tomato}\} \to \{\text{red}, \text{green}, \text{orange}, \text{blue}\}$, which assigns each fruit its corresponding color is certainly not a function.[2]
(4) The map $u\colon \mathbb{R}^2 \to \mathbb{R}$, $(x_1, x_2) \mapsto x_1 + 2x_2$, is a function, but in the context of linear algebra (cf. Chapter 3) we shall stick to calling it a map. (Actually, a *linear* map, because it is linear.)

[2] Yes, tomatoes are fruit and the question whether an apple is red or green (and similar questions) is something every reader may grapple with on their own. The author of these notes was too lazy to think of a better example.

0.2.4. Sequences. Functions $a\colon N \to B$ defined on some subset $N \subseteq \mathbb{N}_0$ which contains all sufficiently large natural numbers are also called ***sequences***. (In most cases of interest one has $N = \mathbb{N}_0$ or $N = \mathbb{N}$, but we wish to reserve the right to, say, speak of the sequence defined by $n \mapsto 1/(n-5)$, which makes sense for $n = 6, 7, 8, \ldots$, but not for 5.) When speaking in terms of sequences, the value $a(n) \in B$ for $n \in N$ is usually written as a_n. The sequence a is then often denotes by $(a(n))_{n\in N}$, or $(a(n))_n$ for short. In the latter notation, the range N for n is not specified, but usually one is only interested in the behaviour of the sequence for sufficiently large arguments/indices n. For that reasons it is mostly irrelevant which N one takes. Similarly, the precise choice of B is also usually not specified. In applications it will almost always be $\mathbb{R}$, $\mathbb{C}$, $\mathbb{R}^m$ or $\mathbb{C}^m$ for some $m \in \mathbb{N}$.

Example. The sequence $(1/n)_n$ means the function $\mathbb{N} \to \mathbb{R}$, $n \mapsto 1/n$, subject to the aforementioned ambiguities in terms of the domain of definition and the target set.

0.2.5. Composition. Given two maps $f : A \to B$ and $g\colon B \to C$, one may define their ***composition***, denoted by $g \circ f$. This is a map from A to C, given by $(g \circ f)(a) := g(f(a))$. (One particular application of the composition of maps is the chain rule which may allow one to compute the total differential of a complicated function, see § 5.3.)

Example. Consider $f : \mathbb{N} \to \mathbb{Z}$, $x \mapsto 2-3x$, and $g\colon \mathbb{Z} \to \mathbb{R}$, $y \mapsto \exp(y^5)$. Then $g \circ f$ is the map $\mathbb{N} \to \mathbb{R}$ which maps x to $g(f(x)) = g(2-3x) = \exp((2-3x)^5)$.

Lemma 0.1. *Composition of maps is associative, i.e., if $f\colon A \to B$, $g\colon B \to C$, and $h\colon C \to D$ are maps, then $(h \circ g) \circ f = h \circ (g \circ f)$.*

Proof. Both $(h \circ g) \circ f$ and $h \circ (g \circ f)$ are maps from A to D and upon expanding their definition, one finds that both of them map elements $a \in A$ to $h(g(f(a)))$. Hence, both maps coincide. To accomplish the aforementioned expansion, observe that

$$((h \circ g) \circ f)(a) = (h \circ g)(f(a)) = h(g(f(a))),$$

and

$$(h \circ (g \circ f))(a) = h((g \circ f)(a)) = h(g(f(a))). \qquad \square$$

One can use Lemma 0.1 to justify that we may omit parentheses altogether when composing maps. Thus, in the setting of Lemma 0.1 one may write $h \circ g \circ f$ to mean $(h \circ g) \circ f$ or, what is the same, $h \circ (g \circ f)$. Similar comments apply to when composing more than three maps.

0.2.6. Injectivity, surjectivity, and bijectivity. One defines many properties which a map may or may not have. For instance, in Chapter 3 we encounter *linear* maps, in Chapter 5 we encounter *differentiable* maps, and so on. Three of the most basic properties that a map may or may not have are *injectivity*, *surjectivity* and *bijectivity*. We now define these. Let $f : A \to B$ be any map. Then f is called ***injective*** if for any $b \in B$ there is *at most one* $a \in A$ such that $f(a) = b$. (Note that 'at most

one' does allow for the possibility that there is no such a at all.) An alternative way of phrasing the definition of injectivity is saying that whenever $a, a' \in A$ are such that $f(a) = f(a')$, then $a = a'$. The map f is called ***surjective*** if for any $b \in B$ there is *at least one* $a \in A$ such that $f(a) = b$. The map f is called ***bijective*** if it is both injective and surjective. (Equivalently, f is bijective if any only if for every $b \in B$ there is one and only one element $a \in A$ such that $f(a) = b$.)

Examples.

(1) The map $f\colon \mathbb{R} \to \mathbb{R}$, $x \mapsto x^2$, is neither injective ($f(-1) = 1 = f(1)$, but $-1 \neq 1$) nor surjective (there is no $x \in \mathbb{R}$ such that $f(x) = -42$) and, in particular, not bijective.
(2) The map $g\colon \mathbb{R} \to \mathbb{R}_{\geq 0}$, $x \mapsto x^2$ is surjective, because every non-negative real number y admits a square root $\sqrt{y} \in \mathbb{R}$. For this square root one has $g(\sqrt{y}) = y$. However, the map g is not injective (for the same reason as for f above).
(3) The map $h\colon \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$, $x \mapsto x^2$, is both injective and surjective (thus, bijective).
(4) The map $k\colon \mathbb{R}_{+} \to \mathbb{R}_{\geq 0}$, $x \mapsto x^2$, is not surjective, as there is no $x \in \mathbb{R}_{+}$ such that $k(x) = 0$.

Proposition 0.2. *Let $f\colon A \to B$ be a map between two non-empty sets A and B. Then the following statements hold:*

(1) f is injective if and only if there exists a map $l\colon B \to A$ such that $l \circ f = \mathrm{id}_A$.
(2) f is surjective if and only if there exists a map $r\colon B \to A$ such that $f \circ r = \mathrm{id}_B$.
(3) f is bijective if and only if there exists a map $g\colon B \to A$ such that $g \circ f = \mathrm{id}_A$ and $f \circ g = \mathrm{id}_B$. In this case, the map g is uniquely determined by f.

Proof. For the proof of (1), suppose first that f is injective. Then we may define the desired map l as follows: for every $b \in B$ for which there exists some $a \in A$ with $f(a) = b$, define $l(b) \coloneqq a$. For every other $b \in B$, define $l(b)$ to be an arbitrary element of A (this is possible, because A is assumed to be non-empty). This defines l. We contend that $l \circ f = \mathrm{id}_A$. Indeed, for any $a' \in A$, one has $(l \circ f)(a') = l(f(a'))$. Now for $b = f(a')$ there exists some $a \in A$ such that $f(a) = b$, namely one may pick $a = a'$. Suppose that a is as in the definition of $l(b)$. Then $l(b) = a$ and we have $(l \circ f)(a') = a$. It remains to show that $a = a'$. However, this is obviously true, because f is assumed to be injective and we have $f(a') = b = f(a)$, so we must have $a = a'$.

Conversely, suppose that one has a map $l\colon B \to A$ such that $l \circ f = \mathrm{id}_A$. We intend to show that f is injective. To this end, suppose that $a, a' \in A$ satisfy $f(a) = f(a')$. Then $l(f(a)) = l(f(a'))$. However, because of $l \circ f = \mathrm{id}_A$, this already shows $a = a'$, as desired. Hence, f is injective.

The proof of (2) is similar. We just hint that, in order to define r if f is assumed to be surjective, pick $b \in B$, then find any $a \in A$ such that $f(a) = b$ (which is possible by surjectivity of f) and let $r(b) \coloneqq a$. The details are left to the reader.

To show (3), note that if some $g\colon B \to A$ satisfying $g \circ f = \mathrm{id}_A$ and $f \circ g = \mathrm{id}_B$ exists, then g may take the roles of l and r from (1) and (2), so f must be both injective and surjective, hence, bijective. Conversely, suppose that f is bijective. Let l and r be as in (1) and (2). It then suffices to show that $l = r$, because then one may take g to be l (or, what is the same, r). To this end, observe that, using Lemma 0.1,

$$r = \mathrm{id}_A \circ r = (l \circ f) \circ r = l \circ (f \circ r) = l \circ \mathrm{id}_B = l.$$

To show that g, in case it exists, is determined uniqueless by f, suppose that $\tilde{g}\colon B \to A$ is another map such that $\tilde{g} \circ f = \mathrm{id}_A$ and $f \circ \tilde{g} = \mathrm{id}_B$. Then $\tilde{g} \circ f = \mathrm{id}_A = g \circ f$. Now, composing with g from the right, and suitably adding parentheses (using Lemma 0.1 once more), we find that $\tilde{g} \circ (f \circ g) = g \circ (f \circ g)$. As $f \circ g = \mathrm{id}_B$, we infer $\tilde{g} \circ \mathrm{id}_B = g \circ \mathrm{id}_B$, so that $\tilde{g} = g$. Hence, g is uniquely determined. □

In the setting of Proposition 0.2 (3), the map is also denoted by f^{-1} and called the ***inverse map*** of f. Observe that there is a slight ambiguity with respect to the *preimage* $f^{-1}(A_0)$ of a subset $A_0 \subseteq A$ under f (see § 0.2.2). However, confusion does usually not arise, for we generally do not consider maps $f\colon A \to B$ where there exists some A_0 which is both a subset *and* an element of A at the same time. Assuming now that f is bijective, writing f^{-1} for the map g from Proposition 0.2 (3), and write $f^{-\underline{1}}$ for the preimage as defined in § 0.2.2 with the extra underline added (just for now) for distinguishing both notions. Then, for any $A_0 \subseteq A$, one has

$$f^{-\underline{1}}(A_0) = \{ f^{-1}(a) : a \in A_0 \} = \text{image of } A_0 \text{ under } f^{-1}.$$

Example. Consider the map $f\colon \mathbb{R} \to \mathbb{R}_{\geq 0}$, $x \mapsto x^2$. The map $r\colon \mathbb{R}_{\geq 0} \to \mathbb{R}$, $y \mapsto \sqrt{y}$, satisfies $f \circ r = \mathrm{id}_{\mathbb{R}_{\geq 0}}$, because $(f \circ r)(y) = f(r(y)) = f(\sqrt{y}) = (\sqrt{y})^2 = y$. However, $(r \circ f)(-1) = r(f(-1)) = r((-1)^2) = r(1) = \sqrt{1} = 1 \neq -1$, so that $r \circ f \neq \mathrm{id}_{\mathbb{R}}$.

Example. Consider the set $A = \prod_{n \in \mathbb{N}} \mathbb{Z}$. Elements of A are given by sequences $(a_1, a_2, a_3, \ldots)$ of integers, indexed by the natural numbers. We define $l\colon A \to A$ to be the ***left-shift*** on A, given by $l(a_1, a_2, a_3, \ldots) = (a_2, a_3, \ldots)$ (note that a_1 is omitted on the right hand side). Similarly, we define the ***right-shift*** $r\colon A \to A$, given by $r(a_1, a_2, a_3, \ldots) = (0, a_1, a_2, a_3, \ldots)$. Clearly we have

$$(l \circ r)(a_1, a_2, \ldots) = l(r(a_1, a_2, \ldots)) = l(0, a_1, a_2, \ldots) = (a_1, a_2, \ldots).$$

Hence, $l \circ r = \mathrm{id}_A$. Note that l is surjective and r is injective. However, l is not injective and r is not surjective. We also note that $(r \circ l)(a_1, a_2, \ldots) = (0, a_2, \ldots)$, so $r \circ l \neq \mathrm{id}_A$. In fact, there is no map $g\colon A \to A$ satisfying $r \circ g = \mathrm{id}_A$ or $g \circ l = \mathrm{id}_A$.

We have seen that injectivity, surjectivity and bijectivity depend in an essential way on the domain and target set of a given map. Note that, for any map $f\colon A \to B$ and any subset A_0 of A, one naturally gets a map $f_0\colon A_0 \to B$ given by $f_0(a) = f(a)$. One may be inclined to write '$f_0 = f$' (in fact, some authors do, depending on the context). We, however, generally try to distinguish f_0 and f, because their domains are different (at least, unless $A_0 = A$). We write $f|_{A_0}$ for f_0 and say that $f|_{A_0}$ is the ***restriction*** of f to A_0.

Example. In the previous example, $h = g|_{\mathbb{R}_{\geq 0}}$. Note that the map h is bijective, but g is not. This is one reason why one should consider g and h to be different maps, although both of them are superficially the same in that both are given by squaring their argument.

0.3. Sums and products

0.3.1. Sums. Often one wants to sum many numbers. The basic notation for this is

$$\sum_{\text{subscript}}^{\text{superscript}} \text{expression}.$$

This asks one to sum the given expression, usually depending on some variable, where the variable and range of summation are specified by the superscripts and subscripts. One uses many variants of this, which will usually be obvious from the given context, so in place of trying to give a definition, we only provide examples.

Examples.

(1) $\sum_{i=1}^{4} i^2$ means $1^2 + 2^2 + 3^2 + 4^2$.

(2) $\sum_{j=0}^{4} j^2$ means $0^2 + 1^2 + 2^2 + 3^2 + 4^2$.

(3) $\sum_{k=2}^{4} 1$ means $1 + 1 + 1 = 3$.

(4) $\sum_{u=-3}^{3} u$ means $(-3) + (-2) + (-1) + 0 + 1 + 2 + 3 = 0$.

(5) $\sum_{i \in \{1,2\}} i$ means $1 + 2$.

(6) $\sum_{|i| \leq 1} i^2$ means $0^2 + (-1)^2 + 1^2$.

(7) $\sum_{0 < |i| \leq 1} i^3$ means $(-1)^3 + 1^3$.

Example. A result often ascribed to Gauß (when he was nine years old!), but known much earlier, is the following explicit evaluation of the sum of the first n positive integers:

$$S(n) := \sum_{k=1}^{n} k = \frac{n(n+1)}{2} \quad (\text{for } n = 1, 2, 3, \ldots). \tag{0.1}$$

One way of seeing this, is via the following computation. We start with the trivial identity

$$2S(n) = \sum_{k=1}^{n} k + \sum_{k=1}^{n} k.$$

In the last sum, we reverse the order of the terms, i.e., we replace $1+2+\ldots+(n-1)+n$ by $n+(n-1)+\ldots+2+1$, which evidently does not change the result. Thus,

$$2S(n) = \sum_{k=1}^{n} k + \sum_{k=1}^{n} (n+1-k) = \sum_{k=1}^{n} [k+(n+1-k)] = \sum_{k=1}^{n} (n+1) = n(n+1).$$

Upon dividing both sides of the equation by two, we obtain (0.1). The argument employed here may be illustrated by means of the following picture which illustrates the formula for small values of n:

$n=1$: , $n=2$:  , $n=3$:  , $n=4$:  , $\ldots$

(The number of squares is $2S(n)$, the total number of white squares or black squares representing one copy of $S(n)$ respectively. The squares build up a rectangle with n squares in the vertical and $n+1$ squares in the horizontal direction, thus $2S(n) = n(n+1)$.)

0.3.2. Infinite series. We shall also encounter various 'infinite sums' (called ***(infinite) series***), such as power series

$$\sum_{n=0}^{\infty} a_n x^n$$

(see § 0.5) or Fourier series

$$\sum_{k=-\infty}^{\infty} \hat{f}(k) e^{2\pi \mathrm{i}(k-\ell)x}$$

(see Chapter 4). The way in which one makes sense of these involve 'limits' and we refer to the relevant sections for the details.

Evaluating infinite series is, in general, a hard problem. We illustrate this using two examples, the first one being easy and the second one being rather more deep. (Deep enough, in fact, so that we do not hazard trying to give an explanation of this result.)

Example. One has

$$\sum_{n=1}^{\infty} \frac{1}{n(n+1)} = 1.$$

Indeed, one has the partial fraction decomposition (cf. § 2.5)

$$\frac{1}{n(n+1)} = \frac{1}{n} - \frac{1}{n+1},$$

so that

$$\sum_{n=1}^{N} \frac{1}{n(n+1)} = \sum_{n=1}^{N} \frac{1}{n} - \sum_{n=1}^{N} \frac{1}{n+1} = \frac{1}{1} - \frac{1}{N+1},$$

where the last equation follows by writing out both sums and seeing that most terms cancel, leaving behind only the first term from the first sum and the last term from the last sum. Thus,

$$\sum_{n=1}^{\infty} \frac{1}{n(n+1)} = \lim_{N\to\infty} \sum_{n=1}^{N} \frac{1}{n(n+1)} = \lim_{N\to\infty} \left(\frac{1}{1} - \frac{1}{N+1} \right) = 1 - 0 = 1.$$

Example (Euler, 1735)**.** The explicit evaluation of the series on the left hand side of the following equation was known as the ***Basel problem*** and was solved by Leonhard Euler in 1735, much to the amazement of his contemporaries:

$$\sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6}.$$

Today, many proofs of this result are known. We may encounter one using Fourier analysis in the course of the exercises, but at this point we leave it at that and only note that even innocuously looking infinite series can produce rather complicated values.[3]

0.3.3. Products. The discussion concerning sums applies equally to products with the obvious modifications. The basic notation for products is

$$\prod_{\text{subscript}}^{\text{superscript}} \text{expression}.$$

We just give one example of this, as comparison with the above examples for sums should be sufficient.

Example. $\prod_{i=1}^{4} i^2$ means $1^2 \cdot 2^2 \cdot 3^2 \cdot 4^2$.

Example. $\prod_{k=1}^{n} \frac{k+1}{k} = \frac{2}{1} \cdot \frac{3}{2} \cdots \frac{n}{n-1} \cdot \frac{n+1}{n} = \frac{n+1}{1} = n+1.$

[3]For those, who want to argue that the answer, $\pi^2/6$, does look rather neat and tidy: note that π is (or, at least, may be) defined as the ratio of any circle's circumference by its diameter. Thus, to find Euler's result, one ought to connect the inifnite sum over $1/n^2$ to circles somehow. It should be obvious, that this task is non-trivial.

0.4. Limits

Limits are an essential tool in analysis. They are used, for instance, to sum infinitely many terms (often used to pass from approximations to the 'real' object of interest by means of successively refining approximations), to define differentiability (say, in the one-dimensional setting, by passing from slopes of secant lines to slopes of tangents), or to successfully work with integrals (see, e.g., the arguments in Chapter 7 with regard to Gauß's theorem).

In practice, one encounters many different notions of limits. There are technical notions which allow one to treat most of these limits in convenient generality (e.g., so-called ***metric spaces***), but we do not introduce these for fear of making things too abstract. Consequently, our discussion here is somewhat clumsy in that it requires a lot of case work. In one further attempt at keeping things simple, we only discuss the one-dimensional real situation, although in these notes we definitely also require the case of higher-dimensions and the complex case. In both of these cases the distinction between $-\infty$ and $+\infty$ needs to be dropped, and for the higher-dimensional case one ought to replace occurences of $|\cdot|$ by $\|\cdot\|$.

Suppose that $f\colon X \to Y$ is a function with X and Y both being subsets of $\mathbb{R}$. Moreover suppose one of the following:

(1) $x_0 \in \mathbb{R}$, and for every $\delta > 0$ there exists some $x \in X$ such that $|x_0 - x| < \delta$.
(2) $x_0 = +\infty$, and for every $\delta > 0$ there exists some $x \in X$ such that $x > \delta^{-1}$.
(3) $x_0 = -\infty$, and for every $\delta > 0$ there exists some $x \in X$ such that $-x > \delta^{-1}$.

Roughly speaking, these conditions ensure that one can get arbitrarily close to x_0 from within the set X. In the last two cases, getting 'arbitrarily close to $\pm\infty$' from within X means that X ought to contain arbitrarily large elements (while taking signs into account suitably). In each case, the set of $x \in X$ for which the last inequality specified in the case holds, we shall call a ***δ-neighbourhood of x_0 for X*** (caution: this is not standard terminology). The inclusion of the inversion (δ^{-1} instead of δ) in the last two cases is there only for cosmetic reasons, as one usually likes to think of taking δ smaller when trying to select smaller neighbourhoods.

Example. Let f and x_0 be as above. Suppose that f is constant, i.e., there exists some $y_* \in Y$ such that for every $x \in X$ one has $f(x) = y_*$. Then $\lim_{x \to x_0} f(x) = y_*$. (The readers should try to show this for themselves.)

Example. Consider a sequence $(a_n)_{n\in\mathbb{N}}$, i.e., a function $a\colon \mathbb{N} \to \mathbb{R}$. Then $\mathbb{N}$, the domain of definition of a, contains arbitrarily large (positive) numbers, so we may consider the case $x_0 = +\infty$ in the above. For $\pi = 3.1415\ldots$, a π^{-1}-neighbourhood of x_0 is given by $\{4, 5, 6, \ldots\} \subseteq \mathbb{N}$. On the other hand, a $(1/100)$-neighbourhood of x_0 is given by $\{101, 102, 103, \ldots\} \subseteq \mathbb{N}$.

Example. Consider a sequence $(a_n)_{n\in\mathbb{N}}$. Then $x_0 \in \mathbb{R}$ satisfies case (1) (with $X = \mathbb{N}$) if and only if $x_0 \in \mathbb{N}$. Moreover, in that case, any δ-neighbourhood of x_0 consists

only of x_0 provided that $0 < \delta \leq 1$. For instance, the 2-neighbourhood of $x_0 = 5$ is $\{4, 5, 6\}$, while the 1-neighbourhood of x_0 is $\{5\}$.

Example. Consider a function $f : \mathbb{R} \to \mathbb{R}$. Then the δ-neighbourhood of $x_0 \in \mathbb{R}$ is the interval $(x_0 - \delta, x_0 + \delta)$.

Suppose now that f and x_0 are as above, satisfying one of the cases (1)–(3). If there exists some $y_0 \in Y \cup \{-\infty, +\infty\}$ such that for every $\epsilon > 0$ one finds a $\delta > 0$ such that for every x in the δ-neighbourhood of x_0 one has that $f(x)$ lies in the ϵ-neighbourhood of y_0 for $\mathbb{R}$, then one calls y_0 the ***limit of $f(x)$ as $x \to x_0$***. (One can show that there exists at most one y_0 satisfying the above, so speaking of *the* limit rather than *a* limit is justified.) Roughly speaking, the definition of y_0 being the limit of $f(x)$ as $x \to x_0$ requires that, for every $\epsilon > 0$, there exists a δ-neighbourhood of x_0 for X whose image under f is contained in the ϵ-neighbourhood of y_0 for $\mathbb{R}$.

When y_0 satisfying the above exists *and* $y_0 \in Y$ (i.e., $y_0 \neq \pm\infty$), then one says that $f(x)$ ***converges*** to y_0. Otherwise (i.e., when no y_0 as above exists, or if such a y_0 exists, but $y_0 = +\infty$ or $y_0 = -\infty$), then one says that $f(x)$ ***diverges***. If $y_0 = +\infty$ or $y_0 = -\infty$ is the limit of $f(x)$ as $x \to y_0$, then one says that $f(x)$ ***diverges to $\pm\infty$*** (with appropriate choice of sign). In the case of convergence, one also says that 'the limit of $f(x)$ as $x \to x_0$ exists'.

Remark. While we have tried to treat limits $x \to x_0 \in \mathbb{R}$ and limits $x \to \pm\infty$ in a unified fashion, the distinction between convergence and divergence, and a limit existing or not, introduced here is utterly and bewilderingly confusing. This distinction is, however, somewhat justified, because arithmetic with limits works much cleaner when one has convergence (see below).

Example. Consider the sequence $(1/n)_n$, represented by the function $f : \mathbb{N} \to \mathbb{R}$, $n \mapsto 1/n$. Then

$$\lim_{n\to\infty} (1/n) = 0.$$

To see this, note that for every $\epsilon > 0$ the ϵ-neighbourhood of ∞ (for $\mathbb{N}$) is $\{n_\epsilon, n_\epsilon + 1, n_\epsilon + 2, \ldots\} \subseteq \mathbb{N}$, where n_ϵ is the smallest positive integer satisfying

$$n_\epsilon > \epsilon^{-1}. \tag{0.2}$$

Take $\delta = \epsilon$. Then the δ-neighbourhood of 0 (for $\mathbb{R}$) is the interval $(-\delta, \delta)$. Now it suffices to observe that, due to (0.2), every n in the ϵ^{-1}-neighbourhood of ∞ (for $\mathbb{N}$) satisfies $|f(n)| = |1/n| = 1/n < \epsilon = \delta$. Hence, $f(n)$ is contained in the δ-neighbourhood of 0 (for $\mathbb{R}$). Therefore, 0 is the limit of $f(n) = 1/n$ as $n \to \infty$.

Example. Consider the function $f : \mathbb{R}_+ \to \mathbb{R}$, $x \mapsto 1/x$. Then

$$\lim_{x\to\infty} (1/x) = 0.$$

The justification for this equation is basically the same as in the previous example; the ϵ-neighbourhood now needs to be computed inside $\mathbb{R}_+$ rather than $\mathbb{N}$, but the underlying argument using inequalities remains unchanged.

In practice, one basically never computes limits by appealing to the definition. Instead, one uses theorems, which allow one to break up the computation of the desired limit into the computation of limits that one already knows the value of. For instance, let $f\colon X \to Y$ and $g\colon X' \to Y'$ be two functions, $X, X', Y, Y' \subseteq \mathbb{R}$ and suppose that $x \in \mathbb{R} \cup \{-\infty, +\infty\}$ satisfies one of the cases (1)–(3) with $X \cap X'$ taking the role of X in the cases. We define $f+g$ and $f \cdot g$ on $X \cap X'$ in a point-wise fashion, i.e., $(f+g)(x) := f(x) + g(x)$ and $(f \cdot g)(x) := f(x) \cdot g(x)$. Then, provided that $f(x)$ and $g(x)$ both converge as $x \to x_0$, also $(f+g)(x)$ converges as $x \to x_0$, and one has

$$\lim_{x \to x_0} (f+g)(x) = \big(\lim_{x \to x_0} f(x)\big) + \big(\lim_{x \to x_0} g(x)\big).$$

A similar statement holds for $f \cdot g$. Moreover, one can also prove a similar statement for f/g (again, defined point-wise), provided that g does not vanish on some δ-neighbourhood of x_0 for X'. (Here the last condition is to make sense of the division.)

Example. Write lim for $\lim_{n\to\infty}$ for brevity's sake. Then

$$\begin{aligned}\lim \frac{n^2+2}{n^2-1} &= \lim \frac{n^2(1+2/n^2)}{n^2(1-1/n^2)} = \lim \frac{1+2/n^2}{1-1/n^2} = \frac{\lim(1+2/n^2)}{\lim(1-1/n^2)} \\ &= \frac{(\lim 1) + (\lim[2/n^2])}{(\lim)1 + (\lim[-1/n^2])} = \frac{(\lim 1) + (\lim 2)(\lim[1/n])(\lim[1/n])}{(\lim 1) + (\lim[-1])(\lim[1/n])(\lim[1/n])} \\ &= \frac{1+2\cdot 0\cdot 0}{1+(-1)\cdot 0\cdot 0} = 1,\end{aligned}$$

where the chain of equalities actually ought to be read 'from bottom to top' as to justify the existence of the involved limits in the process. To go from the last line up, we have used our previous discussions of limits of constant functions and of the sequence $(1/n)_n$.

Example. Care must be taken with divergence to $\pm\infty$. For instance, writing lim for $\lim_{n\to\infty}$ for brevity's sake, once more, consider the following two lines:

$$\text{`}0 = \lim(n-n) = (\lim n) - (\lim n) = \infty - \infty\text{'},$$

and

$$\text{`}\infty = \lim(2n-n) = (\lim 2n) - (\lim n) = \infty - \infty\text{'}.$$

These lines show that care ought to be taken when dealing with divergence. The quotation marks are meant to indicate that anyone who carelessly writes such nonsense will have their maths license revoked.

Let $f\colon X \to Y$ be a function with $X, Y \subseteq \mathbb{R}$ and $x_0 \in X$. Then f is called ***continuous at x_0*** if

$$\lim_{x \to x_0} f(x) = f(x_0).$$

Moreover, f is called ***continuous*** if it is continuous at every point $x_0 \in X$. Upon noting that $\lim_{x\to x_0} x = x_0$, we see that the above may be re-written as

$$\lim_{x\to x_0} f(x) = f\big(\lim_{x\to x_0} x\big).$$

This shows that continuity is a sort of compatibility condition between functions and limits. The following proposition should indicated the one has continuity in many reasonable cases. (Much more can [and should] be said, but we lack the capacity to do so here.)

Proposition 0.3. *The following functions are continuous:*

(1) polynomial functions (defined on $\mathbb{R}$),

(2) $\exp, \sin, \cos\colon \mathbb{R} \to \mathbb{R}$.

Proof. Omitted, as it would occupy us unduly. □

Example. Consider the sequence $(\exp(1/n))_n$. We already know that $1/n$ converges to 0 as $n \to \infty$, so by continuity of the exponential function, we find that

$$\lim_{n\to\infty} \exp(1/n) = \exp(0) = 1.$$

Example. The function $H\colon \mathbb{R} \to \mathbb{R}$ defined by $H(x) = 1$ for $x \geq 0$ and $H(x) = 0$ for $x < 0$ is continuous at every point $x_0 \in \mathbb{R} \setminus \{0\}$, but not continuous at $x_0 = 0$.

0.5. Power series

Polynomials are an easy source of functions and much time in the mathematics education in school is concerned with attaining some understanding of these functions. Mathematicians like polynomials, because they obey very rigid structural rules and are, therefore, easy to work with; 'easier' than most other functions, that is. Things liked by people tend to pop up at various places. We shall see later, in § 6.2, that many other functions can also be approximated using polynomials. Polynomials have the form

$$a_0x^0 + a_1x^1 + \ldots + a_nx^n = \sum_{k=0}^{n} a_kx^k,$$

where the a_k are coefficients (real numbers or complex numbers, for instance) and 'x' is either some variable that waits for being substituted by some number or already some number we have decided on.

If one approximates a function by polynomials

$$\sum_{k=0}^{n} a_kx^k, \quad \sum_{k=0}^{m} b_kx^k, \quad \sum_{k=0}^{r} c_kx^k, \quad \ldots,$$

the coefficients a_k, b_k, c_k, … need not have anything to do with each other. However, one particularly nice case arises when the above are

$$a_0x^0, \quad a_0x^0 + a_1x^1, \quad a_0x^0 + a_1x^1 + a_2x^2, \quad \ldots .$$

This produces so-called ***power series***

$$\sum_{k=0}^{\infty} a_k x^k = \lim_{n\to\infty} \sum_{k=0}^{n} a_k x^k, \tag{0.3}$$

where such a series denotes both the left hand side interpreted as a formal expression where x is viewed a variable, and the *value* produced by the limit on the right if x is given some numeric value. The limit on the right hand side may or may not exist for $x \neq 0$. If it does, the power series is said to ***converge*** at x. Otherwise it is said to ***diverge*** at x. We illustrate this by an example which is probably the easiest power series (except for polynomials) one can write down, that is, the power series obtained by setting all coefficients a_k to 1.

Example. For any real x we have

$$\begin{aligned}(1-x)(1+x+x^2) &= (1+x+x^2) - x(1+x+x^2)\\ &= 1+x+x^2-x-x^2-x^3.\end{aligned}$$

Here all but the 'outer' two terms cancel and the result is $1-x^3$. Generalising this, we find that, for $n = 1, 2, 3, 4, \ldots$, we have

$$\begin{aligned}(1-x)(x^0+x^1+\ldots+x^n) &= (x^0+x^1+\ldots+x^n) - x(x^0+\ldots+x^{n-1}+x^n)\\ &= x^0 + \underline{x^1} + \ldots + \underline{\underline{x^n}} - \underline{x^1} - \ldots - \underline{\underline{x^n}} - x^{n+1}\\ &= x^0 - x^{n+1}.\end{aligned}$$

If $x \neq 1$, we may divide by $1-x$ and obtain (noting that $x^0 = 1$)

$$\sum_{k=0}^{n} x^k = \frac{1-x^{n+1}}{1-x}.$$

(The left hand side of the above is called a ***geometric sum***.) We have

$$\lim_{n\to\infty} \frac{1-x^{n+1}}{1-x} \begin{cases} = \dfrac{1}{1-x} & \text{if } |x| < 1,\\ = \text{divergent} & \text{if } |x| \geq 1,\ x \neq 1.\end{cases}$$

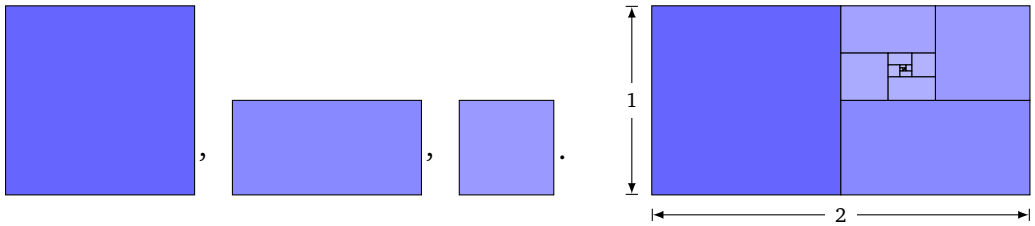
Consequently, the infinite series (called a ***geometric series***)

$$\sum_{k=0}^{\infty} x^k$$

has the value $1/(1-x)$ for $|x| < 1$. For $|x| \geq 1$ the limit

$$\lim_{n\to\infty} \sum_{k=0}^{n} x^k$$

diverges. (For $x \neq 1$ this follows from the above, but for $x = 1$ it is also clear, because $\sum_{k=0}^{n} 1^k = n+1$ and this diverges, too, as $n \to \infty$.)



(a) The first three pieces. **(b)** All pieces together.

Figure 2. Geometrical reasoning for $1+1/2+1/4+1/8+\ldots=2$.

Example. Letting $x=1/2$ in the power series from the previous example, we obtain

$$\sum_{k=0}^{\infty}(1/2)^k=\frac{1}{1-1/2}=2.$$

One can give a geometric 'proof' of this fact. We look at a rectangle (square) all sides having length 1. It has area 1. Now we look at the rectangle obtained by halving the height of our square. It has area $1/2$. The rectangle obtained from this after halving the width has area $1/4$ (see Figure 2 (a)). We continue this process to infinity and ask for the total area of all rectangles thus obtained. Clearly the answer ought to be

$$1+(1/2)+(1/4)+(1/8)+\ldots=\sum_{k=0}^{\infty}(1/2)^k.$$

Figure 2 (b) suggests that this answer should indeed be 2.

There is a full theory regarding the convergence of power series. We discuss none of this here. We only mention a few power series that one should know.

0.5.1. Exponential function. First, one should know that the ***exponential function*** $\exp\colon\mathbb{R}\to\mathbb{R}$ is given by

$$\exp(x)=\sum_{k=0}^{\infty}\frac{1}{k!}x^k.$$

(One can show that the limit defining the value of this power series exists for all real numbers x.) It satisfies

- $\exp(x+y)=\exp(x)\exp(y)$, (***functional equation*** of exp)
- $\exp(-x)=\exp(x)^{-1}$,
- $\exp(x)>0$,
- $\exp(0)=1$,
- $\exp'(x)=\exp(x)$, (exp is its own derivative)

for all $x,y\in\mathbb{R}$. (A more general version, using complex numbers, is given by Theorem 1.3.)

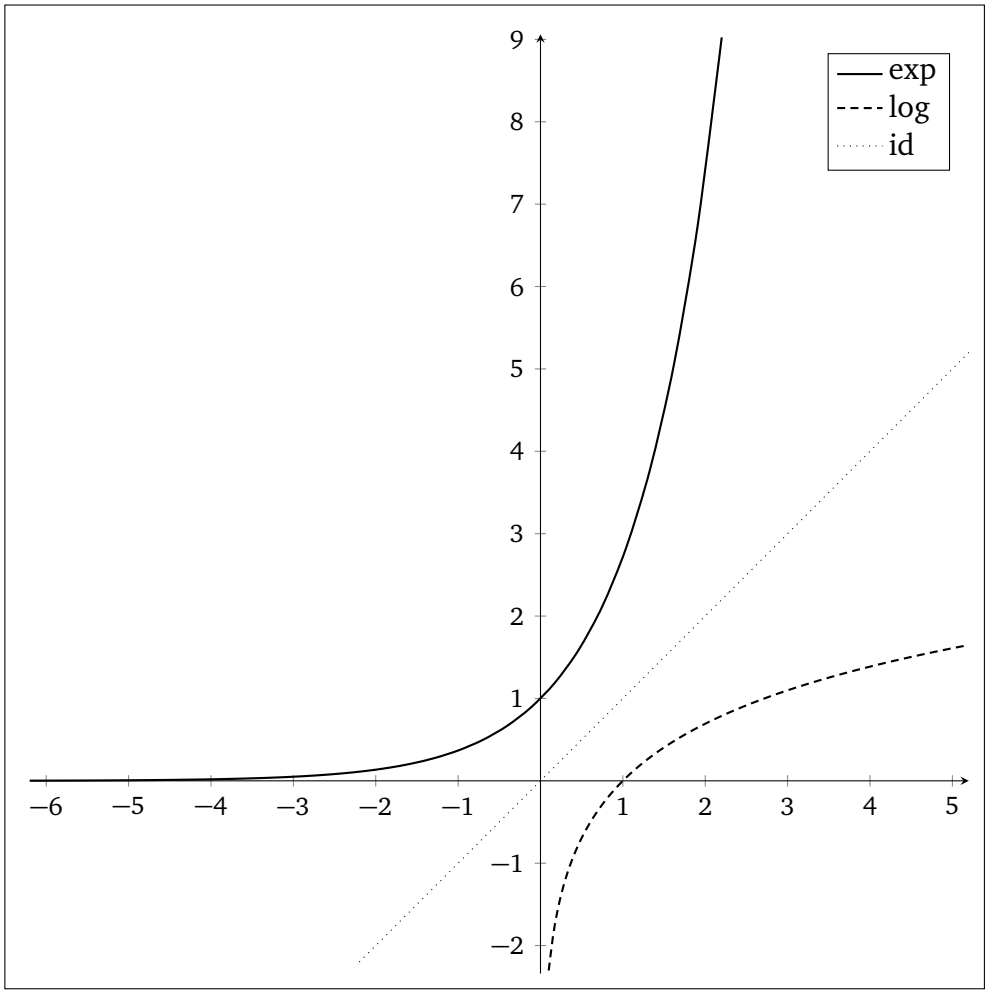


Figure 3. Graph of the exponential function and its inverse, the logarithm function. Observe the symmetry of both graphs with respect to the graph of $x \mapsto x$. (This feature is characteristic for plotting a real-valued function and its inverse.)

0.5.2. Sine and cosine. Also the ***sine and cosine functions*** can be written as power series. For $x \in \mathbb{R}$ one has

$$\sin(x) = \sum_{m=0}^{\infty} \frac{(-1)^m}{(2m+1)!} x^{2m+1} \quad \text{and} \quad \cos(x) = \sum_{m=0}^{\infty} \frac{(-1)^m}{(2m)!} x^{2m}.$$

Here the exponent of x does not match the form given in (0.3), but we may write

$$\sin(x) = \sum_{k=0}^{\infty} a_k x^k, \quad \text{where} \quad a_k = \begin{cases} \dfrac{(-1)^{(k-1)/2}}{k!} & \text{if } k \text{ is odd,} \\ 0 & \text{if } k \text{ is even,} \end{cases}$$

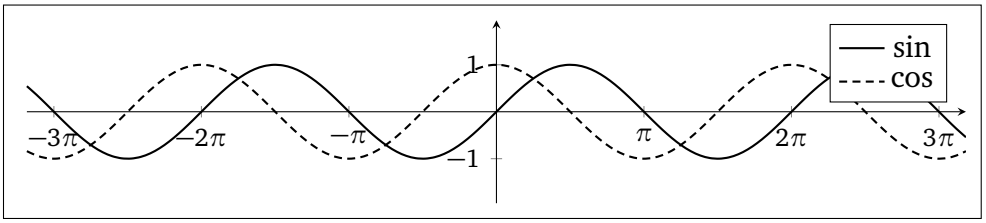


Figure 4. Graphs of the sine and cosine functions.

which is of the form (0.3).

0.5.3. Natural logarithm. There is a unique function $f : (0, \infty) \to \mathbb{R}$ such that $f \circ \exp = \mathrm{id}_{\mathbb{R}}$ and $\exp \circ f = \mathrm{id}_{(0,\infty)}$, called the ***(natural) logarithm function***. It is denoted by 'log' (sometimes also by 'ln'). One can show that there is no power series of the form (0.3) that represents the logarithm function. One can, however represent $x \mapsto \log(1+x)$ by a power series:

$$\log(1+x) = \sum_{k=1}^{\infty} \frac{(-1)^{k+1}}{k} x^k \quad (\text{for } |x| < 1).$$

The restriction to $|x| < 1$ is perhaps not too surprising, because as $x > -1$ approaches -1, the term $\log(1+x)$ diverges to $-\infty$ (compare § 0.5.1), so clearly something strange happens here. In fact, the power series on the right hand side of the above equation can be seen to diverge for any x with $|x| > 1$ and for $x = -1$. (For $x = 1$ it converges to the limit $\log 2$, but we prove none of these claims.) From the above, and using the functional equation for the logarithm[4]

$$\log(xy) = \log(x) + \log(y) \quad (x, y > 0),$$

one can deduce that

$$\log(r+x) = \log(r) + \log(1 + x/r) = \log(r) + \sum_{k=1}^{\infty} \frac{(-1)^{k+1}}{r^k k} x^k \quad (\text{for } |x| < r).$$

Hence, by choosing r large, one can obtain a power series representation for (a shifted version of) the logarithm function which is valid for a wide range of x.

Remark (About working with power series)**.** One should not be afraid of power series. In some sense, they are the nicest functions one can possibly encounter. For all practical purposes, the reader may think of power series as 'just polynomials'. Most of the things the reader would feel comfortable doing with polynomials, like adding them,[5]

$$\sum_{k=0}^{n} a_k x^k + \sum_{k=0}^{m} b_k x^k = \sum_{k=0}^{\max\{n,m\}} (a_k + b_k) x^k$$

[4]This can be deduced, for instance, from the functional equation of the exponential function.

[5]For $n > m$ we assume that $b_k = 0$ for $k > m$ and similarly with a_k instead of b_k when $n < m$.

generally also works for power series:

$$\sum_{k=0}^{\infty} a_k x^k + \sum_{k=0}^{\infty} b_k x^k = \sum_{k=0}^{\infty} (a_k + b_k) x^k. \tag{0.4}$$

Easy examples like

$$\underbrace{\sum_{k=0}^{\infty} (-1) x^k}_{\text{not convergent for } x \neq 0} + \underbrace{\sum_{k=0}^{\infty} (+1) x^k}_{\text{not convergent for } x \neq 0} \stackrel{?}{=} \underbrace{\sum_{k=0}^{\infty} 0 x^k}_{\text{convergent for all } x}$$

show that this is *not really* true in general, but one can show that (0.4) holds for those x for which both power series on the left hand side of (0.4) converge. Therefore, some comments concerning convergence of power series would be in order. We do, however, make no further such comments here.

0.6. Differentiation

(This section is essentially identical to § 1.4.)

Given a real-valued function f defined on a neighbourhood[6] of a point $x_0 \in \mathbb{R}$ one defines its derivative as

$$f'(x_0) := \frac{\mathrm{d}f}{\mathrm{d}x}(x_0) := \lim_{\substack{h \to 0 \\ h \neq 0}} \frac{f(x_0 + h) - f(x_0)}{h} \quad \text{(if the limit exists!)}.$$

One has the rules

- $\dfrac{\mathrm{d}}{\mathrm{d}x}(\lambda f(x) + \mu g(x)) = \lambda f'(x) + \mu g'(x),$ (linearity)
- $\dfrac{\mathrm{d}}{\mathrm{d}x}(f(x)g(x)) = f'(x)g(x) + f(x)g'(x),$ (***product rule***)
- $\dfrac{\mathrm{d}}{\mathrm{d}x}\dfrac{f(x)}{g(x)} = \dfrac{f'(x)g(x) - f(x)g'(x)}{g(x)^2},$ (***quotient rule***)
- $\dfrac{\mathrm{d}}{\mathrm{d}x} f(g(x)) = f'(g(x))g'(x).$ (***chain rule***)

(The above rules are subject to "the obvious" assumptions, i.e., differentiability of f at the point $g(x)$ and that of g at x for the chain rule, f and g being differentiable at x and $g(x) \neq 0$ for the quotient rule etc.) Polynomials are differentiated as one expects as well:

$$\frac{\mathrm{d}}{\mathrm{d}x} x^n = n x^{n-1}. \tag{0.5}$$

(This formula also holds for every integer $n \neq 0$.)

[6] This means that f is defined on some interval $(x_0 - \epsilon, x_0 + \epsilon)$ about x_0 for some (potentially very small) $\epsilon > 0$.

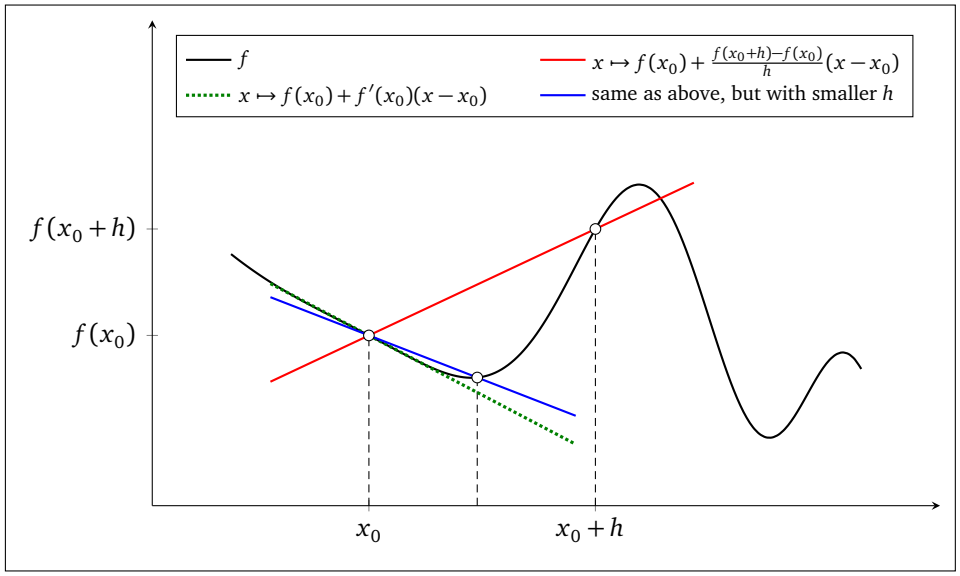


Figure 5. Visually speaking, computing the derivative $f'(x_0)$ at a point x_0 means to compute the *slope* of the tangent line (green, dotted) to the graph of f at the point $(x_0, f(x_0))$. This slope is approximated using the secant through the points $(x_0, f(x_0))$ and $(x_0 + h, f(x_0 + h))$ (red and blue lines); the parameter h is then sent to zero using a limiting process.

Example. For any non-zero number x we have

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}x}\frac{x^2+x-3}{x^2} &= \frac{(x^2+x-3)'x^2-(x^2+x-3)(x^2)'}{(x^2)^2}\\ &= \frac{(2x+1)x^2-(x^2+x-3)(2x)}{x^4}\\ &= \frac{2x^3+x^2-(2x^3+2x^2-6x)}{x^4} = \frac{6-x}{x^3}.\end{aligned}$$

Alternatively, one can use linearity first, getting

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}x}\frac{x^2+x-3}{x^2} &= \frac{\mathrm{d}}{\mathrm{d}x}\frac{x^2}{x^2} + \frac{\mathrm{d}}{\mathrm{d}x}\frac{x}{x^2} - 3\frac{\mathrm{d}}{\mathrm{d}x}\frac{1}{x^2}\\ &= \frac{\mathrm{d}}{\mathrm{d}x}1 + \frac{\mathrm{d}}{\mathrm{d}x}x^{-1} - 3\frac{\mathrm{d}}{\mathrm{d}x}x^{-2}\\ &= (-1)x^{-2} - 3(-2)x^{-3} = \frac{6-x}{x^3}.\end{aligned}$$

Another important rule for differentiation is the rule for differentiation of inverse functions. If $f\colon U \to V$ is a bijective (1:1) function between two sets $U, V \subseteq \mathbb{C}$ (or $U, V \subseteq \mathbb{R}$ in the real case), then there exists a (unique) function $f^{-1} : V \to U$

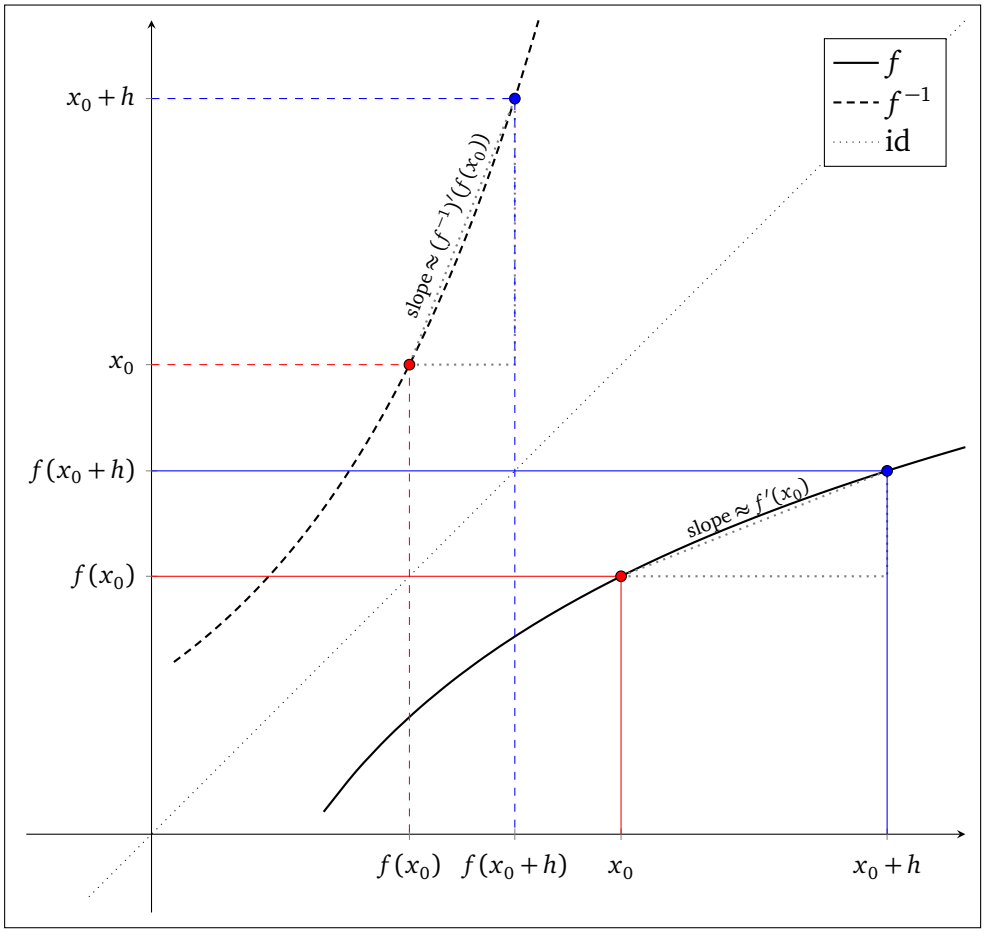


Figure 6. Illustration of (0.6). Visually, $f'(x_0)$ is the slope of the tangent at $(x_0, f(x_0))$ to the graph of f. One computes this slope using a limiting argument with secants; (a part of) each such secant can be found as the hypotenuse of a suitable triangle as drawn in the above picture. It turns out that, when working to compute $(f^{-1})'(f(x_0))$, the same triangles show up, but *flipped*. Consequently, this inverts the slope of the secant; thus (0.6).

such that $f^{-1} \circ f = \mathrm{id}_U$ and $f \circ f^{-1} = \mathrm{id}_V$. Moreover, if any point of U admits a neighbourhood contained in U and f is differentiable, then so is f^{-1}. Furthermore, given the derivative of f, the derivative of f^{-1} is also easy to find: we have

$$(f^{-1})'(y) = \frac{1}{f'(f^{-1}(y))}, \tag{0.6}$$

if $f'(f^{-1}(y))$ is non-zero. (Provided that one already believes in the differentiability of f^{-1}, one can arrive at the above formula by applying the chain rule to both sides

of $f^{-1} \circ f = \mathrm{id}_U$ and rearranging the resulting expression.) We give three examples of this differentiation rule.

Example (Derivative of the square root)**.** The function $f\colon \mathbb{R}_{>0} \to \mathbb{R}_{>0}$, $x \mapsto x^2$, is bijective. Its inverse function is given by $f^{-1}\colon \mathbb{R}_{>0} \to \mathbb{R}_{>0}$, $y \mapsto \sqrt{y}$. Using $\sqrt{y} = y^{1/2}$ and the differentiation rule (0.5), we have

$$\frac{\mathrm{d}}{\mathrm{d}y}\sqrt{y} = \frac{1}{2}y^{1/2-1} = \frac{1}{2\sqrt{y}}.$$

On the other hand, we can also use the aforementioned rule for finding the derivative of f^{-1}. Indeed,

$$\frac{\mathrm{d}}{\mathrm{d}y}\sqrt{y} = \frac{1}{f'(f^{-1}(y))} = \frac{1}{2f^{-1}(y)} = \frac{1}{2\sqrt{y}}.$$

The previous example is meant primarily for illustration's sake, since (as we have pointed out above) the desired result also follows from (0.5). The next two examples may seem more interesting.

Example (Derivative of the natural logarithm)**.** We wish to find the derivative of the natural logarithm $\log\colon \mathbb{R}_{>0} \to \mathbb{R}$. From school one should already know that $\log'(y) = 1/y$. To derive this, we observe that the natural logarithm is obtained as inverse function to the exponential function. Hence,

$$\log' y = \frac{1}{\exp'(\log y)} = \frac{1}{\exp(\log y)} = \frac{1}{y}.$$

Example (Derivative of the arcus tangent)**.** We wish to find the derivative of

$$\arctan\colon \mathbb{R} \to (-\pi/2, \pi/2),$$

the inverse function of $\tan : (-\pi/2, \pi/2) \to \mathbb{R}$. In preparation for this, we first find the derivative of the tangent function. We have

$$\tan'(x) = \frac{\mathrm{d}}{\mathrm{d}x}\frac{\sin x}{\cos x} = \frac{(\sin' x)(\cos x) - (\sin x)(\cos' x)}{(\cos x)^2} = \frac{(\cos x)^2 + (\sin x)^2}{(\cos x)^2}.$$

Using $(\cos x)^2 + (\sin x)^2 = 1$, the above expression could now be simplified to $\tan'(x) = 1/(\cos x)^2$. However, this is less useful for what we have in mind. Instead, we write

$$\tan'(x) = \frac{(\cos x)^2}{(\cos x)^2} + \frac{(\sin x)^2}{(\cos x)^2} = 1 + (\tan x)^2.$$

Hence,

$$\arctan'(y) = \frac{1}{\tan'(\arctan y)} = \frac{1}{1 + (\tan(\arctan y))^2} = \frac{1}{1+y^2}.$$

The above example is interesting, because arctan is arguably rather more complicated than its derivative, which happens to be a rational function (a quotient of polynomials). This shows two things: first, the derivative of a complicated function can be quite easy. Second, and more crucially, integrals of easy looking functions may turn out to be quite more delicate. This hints that computing integrals often requires some sort of ingenuity in comparison to differentiation.

0.7. Integration

We give a quick review of integration, because it plays a prominent rôle in Chapter 2 and Chapter 7 when working on examples. Our goal here is to recall some techniques for *computing* integrals. A more thorough discussion of integrals will be given in Chapter 7. We assume that the reader already knows what the integral of a (sufficiently well-behaved) real-valued function $f\colon [a,b] \to \mathbb{R}$ is.

A reader familiar with complex numbers[7] (see Chapter 1) may also wish to integrate complex-valued functions. This is done as follows. Let $f\colon [a,b] \to \mathbb{C}$ be a sufficiently well-behaved, complex-valued function. Its integral is defined via the formula

$$\int_a^b f(x)\,\mathrm{d}x := \int_a^b (\operatorname{Re} f(x))\,\mathrm{d}x + \mathrm{i}\int_a^b (\operatorname{Im} f(x))\,\mathrm{d}x.$$

It should be noted, however, the most of the time one does not need to remember this definition, as integrals are usually computed by other techniques. (It turns out that these techniques generally work equally well for complex as for real-valued functions.)

Later, in § 7.6, we shall attach some meaning to the symbol '$\mathrm{d}x$'. At this stage, however, the variable x is completely arbitrary and can be interchanged for any other unused symbol:

$$\int_a^b f(x)\,\mathrm{d}x = \int_a^b f(t)\,\mathrm{d}t = \int_a^b f(\theta)\,\mathrm{d}\theta = \int_a^b f(\checkmark)\,\mathrm{d}\checkmark.$$

(Okay, the last one is perhaps a bit too silly.)

Integration is *linear*. This means that, for any two (sufficiently well-behaved) functions $f, g\colon [a,b] \to \mathbb{C}$, and numbers $\lambda, \mu \in \mathbb{C}$, one has

$$\int_a^b (\lambda f(x) + \mu g(x))\,\mathrm{d}x = \lambda \int_a^b f(x)\,\mathrm{d}x + \mu \int_a^b g(x)\,\mathrm{d}x.$$

[7] A reader who is not familiar with complex numbers should just skip this paragraph and replace all occurrences of $\mathbb{C}$ by $\mathbb{R}$ in the subsequent discussion.

0.7.1. Anti-derivatives. We now vary the upper integration limit b. More precisely, we shall look at the function

$$F\colon (a,\xi) \to \mathbb{R}, \quad \xi \mapsto \int_a^\xi f(x)\,\mathrm{d}x.$$

If f is 'sufficiently nice' (*continuous*; we shall return to this notion in Chapter 5), then one can show that F is differentiable, and one has

$$\begin{aligned} F'(\xi) &= \lim_{h\to\infty} \frac{F(\xi+h)-F(\xi)}{h} \\ &= \lim_{h\to\infty} \frac{1}{h}\left(\int_a^{\xi+h} f(x)\,\mathrm{d}x - \int_a^\xi f(x)\,\mathrm{d}x\right) \\ &= \lim_{h\to\infty} \frac{1}{h}\int_\xi^{\xi+h} f(x)\,\mathrm{d}x \\ &= f(\xi). \end{aligned}$$

(The last equation is for what we need f to be nice and the next-to-last equation uses a property of integrals which we have not discussed.) The above shows that, given a nice function f, integration can be used to produce a function whose derivative is f. Any such function is called an anti-derivative and one often writes

$$F(x) = \int f(x)\,\mathrm{d}x$$

for this (without upper and lower bounds following the integral sign) and sometimes also writes

$$F(x) = \int f(x)\,\mathrm{d}x + \text{const.}, \quad \text{or} \quad F(x) = \int f(x)\,\mathrm{d}x + c$$

instead. The latter is done to draw attention to the fact that if F is a function with derivative f, then so is the function $F + c$ given by $x \mapsto F(x) + c$ for any fixed c. There are some notational issues with this. Jänich [5, pp. 59–59] goes on a nice rant regarding this; see also [5, § 2.2].

0.7.2. Fundamental theorem of calculus. One of the most important tools for computing integrals is the ***fundamental theorem of calculus***. It states that for a differentiable function $F\colon [a,b] \to \mathbb{C}$ we have

$$\int_a^b F'(x)\,\mathrm{d}x = F(b) - F(a) =: F(x)\Big|_{x=a}^{b} =: F\Big|_a^b.$$

Example. For $n \in \mathbb{R} \setminus \{-1\}$ we have

$$\int_a^b x^n\,\mathrm{d}x = \int_a^b \left(\frac{\mathrm{d}}{\mathrm{d}x}\frac{1}{n+1}x^{n+1}\right)\mathrm{d}x = \frac{1}{n+1}b^{n+1} - \frac{1}{n+1}a^{n+1}. \tag{0.7}$$

Example. The following is to make a general point. Observe that

$$\frac{\mathrm{d}}{\mathrm{d}x}\frac{1}{n+1}x^{n+1} = x^n, \quad \text{but also} \quad \frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{1}{n+1}x^{n+1}+42\right) = x^n.$$

In fact, 42 could be replaced by an arbitrary constant c. (Even more generally, if $F' = f$, then also $F+c\colon x \mapsto F(x)+c$ is a function with $(F+c)' = f$.) Hence, looking at the previous example, we may compute

$$\begin{aligned}\int_a^b x^n\,\mathrm{d}x &= \int_a^b \frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{1}{n+1}x^{n+1}+42\right)\mathrm{d}x \\ &= \left(\frac{1}{n+1}b^{n+1}+42\right)-\left(\frac{1}{n+1}a^{n+1}+42\right).\end{aligned}$$

Here both constants 42 cancel each other due to the opposing signs, and we recover the result from (0.7).

Example. In the previous example we had to exclude the case $n=-1$. We shall now consider what happens in this case. For $x \neq 0$ one has

$$\frac{\mathrm{d}}{\mathrm{d}x}\log|x| = \frac{1}{x},$$

as one can see by considering the cases $x<0$ (then $|x| = -x$) and $x>0$ (then $|x| = x$) separately. Consequently, for any interval $[a,b]$ that does not contain 0 we have

$$\int_a^b \frac{1}{x}\,\mathrm{d}x = \log|b| - \log|a| = \log\frac{|b|}{|a|}.$$

Example. For $k \neq 0$ we have

$$\int_a^b \exp(kx)\,\mathrm{d}x = \int_a^b \left(\frac{\mathrm{d}}{\mathrm{d}x}\frac{1}{k}\exp(kx)\right)\mathrm{d}x = \frac{1}{k}\exp(kb) - \frac{1}{k}\exp(ka).$$

Example. Using linearity, we can also integrate polynomials. For example,

$$\int_0^1 (6x^2-8x)\,\mathrm{d}x = 6\int_0^1 x^2\,\mathrm{d}x - 8\int_0^1 x\,\mathrm{d}x = 6\frac{1}{3}x^3\Big|_{x=0}^1 - 8\frac{1}{2}x^2\Big|_{x=0}^1 = 2-4 = -2.$$

0.7.3. Integration via substitution. Recall the formula for ***integration via substitution***:

$$\int_{\varphi(a)}^{\varphi(b)} f(x)\,\mathrm{d}x = \int_a^b f(\varphi(y))\varphi'(y)\,\mathrm{d}y.$$

As it turns out, a generalisation of the above formula to higher-dimensions shall underpin much of our work in Chapter 7 (see § 7.2) when we discuss the theory of integration in more depth. Incidentally, integration via substitution can be a very powerful technique for evaluating integrals, but usually its power stems from choosing φ in a fashion tailored to possible "symmetries" of f. However, unearthing such

symmetries often requires some insights into the problem at hand and insight seldom presents itself willingly. For the reason, applications of integration via substitution often seem quite magical: the reader is presented with a pre-chosen function φ and everything falls nicely into place. The much harder question as to how one would go about selecting "the right" φ from the myriad of admissible functions is often left unanswered. Here we just mention that,—apart from experience and a few "well-known" examples that one may remember—in applications, it often seems that the problem that gave rise to the desire of computing an integral of some function f may serve to provide hits as to what substitutions would seem to address inherent symmetries correctly. Suffice it to say, that at this point we do not have the time to spare to consider any such examples in depth. Instead, we just give a very simple example in the sense that the choice of φ should not be too surprising here. Nonetheless, this computation will reappear in Example 7.3, when we compute an integral which will have—by then—appeared in related form as the Dirichlet integral during the proof of Theorem 2.2 and in § 2.4 as a normalising factor when discussing the Dirac delta distribution.

Example. We wish to compute

$$I = \int_0^\infty r e^{-r^2}\,\mathrm{d}r = \lim_{R\to\infty}\int_0^R r e^{-r^2}\,\mathrm{d}r.$$

Letting $\varphi(r) = r^2$, we have

$$I = \lim_{R\to\infty}\int_0^R \frac{1}{2}\varphi'(r)e^{-\varphi(r)}\,\mathrm{d}r = \lim_{R\to\infty}\int_{\varphi(0)}^{\varphi(R)} \frac{1}{2}e^{-x}\,\mathrm{d}x = \lim_{R\to\infty}\int_0^{R^2} \frac{1}{2}e^{-x}\,\mathrm{d}x.$$

Now one easily finds that the derivative of $-\frac{1}{2}e^{-x}$ is the expression being integrated, so that the integral may be computed using the fundamental theorem of calculus. Indeed, we get $-\frac{1}{2}e^{-R^2} - (-\frac{1}{2}e^{-0})$ and computing the limit finally yields $I = \frac{1}{2}$. Later, in Example 7.3, we shall use the above to compute

$$\int_{-\infty}^\infty \exp(-x^2)\,\mathrm{d}x = \sqrt{\pi}.$$

(This integral pops up in various statistical contexts.)

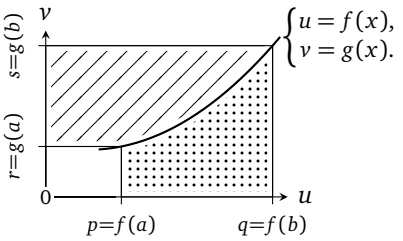
Example. We wish to compute

$$I = \int_0^\xi \tan(x)\,\mathrm{d}x = \int_0^\xi \frac{\sin(x)}{\cos(x)}\,\mathrm{d}x.$$

Observing that $\cos'(x) = -\sin(x)$, we may take $\varphi = \cos$ to get

$$I = \int_0^\xi \frac{-1}{\varphi(x)}\varphi'(x)\,\mathrm{d}x = \int_{\cos(0)}^{\cos(\xi)} \frac{-1}{t}\,\mathrm{d}t = -\log|t|\Big|_{t=1}^{\cos(\xi)}$$

Therefore, $I = -\log|\cos(\xi)| - (-\log|1|) = -\log|\cos(\xi)|$.

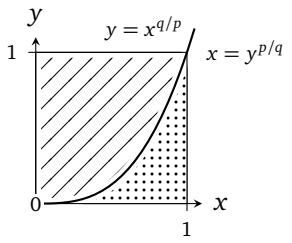


$$\text{Area}\boxed{\diagup} + \text{Area}\boxed{\vdots} = qs - pr,$$

$$\int_r^s u\,\mathrm{d}v + \int_p^q v\,\mathrm{d}u = uv\Big|_{(p,r)}^{(q,s)},$$

$$\int_a^b f(x)g'(x)\,\mathrm{d}x = f(x)g(x)\Big|_a^b - \int_a^b g(x)f'(x)\,\mathrm{d}x\,.$$

—Richard Courant



$$\int_0^1 \left(t^{p/q} + t^{q/p}\right)\mathrm{d}t = 1.$$

—Peter R. Newbury

Figure 7. Two 'proofs without words' related to integration by parts. Can you tell what is being 'proved'? (Adapted from [8].)

0.7.4. Integration by parts. Lastly, recall the formula for ***integration by parts***:

$$\int_a^b f'(x)g(x)\,\mathrm{d}x = f(x)g(x)\Big|_{x=a}^b - \int_a^b f(x)g'(x)\,\mathrm{d}x. \tag{0.8}$$

When do we hope to apply integration by parts successfully? One should remember the following mantra:

> *integration by parts is useful whenever one is to integrate a product of two functions where one of the function is easy to integrate and the other function is easy to differentiate*.

However, the reader should keep in mind that the term

$$f(x)g(x)\Big|_{x=a}^b$$

should be considered to be "easy": one *just* has to *evaluate* functions to compute it. The crucial part is the evaluation of the integral

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

Now if $f(x)g'(x)$ is somehow easier to integrate than $f'(x)g(x)$, then one is likely to win by integration by parts. We illustrate this with an example:

Example. For $k \neq 0$, we wish to compute

$$\int_0^1 x \exp(kx)\,\mathrm{d}x. \tag{0.9}$$

The idea is to apply integration by parts. There are two obvious choices for f' and g in the formula (0.8):

(1) $(f'(x), g(x)) = (x, \exp(kx))$, so that $(f(x), g'(x)) = (\frac{1}{2}x^2, k\exp(kx))$, or
(2) $(f'(x), g(x)) = (\exp(kx), x)$, so that $(f(x), g'(x)) = (\frac{1}{k}\exp(kx), 1)$.

On first inspection, both choices seem to have some merit to them. However, the second of the two turns out to be distinctly more useful. The reason is that the product $f(x)g'(x) = \frac{1}{k}\exp(kx)1$ is easy to integrate. On the other hand, for the first choice $f(x)g'(x) = \frac{1}{2}x^2 k\exp(kx)$ seems even less easy to integrate than (0.9). Hence, picking the first case, one seems to be end up in a worse position than one has started with, while picking the second case, things seem to have improved. Indeed,

$$\begin{aligned}
\int_0^1 x\exp(kx)\,\mathrm{d}x &= \frac{1}{k}\exp(kx)x\Big|_{x=0}^1 - \int_0^1 \frac{1}{k}\exp(kx)1\,\mathrm{d}x \\
&= \frac{1}{k}\exp(k1)1 - \frac{1}{k}\exp(k0)0 - \frac{1}{k^2}\exp(kx)\Big|_{x=0}^1 \\
&= \frac{1}{k}\exp(k) - \frac{1}{k^2}\exp(k1) + \frac{1}{k^2}\exp(k0) \\
&= \frac{k-1}{k^2}\exp(k) + \frac{1}{k^2}.
\end{aligned}$$

The reader should note that the above example readily generalises to integrating $P(x)\exp(kx)$ where P is an arbitrary polynomial. The underlying insight is that partial integration reduces this problem to integrating $P'(x)\frac{1}{k}\exp(kx)$ and P' is a polynomial of degree $(\deg P) - 1$. One then repeats this process, differentiating P further, until the resulting polynomial is constant. The reader should try this approach to verify that

$$\int_0^1 x^2\exp(kx)\,\mathrm{d}x = \frac{k^2-2k+2}{k^3}\exp(k) + \frac{2}{k^3}.$$

Example. Let $\xi > 1$. We wish to compute

$$I = \int_1^\xi \log(x)\,\mathrm{d}x.$$

To this end, we employ integration by parts, albeit in a slightly non-obvious fashion. Indeed, we use (0.8) with $(f'(x), g(x)) = (1, \log x)$. Thus,

$$I = x\log x\Big|_{x=1}^\xi - \int_1^\xi x\log'(x)\,\mathrm{d}x.$$

Using $\log'(x) = 1/x$, we find that

$$I = \xi \log \xi - 1 \log 1 - \int_1^\xi x \frac{1}{x}\,\mathrm{d}x = \xi \log \xi - (\xi - 1). \tag{0.10}$$

We now write $I(\xi)$ for I to draw attention to the effect of ξ. As we have seen in § 0.7.1, I, viewed as a function of ξ is differentiable and satisfies

$$I'(\xi) = \log \xi \quad \text{(the integrand evaluated at } \xi\text{)}.$$

The constant $-(-1)$ at the right hand side of (0.10) is annihilated by differentiation, so we may as well omit it in the first place when differentiating.[8] Hence, we arrive at the formula

$$\frac{\mathrm{d}}{\mathrm{d}\xi}(\xi \log \xi - \xi) = \log \xi. \tag{0.11}$$

Note that we found this by partial integration. Just for fun, let us check the above equation also simply by working out the derivative. Indeed, using linearity, the product rule and the fact that $\log' \xi = 1/\xi$, we find that

$$\frac{\mathrm{d}}{\mathrm{d}\xi}(\xi \log \xi - \xi) = 1 \log \xi + \xi \frac{1}{\xi} - 1 = 1 \log \xi + 1 - 1 = \log \xi.$$

This confirms (0.11) once more.

[8] Although we cannot omit the constant $-(-1)$ from (0.10) itself, as this would invalidate the equation.

CHAPTER 1

Complex numbers

Complex numbers are indispensable for modern mathematics. In this course we shall glimpse at some of their marvels. Whilst the present chapter is only concerned with the basics of complex numbers, we shall already see that they give rise to a unifying bridge between the exponential and the trigonometric functions. While this alone may not strike a potential reader as utterly important, we shall later encounter complex numbers in our discussion of the Laplace transform in Chapter 2, where we sketch how they help with solving certain differential equations. Even later, in Chapter 4, where practical needs originating, e.g., from solid state physics, lead us to consider periodic functions by means of Fourier analysis, complex numbers are ubiquitous. Whilst they technically *could* be avoided there (which we will not!), the initial effort of learning about them is rewarded handsomely with much nicer formulae.

> *"The shortest path between two truths in the real domain passes through the complex domain."*
>
> —Jacques Hadamard

1.1. Motivation via differential equations

We shall start our investigation with a problem that arises from classical mechanics (see Remark 1.1); such problems also arise in modelling vibrations in solids, for instance. Let a and b be two real numbers. We shall be interested in finding solutions to the differential equation

$$\ddot{x} + a\dot{x} + bx \stackrel{!}{=} 0 \tag{1.1}$$

on $\mathbb{R}$, i.e., twice-differentiable functions $x\colon \mathbb{R} \to \mathbb{R}$ such that

$$\frac{\mathrm{d}^2 x}{\mathrm{d}t^2}(t) + a\frac{\mathrm{d}x}{\mathrm{d}t}(t) + bx(t) = 0$$

holds for all $t \in \mathbb{R}$. A *trivial solution* would be the constant function $x\colon \mathbb{R} \to \mathbb{R}$, $t \mapsto 0$. To find *other* solutions, we plug in the function $x_\lambda\colon t \mapsto \exp(\lambda t)$ for x in (1.1). (The idea here is that $\exp' = \exp$ gives control over the derivative and it remains to be determined whether this control was bought at the cost of not finding any solutions.) Then the left hand of (1.1) at the point $t \in \mathbb{R}$ turns into

$$\lambda^2 \exp(\lambda t) + a\lambda \exp(\lambda t) + b\exp(\lambda t) = \underbrace{\exp(\lambda t)}_{\neq 0}\underbrace{(\lambda^2 + a\lambda + b)}_{=0\,?}.$$

This solves the differential equation (1.1) if and only if

$$\lambda^2 + a\lambda + b = 0. \tag{1.2}$$

We shall now consider two special cases of this problem.

- For instance, for $(a, b) = (0, -1)$, we obtain $\lambda \in \{1, -1\}$ (solutions of $\lambda^2 - 1 = 0$) and this yields the two solutions
$$x_1 \colon t \mapsto \exp(t) \quad \text{and} \quad x_{-1} \colon t \mapsto \exp(-t).$$
In fact, any solution of (1.1) with $(a, b) = (0, -1)$ turns out to be of the form $\mu x_1 + \tilde{\mu} x_{-1}$ for suitable scalars μ and $\tilde{\mu}$.
- On the other hand, for $(a, b) = (0, 1)$, the above strategy *seemingly* fails: there are no real numbers λ satisfying the equation
$$\lambda^2 + 1 = 0. \tag{1.3}$$
However, if there were a number, which we shall denote by "i", such that $\mathrm{i}^2 + 1 = 0$, then also $(-\mathrm{i})^2 + 1 = 0$. We expect to get solutions
$$x_{\mathrm{i}} \colon t \mapsto \exp(\mathrm{i}t) \quad \text{and} \quad x_{-\mathrm{i}} \colon t \mapsto \exp(-\mathrm{i}t).$$
It turns out that any solution of (1.1) with $(a, b) = (0, 1)$ is of the form $\mu x_{\mathrm{i}} + \tilde{\mu} x_{-\mathrm{i}}$ with suitable scalars μ and $\tilde{\mu}$. For instance, we obtain real-valued solutions
$$\cos = \tfrac{1}{2} x_{\mathrm{i}} + \tfrac{1}{2} x_{-\mathrm{i}} \quad \text{and} \quad \sin = \tfrac{1}{2\mathrm{i}} x_{\mathrm{i}} - \tfrac{1}{2\mathrm{i}} x_{-\mathrm{i}}.$$

Remark 1.1 (Physical interpretation)**.** The differential equation $\ddot{x} + x = 0$ whose solutions we have determined just now appears, for example, in mechanics when modelling a mass on a spring, ignoring friction. Indeed, suppose that the mass is constrained to move only in one dimension and let $x(t)$ denote its displacement from its equilibrium position at time t. Then, combining Newton's second law (force is mass times acceleration, the latter being $\ddot{x}$), and Hooke's law (an extended or compressed spring exerts a force proportional to the displacement from its equilibrium position in the opposite direction), one derives that

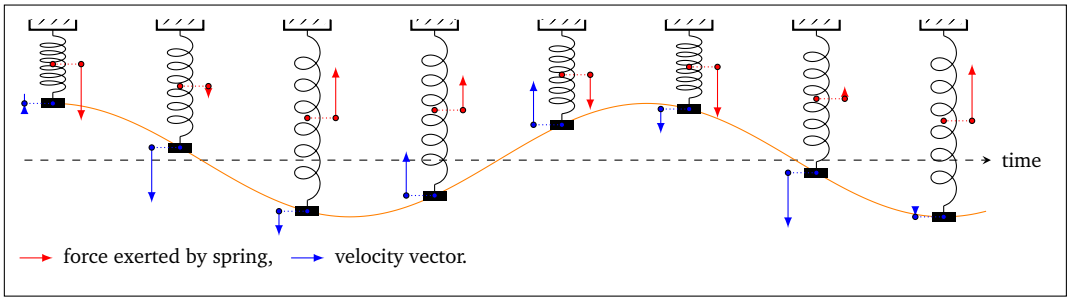
$$(\text{mass}) \times \ddot{x} = -(\text{stiffness of the spring}) \times x.$$

Abstracting away the mass and the stiffness of the spring, we get the equation $\ddot{x} = -x$, which can also be re-written as $\ddot{x} + x = 0$. It should also be noted that taking friction into account, which is proportional to the velocity of the mass, would yield an additional term involving $\dot{x}$.

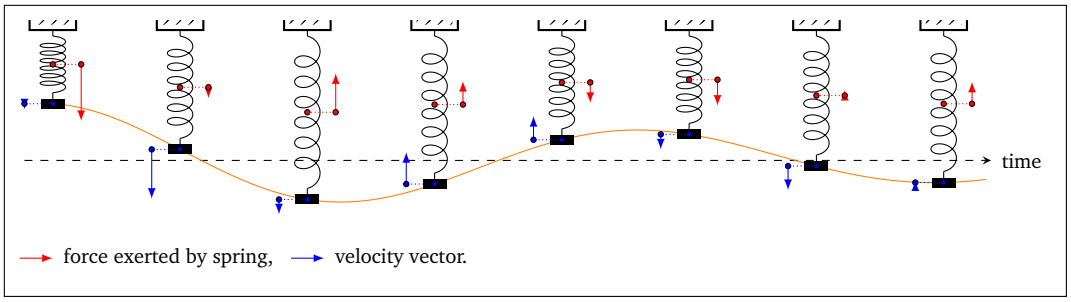
Remark (Variant of the method)**.** Suppose that $\lambda^2 + a\lambda + b = (\lambda - \lambda_0)^2$, i.e., the polynomial has a double root. Then one obtains two solutions of (1.1)

$$x_{\lambda_0} \colon t \mapsto \exp(\lambda_0 t) \quad \text{and} \quad t \mapsto t x_{\lambda_0}(t).$$

A natural generalisation of this allows for the solution of arbitrary linear ordinary differential equations with constant coefficients.



(a) Without damping.



(b) With damping.

Figure 8. Idealised movement of a mass on a spring.

Remark (Existence and uniqueness theory)**.** The claims we have made above about all solutions of the appearing differential equations being of a certain form are not obvious in a narrow sense; they are a consequence of the celebrated Picard–Lindelöf theorem which can be used to show that initial value problems (see Chapter 2) for systems of ordinary linear differential equations always admit unique solutions (provided some minor technical assumption on the "coefficients" of the system hold). We shall have no reason to expound this theory here.

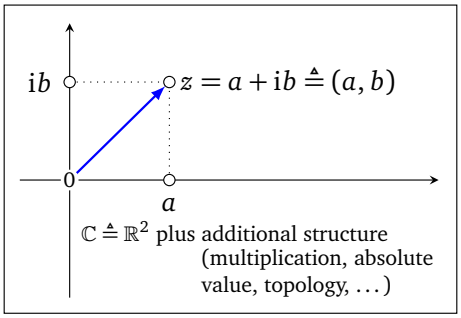
1.2. Definition and properties of complex numbers

We now introduce a set of numbers—the set of ***complex numbers***—which extends the real number system and has the pleasant feature that all non-zero polynomials (in one variable) always factor into products linear factors (see Theorem 1.2 below); in particular, the equation (1.2) is guaranteed to have solutions.

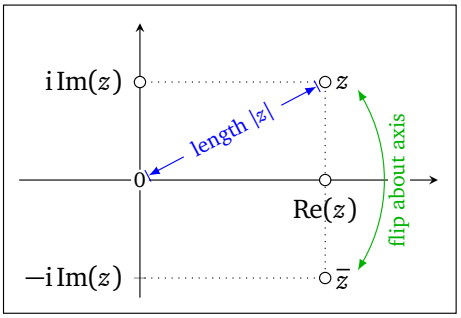
We put

$$\mathbb{C} := \{\, a + \mathrm{i}b : a, b \in \mathbb{R} \,\},$$

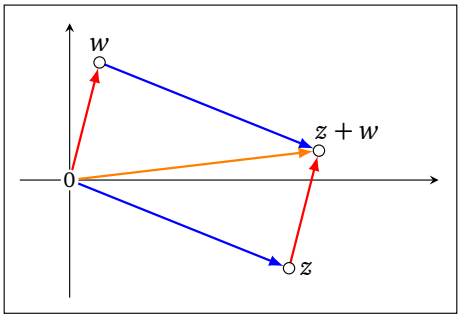
where i is a "formal" variable. The "correct" way of making sense of the expressions $a+\mathrm{i}b$ shall not concern us; here we need only know that two complex numbers $a+\mathrm{i}b$ and $c+\mathrm{i}b$ are equal if and only if $a = c$ and $b = d$. If we write $z = a+\mathrm{i}b$, then we usually understand *implicitly* that a and b are supposed to be *real* numbers. We write $\mathrm{Re}(z) := a$, the ***real part*** of z, and $\mathrm{Im}(z) := b$, the ***imaginary part*** of z.



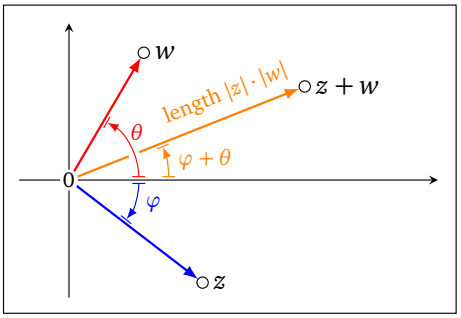
(a) Complex numbers as points or vectors in $\mathbb{R}^2$.



(b) $\mathrm{Re}(z)$, $\mathrm{Im}(z)$, $|z|$ and $\overline{z}$.



(c) Addition of complex numbers.



(d) Multiplication of complex numbers.

Figure 9. Basic notions for dealing with complex numbers.

We define ***addition and multiplication*** of two complex numbers via the formulas

$$(a+\mathrm{i}b)+(c+\mathrm{i}d) := (a+c)+\mathrm{i}(b+d), \tag{1.4}$$

$$(a+\mathrm{i}b)\cdot(c+\mathrm{i}d) := (ac-bd)+\mathrm{i}(ad+bc). \tag{1.5}$$

Addition and multiplication of complex numbers obey the usual rules one is used to from the reals. Namely, for any three complex numbers z, w, ζ we have

- $(z+w)+\zeta = z+(w+\zeta)$, $\quad (z\cdot w)\cdot\zeta = z\cdot(w\cdot\zeta)$, $\quad$ (associative laws)
- $z+w = w+z$, $\quad z\cdot w = w\cdot z$, $\quad$ (commutative laws)
- $(z+w)\cdot\zeta = (z\cdot\zeta)+(w\cdot\zeta)$. $\quad$ (distributive law)

Remark (How to remember complex multiplication)**.** The definition (1.4) of the addition of complex numbers is quite straight-forward: just add the real and imaginary parts separately. On the other hand, the definition (1.5) of the multiplication of complex numbers seems quite awkward. Luckily one basically never has to remember it. One must just remember $\mathrm{i}^2 = -1$ and multiply like one would anyway; indeed,

$$\begin{aligned}(a+\mathrm{i}b)\cdot(c+\mathrm{i}d) &= a(c+\mathrm{i}d)+\mathrm{i}b(c+\mathrm{i}d) = ac+\mathrm{i}ad+\mathrm{i}bc+\mathrm{i}^2bd\\ &= ac+\mathrm{i}ad+\mathrm{i}bc-bd = (ac-bd)+\mathrm{i}(ad+bc).\end{aligned}$$

We define the ***absolute value*** $|z|$ of a complex number $z = a + \mathrm{i}b$ to be

$$|z| := \sqrt{a^2 + b^2}.$$

(Geometrically: length of the vector $(a, b) \in \mathbb{R}^2$.) Imagining the complex numbers as points in a Cartesian coordinate system ($a + \mathrm{i}b \in \mathbb{C} \leftrightsquigarrow (a, b) \in \mathbb{R}^2$), we see that any non-zero $z \in \mathbb{C}$ may be written in ***polar form***

$$z = r(\cos\varphi + \mathrm{i}\sin\varphi)$$

with length $r = |z|$ and angle φ, the so-called ***argument*** of z. (For $z = 0$ we may write $z = 0(\cos\varphi + \mathrm{i}\sin\varphi)$ with *any* φ.) By 2π-periodicity of the cosine and sine functions, the argument φ of z is only defined up to adding integer multiples of 2π. Given $z = a + \mathrm{i}b$, its ***complex conjugate*** $\overline{z}$ is defined to be

$$\overline{(a + \mathrm{i}b)} := a - \mathrm{i}b.$$

We have

$$\overline{(z + w)} = \overline{z} + \overline{w} \quad \text{and} \quad \overline{(z \cdot w)} = \overline{z} \cdot \overline{w}.$$

Here are some special complex numbers: $0 = 0 + \mathrm{i}0 \in \mathbb{C}$ and $1 = 1 + \mathrm{i}0 \in \mathbb{C}$. We have

$$0 \cdot z = z \cdot 0 = 0 \quad \text{and} \quad 1 \cdot z = z \cdot 1 = z \quad \text{for every } z \in \mathbb{C}.$$

Moreover, any real number $a \in \mathbb{R}$ is also a complex number: $a = a + \mathrm{i}0 \in \mathbb{C}$. The number $\mathrm{i} = 0 + \mathrm{i}1$ satisfies $\mathrm{i}^2 = -1$, whereas there is no real number a such that $a^2 = -1$. Thus, $\mathbb{R} \subsetneq \mathbb{C}$. The multiplicative inverse of $a + \mathrm{i}b \neq 0$ is given by

$$(a + \mathrm{i}b)^{-1} = \frac{1}{a + \mathrm{i}b} = \frac{1}{a + \mathrm{i}b}\frac{a - \mathrm{i}b}{a - \mathrm{i}b} = \frac{a - \mathrm{i}b}{a^2 + b^2} = \frac{a}{a^2 + b^2} + \mathrm{i}\frac{-b}{a^2 + b^2}.$$

We record some useful identities:

$$\begin{aligned} z\overline{z} &= |z|^2, \quad z + \overline{z} = 2\,\mathrm{Re}(z), \qquad z - \overline{z} = 2\mathrm{i}\,\mathrm{Im}(z), \\ &(r_1[\cos\varphi_1 + \mathrm{i}\sin\varphi_1]) \cdot (r_2[\cos\varphi_2 + \mathrm{i}\sin\varphi_2]) \\ &\qquad = r_1 r_2[\cos(\varphi_1 + \varphi_2) + \mathrm{i}\sin(\varphi_1 + \varphi_2)], \\ &\frac{r_1(\cos\varphi_1 + \mathrm{i}\sin\varphi_1)}{r_2(\cos\varphi_2 + \mathrm{i}\sin\varphi_2)} = \frac{r_1}{r_2}[\cos(\varphi_1 - \varphi_2) + \mathrm{i}\sin(\varphi_1 - \varphi_2)]. \end{aligned} \tag{1.6}$$

Example.

- $(3 + 4\mathrm{i}) + (2 + \mathrm{i}) = 5 + 5\mathrm{i}$,
- $(3 + 4\mathrm{i})(2 + \mathrm{i}) = 2 + 11\mathrm{i}$,
- $\frac{1}{\mathrm{i}} = -\mathrm{i}$,
- $\overline{(1 + 2\mathrm{i})} = 1 - 2\mathrm{i}$,
- $|1 + 2\mathrm{i}| = \sqrt{1^2 + 2^2} = \sqrt{5}$,
- $\frac{1}{1+2\mathrm{i}} = \frac{1}{1+2\mathrm{i}}\frac{1-2\mathrm{i}}{1-2\mathrm{i}} = \frac{1-2\mathrm{i}}{5} = \frac{1}{5} + \mathrm{i}\frac{-2}{5}$,
- $1 + 2\mathrm{i} = \sqrt{5}(\cos(\arctan 2) + \mathrm{i}\sin(\arctan 2))$.

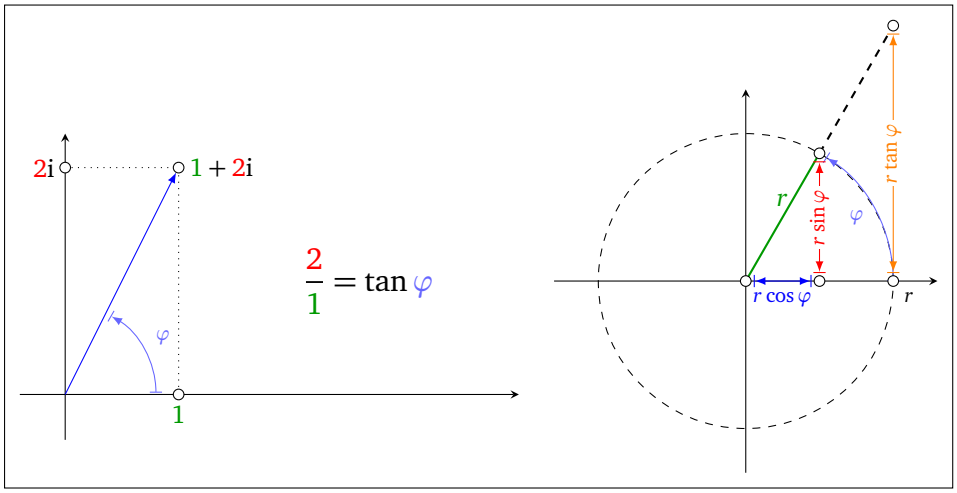


Figure 10. Illustration for finding the polar form of $1+2\mathrm{i}$.

1.3. Fundamental theorem of algebra

It turns out that our desire to find solutions to the polynomial equation (1.3) which has lead us to consider complex numbers in the first place is rewarded with a much farther-reaching result:

Theorem 1.2 (Fundamental theorem of algebra)**.** *Let n be a positive integer. Then every polynomial $f = a_n X^n + \ldots + a_1 X + a_0$ with complex coefficients $a_n, \ldots, a_1, a_0 \in \mathbb{C}$ factors as a product of linear factors:*

$$f = a_n(X - z_1)\cdots(X - z_n).$$

Moreover this factorisation is unique up to rearrangement of the factors and the equation $f(z) \stackrel{!}{=} 0$ has precisely[1] n complex solutions, namely $z_1, \ldots, z_n$.

Example. Using (1.6) we see that the numbers

$$\cos\left(\frac{2\pi k}{n}\right) + \mathrm{i}\sin\left(\frac{2\pi k}{n}\right) \quad \text{with} \quad k = 1, 2, \ldots, n$$

are solutions to the equation $X^n - 1 \stackrel{!}{=} 0$. In view of Theorem 1.2, these are all the solutions.

Example. We want to find the solutions of $1X^2 - 2X + 5 \stackrel{!}{=} 0$. The well-known formula for solving quadratic equations gives

$$\frac{-(-2) \pm \sqrt{(-2)^2 - 4 \cdot 1 \cdot 5}}{2 \cdot 1} = 1 \pm \sqrt{-4}.$$

[1] Here numbers ought to be counted as many times as they appear in the list $z_1, \ldots, z_n$ even if there are repetitions—$X^2 - 2X + 1 = (X-1)^2$.

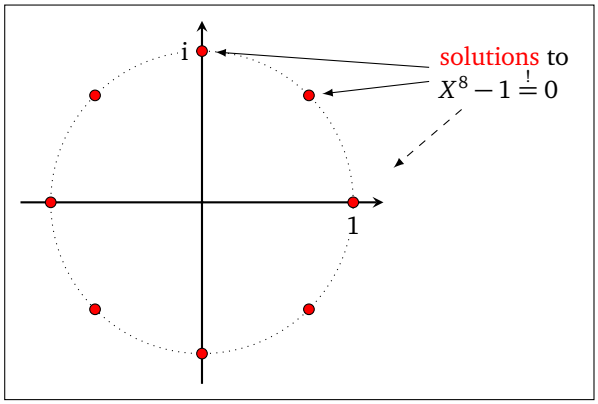


Figure 11. Illustration of the solutions to the equation $X^8 - 1 \stackrel{!}{=} 0$.

In the real numbers we would run into a problem here, because -4 has no real square root (there is no real number a such that $a^2 = -4$, because squares of real numbers are never negative). However, in the complex numbers we find that $-4 = (2\mathrm{i})^2 = (-2\mathrm{i})^2$ so that $1 \pm 2\mathrm{i}$ are the solutions we are looking for. Indeed,

$$X^2 - 2X + 5 = (X - [1 + 2\mathrm{i}])(X - [1 - 2\mathrm{i}]).$$

Example. We want to find the solutions of $X^3 \stackrel{!}{=} 1 + 2\mathrm{i}$. For this we use the polar form of

$$1 + 2\mathrm{i} = \sqrt{5}\,(\cos(\arctan 2) + \mathrm{i}\sin(\arctan 2)),$$

because for a complex number $z = r(\cos\varphi + \mathrm{i}\sin\varphi)$, the powers $z^n = r^n(\cos(n\varphi) + \mathrm{i}\sin(n\varphi))$ are easy to compute. Indeed, this yields the three solutions

$$\sqrt[3]{\sqrt{5}}\left(\cos(\tfrac{k}{3}\arctan 2) + \mathrm{i}\sin(\tfrac{k}{3}\arctan 2)\right)$$

for $k = 1, 2, 3$. (Different integers k also yield solutions, but they repeat because of 2π-periodicity of cosine and sine.)

Remark. The above examples are somewhat misleading. In general, "solving" a polynomial equation $a_n X^n + \ldots + a_1 X + a_0 \stackrel{!}{=} 0$ as explicitly as in the above is not possible. In practice, the guarantee of mere existence of solutions as furnished by Theorem 1.2 is much more important than actually having an "expression" for the solutions. Either way, there are numerical methods (which we do not discuss) for finding good approximations to such solutions.

1.4. Complex differentiation

(This section is essentially identical to § 0.6.)

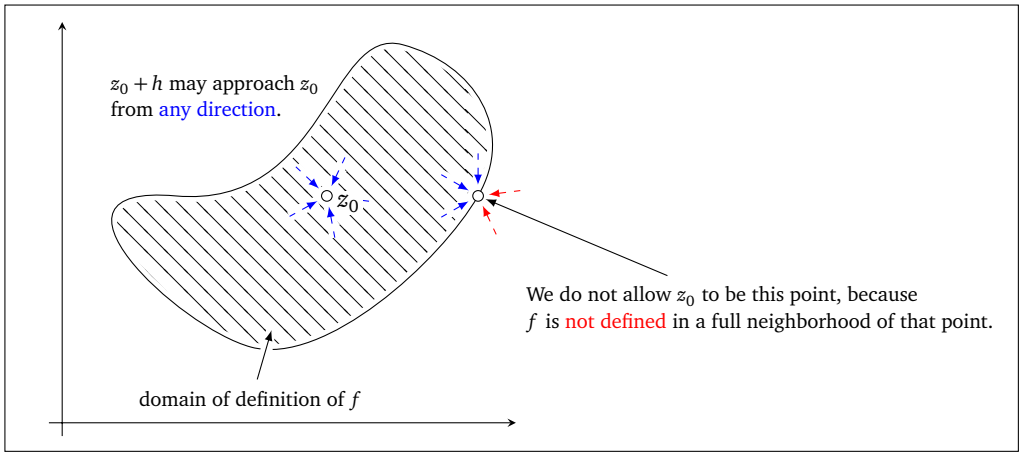


Figure 12. Illustration of the setup for the definition of complex differentiability.

Given a complex-valued function f defined in a neighbourhood[2] of a point $z_0 \in \mathbb{C}$ one defines its derivative just as one would for real-valued functions:

$$f'(z_0) := \frac{\mathrm{d}f}{\mathrm{d}z}(z_0) := \lim_{\substack{h\to 0\\ h\neq 0}} \frac{f(z_0+h)-f(z_0)}{h} \quad \text{(if the limit exists!)}.$$

Here h is taken to be a complex number. Not every function $f\colon \mathbb{C}\to\mathbb{C}$ need be differentiable:

Example. The complex conjugation $\overline{\cdot}\colon \mathbb{C}\to\mathbb{C}$, $a+\mathrm{i}b \mapsto \overline{(a+\mathrm{i}b)} = a-\mathrm{i}b$, is nowhere complex-differentiable. Indeed, for any $z_0 \in \mathbb{C}$ we have

$$\lim_{\substack{h\to 0\\ h\in\mathbb{R}\setminus\{0\}}} \frac{\overline{(z_0+h)}-\overline{z_0}}{h} = \lim_{\substack{h\to 0\\ h\in\mathbb{R}\setminus\{0\}}} 1 = 1 \neq -1 = \lim_{\substack{h\to 0\\ h\in\mathrm{i}\mathbb{R}\setminus\{0\}}} (-1) = \lim_{\substack{h\to 0\\ h\in\mathrm{i}\mathbb{R}\setminus\{0\}}} \frac{\overline{(z_0+h)}-\overline{z_0}}{h},$$

whence the limit

$$\lim_{\substack{h\to 0\\ h\neq 0}} \frac{\overline{(z_0+h)}-\overline{z_0}}{h} \quad \text{does not exist.}$$

However, complex differentiation works essentially as one is used to from real analysis. E.g., one has the rules

- $\dfrac{\mathrm{d}}{\mathrm{d}z}(\lambda f(z)+\mu g(z)) = \lambda f'(z)+\mu g'(z)$, (linearity)
- $\dfrac{\mathrm{d}}{\mathrm{d}z}(f(z)g(z)) = f'(z)g(z)+f(z)g'(z)$, (***product rule***)

[2]This means that there ought to be some disk $\{z\in\mathbb{C} : |z-z_0|<\epsilon\}$ with $\epsilon>0$ that is completely contained in the domain of definition of f; this is some sort of non-degeneracy condition which—roughly speaking—boils down to saying that f be defined on "enough" points close to z_0 so that useful assertions can reasonably be made.

- $\dfrac{\mathrm{d}}{\mathrm{d}z}\dfrac{f(z)}{g(z)} = \dfrac{f'(z)g(z)-f(z)g'(z)}{g(z)^2}$, (***quotient rule***)
- $\dfrac{\mathrm{d}}{\mathrm{d}z}f(g(z)) = f'(g(z))g'(z)$. (***chain rule***)

(The above rules are subject to "the obvious" assumptions, i.e., complex differentiability of f at the point $g(z)$ and that of g at z for the chain rule, f and g being differentiable at z and $g(z) \neq 0$ for the quotient rule etc.) Polynomials are differentiated as one expects as well:

$$\frac{\mathrm{d}}{\mathrm{d}z}z^n = nz^{n-1}. \tag{1.7}$$

(This formula also holds for every real number $n \neq 0$.)

Example. For any non-zero complex number z we have

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}z}\frac{z^2+z-3}{z^2} &= \frac{(z^2+z-3)'z^2-(z^2+z-3)(z^2)'}{(z^2)^2}\\ &= \frac{(2z+1)z^2-(z^2+z-3)(2z)}{z^4}\\ &= \frac{2z^3+z^2-(2z^3+2z^2-6z)}{z^4} = \frac{6-z}{z^3}.\end{aligned}$$

Alternatively, one can use linearity first, getting

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}z}\frac{z^2+z-3}{z^2} &= \frac{\mathrm{d}}{\mathrm{d}z}\frac{z^2}{z^2}+\frac{\mathrm{d}}{\mathrm{d}z}\frac{z}{z^2}-3\frac{\mathrm{d}}{\mathrm{d}z}\frac{1}{z^2}\\ &= \frac{\mathrm{d}}{\mathrm{d}z}1+\frac{\mathrm{d}}{\mathrm{d}z}z^{-1}-3\frac{\mathrm{d}}{\mathrm{d}z}z^{-2}\\ &= (-1)z^{-2}-3(-2)z^{-3} = \frac{6-z}{z^3}.\end{aligned}$$

Another important rule for differentiation is the rule for differentiation of inverse functions. If $f\colon U \to V$ is a bijective (1:1) function between two sets $U, V \subseteq \mathbb{C}$ (or $U, V \subseteq \mathbb{R}$ in the real case), then there exists a (unique) function $f^{-1} : V \to U$ such that $f^{-1} \circ f = \mathrm{id}_U$ and $f \circ f^{-1} = \mathrm{id}_V$. Moreover, if any point of U admits a neighbourhood contained in U and f is differentiable, then so is f^{-1}. Furthermore, given the derivative of f, the derivative of f^{-1} is also easy to find: we have

$$(f^{-1})'(y) = \frac{1}{f'(f^{-1}(y))},$$

if $f'(f^{-1}(y))$ is non-zero. (Provided that one already believes in the differentiability of f^{-1}, one can arrive at the above formula by applying the chain rule to both sides of $f^{-1} \circ f = \mathrm{id}_U$ and rearranging the resulting expression.) We give three examples of this differentiation rule.

Example (Derivative of the square root)**.** The function $f\colon \mathbb{R}_{>0} \to \mathbb{R}_{>0}$, $x \mapsto x^2$, is bijective. Its inverse function is given by $f^{-1}\colon \mathbb{R}_{>0} \to \mathbb{R}_{>0}$, $y \mapsto \sqrt{y}$. Using $\sqrt{y} = y^{1/2}$ and the differentiation rule (1.7), we have

$$\frac{\mathrm{d}}{\mathrm{d}y}\sqrt{y} = \frac{1}{2}y^{1/2-1} = \frac{1}{2\sqrt{y}}.$$

On the other hand, we can also use the aforementioned rule for finding the derivative of f^{-1}. Indeed,

$$\frac{\mathrm{d}}{\mathrm{d}y}\sqrt{y} = \frac{1}{f'(f^{-1}(y))} = \frac{1}{2f^{-1}(y)} = \frac{1}{2\sqrt{y}}.$$

The previous example is meant primarily for illustration's sake, since (as we have pointed out above) the desired result also follows from (1.7). The next two examples may seem more interesting.

Example (Derivative of the natural logarithm)**.** We wish to find the derivative of the natural logarithm $\log\colon \mathbb{R}_{>0} \to \mathbb{R}$. From school one should already know that $\log'(y) = 1/y$. To derive this, we observe that the natural logarithm is obtained as inverse function to the exponential function. Hence,

$$\log' y = \frac{1}{\exp'(\log y)} = \frac{1}{\exp(\log y)} = \frac{1}{y}.$$

Example (Derivative of the arcus tangent)**.** We wish to find the derivative of

$$\arctan\colon \mathbb{R} \to (-\pi/2, \pi/2),$$

the inverse function of $\tan : (-\pi/2, \pi/2) \to \mathbb{R}$. In preparation for this, we first find the derivative of the tangent function. We have

$$\tan'(x) = \frac{\mathrm{d}}{\mathrm{d}x}\frac{\sin x}{\cos x} = \frac{(\sin' x)(\cos x) - (\sin x)(\cos' x)}{(\cos x)^2} = \frac{(\cos x)^2 + (\sin x)^2}{(\cos x)^2}.$$

Using $(\cos x)^2 + (\sin x)^2 = 1$, the above expression could now be simplified to

$$\tan'(x) = 1/(\cos x)^2.$$

However, this is less useful for what we have in mind. Instead, we write

$$\tan'(x) = \frac{(\cos x)^2}{(\cos x)^2} + \frac{(\sin x)^2}{(\cos x)^2} = 1 + (\tan x)^2.$$

Hence,

$$\arctan'(y) = \frac{1}{\tan'(\arctan y)} = \frac{1}{1 + (\tan(\arctan y))^2} = \frac{1}{1+y^2}.$$

The above example is interesting, because arctan is arguably rather more complicated than its derivative, which happens to be a rational function (a quotient of polynomials). This shows two things: first, the derivative of a complicated function can be quite easy. Second, and more crucially, integrals of easy looking functions

may turn out to be quite more delicate. This hints that computing integrals often requires some sort of ingenuity in comparison to differentiation.

1.5. The exponential function

One defines the exponential function exp: $\mathbb{C} \to \mathbb{C}$ in the standard way by means of its power series expansion, namely

$$\exp(z) := \sum_{n=0}^{\infty} \frac{1}{n!} z^n := \lim_{N\to\infty} \sum_{n=0}^{N} \frac{1}{n!} z^n, \tag{1.8}$$

where $n! = 1 \cdot 2 \cdot \ldots \cdot n$. ($0! = 1$, $1! = 1$, $2! = 2$, $3! = 6$, $4! = 24$, ...)

One can show that the above limit converges for every z. Differentiating termwise (which would require justification given the limit over N, but we shall not do this) we get

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}z} \exp(z) = \frac{\mathrm{d}}{\mathrm{d}z} \sum_{n=0}^{\infty} \frac{1}{n!} z^n &\overset{\text{(magic)}}{=} \sum_{n=0}^{\infty} \frac{\mathrm{d}}{\mathrm{d}z} \frac{1}{n!} z^n = \sum_{n=1}^{\infty} \frac{n}{n!} z^{n-1} \\ &= \sum_{n=1}^{\infty} \frac{1}{(n-1)!} z^{n-1} = \sum_{m=1}^{\infty} \frac{1}{m!} z^m = \exp(z). \end{aligned}$$

Therefore, $\exp' = \exp$. Moreover, $\exp(0) = 1$, $\exp(1) =: e = 2.718\,281\,828\ldots$ (***Euler's constant***). We sometimes write e^z for $\exp(z)$.

Theorem 1.3. *The exponential function obeys the following rules for all $z, w \in \mathbb{C}$:*

(1) $\exp(z+w) = \exp(z)\exp(w)$, *(**functional equation** of* exp*)*
(2) $\exp(-z) = \exp(z)^{-1}$,
(3) $\exp(\mathrm{i}z) = \cos z + \mathrm{i}\sin z$, *(**Euler's formula**)*
(4) $\cos z = \frac{1}{2}(\exp(\mathrm{i}z) + \exp(-\mathrm{i}z))$, $\sin z = \frac{1}{2\mathrm{i}}(\exp(\mathrm{i}z) - \exp(-\mathrm{i}z))$,
(5) $\exp(z + 2\pi\mathrm{i}) = \exp(z)$ *and 2π cannot be replaced by a smaller positive number here without invalidating the previous equation,* *(2πi-periodicity)*
(6) $|\exp(\mathrm{i}t)| = 1$ *for every real number t.*

For real numbers $x > 0$, the logarithm $\log x$ of x is defined to be the (unique!) real number ℓ such that $\exp(\ell) = x$. With complex numbers, the problem of taking logarithms is more complicated, because the 2πi-periodicity of the exponential function kills uniqueness of the number ℓ; one *has* to make a choice if one wants to speak of $\log z$ for some complex $z \neq 0$. The subtleties of this are covered in classes on complex analysis.

Proof of Theorem 1.3 (sketch). We give some *hints* as to how to derive the above properties of the exponential function. For the functional equation one uses the

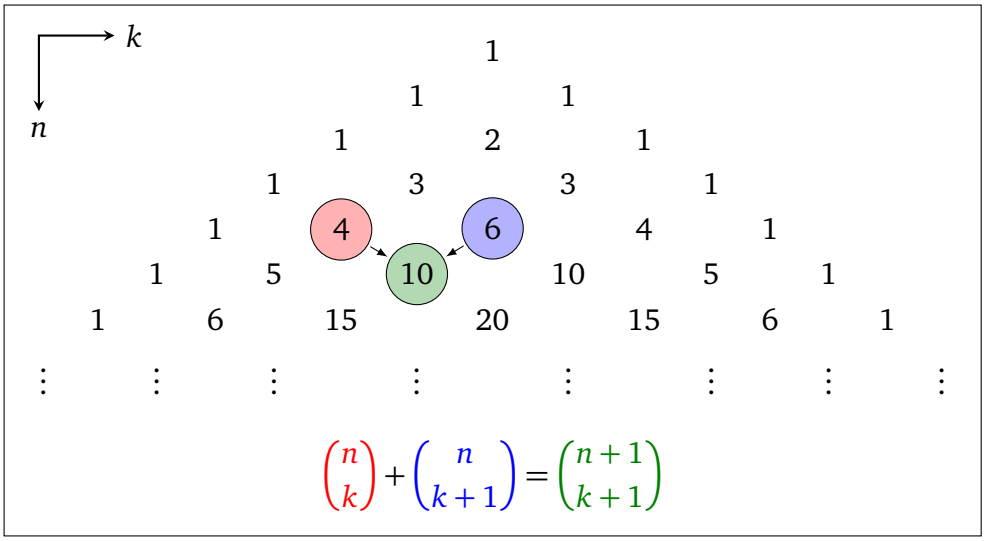


Figure 13. Pascal's triangular arrangement of the binomial coefficients which appear, for instance, in the binomial theorem which is used in our justification of the functional equation of the exponential function. The rows are $\binom{0}{0}$; $\binom{1}{0}, \binom{1}{1}$; $\binom{2}{0}, \binom{2}{1}, \binom{2}{2}$; $\binom{3}{0}, \binom{3}{1}, \binom{3}{2}, \binom{3}{3}$; ...

binomial theorem, i.e.,

$$\begin{aligned}
(z+w)^1 &= z+w,\\
(z+w)^2 &= z^2+2zw+w^2,\\
(z+w)^3 &= z^3+3z^2w+3zw^2+w^3,\\
&\vdots\\
(z+w)^n &= \binom{n}{n}z^n + \binom{n}{n-1}z^{n-1}w + \ldots + \binom{n}{1}zw^{n-1} + \binom{n}{0}w^n.
\end{aligned}$$

This shows that

$$\exp(z+w) = \sum_{n=0}^{\infty}\frac{1}{n!}(z+w)^n = \sum_{n=0}^{\infty}\frac{1}{n!}\sum_{k=0}^{n}\binom{n}{k}z^k w^{n-k}.$$

Then one expands the binomial coefficient and cancels $n!$, getting

$$\exp(z+w) = \sum_{n=0}^{\infty}\frac{1}{n!}\sum_{k=0}^{n}\frac{n!}{k!\,(n-k)!}z^k w^{n-k} = \sum_{n=0}^{\infty}\sum_{k=0}^{n}\frac{z^k}{k!}\frac{w^{n-k}}{(n-k)!}.$$

The last expression can actually be factored in the desired fashion:

$$\exp(z+w) = \sum_{n=0}^{\infty}\sum_{\substack{k,\ell=0\\ \ell+k=n}}^{n}\frac{z^k}{k!}\frac{\ell^{n-k}}{\ell!} = \Big(\sum_{k=0}^{\infty}\frac{z^k}{k!}\Big)\Big(\sum_{\ell=0}^{\infty}\frac{w^\ell}{\ell!}\Big) = \exp(z)\exp(w).$$

Now

$$1 = \sum_{n=0}^{\infty} \frac{1}{n!} 0^n = \exp(0) = \exp(z - z) = \exp(z)\exp(-z),$$

so that $\exp(z)$ cannot be zero and division by $\exp(z)$ yields

$$\exp(-z) = \exp(z)^{-1}.$$

Consequently, this implies that

$$\overline{\exp(\mathrm{i}t)} = \overline{\sum_{n=0}^{\infty} \frac{1}{n!} (\mathrm{i}t)^n} = \sum_{n=0}^{\infty} \overline{\frac{1}{n!} (\mathrm{i}t)^n} = \sum_{n=0}^{\infty} \frac{1}{n!} (-\mathrm{i}t)^n = \exp(-\mathrm{i}t) = \exp(\mathrm{i}t)^{-1}.$$

In particular,

$$|\exp(\mathrm{i}t)| = \exp(\mathrm{i}t)\overline{\exp(\mathrm{i}t)} = \exp(\mathrm{i}t)\exp(\mathrm{i}t)^{-1} = 1.$$

Given the previously stated connection to the cosine and sine function (Euler's formula), the above relation would also follow from the trigonometrical version of the Pythagorean theorem:

$$|\exp(\mathrm{i}t)| = |\cos t + \mathrm{i}\sin t| = \sqrt{(\cos t)^2 + (\sin t)^2} = 1.$$

However, Euler's formula may or may not be somewhat cumbersome to derive due to the fact that we have not bothered to fix a definition of the cosine and sine functions. The quick and dirty way would be to *define*

$$\cos z := \sum_{k=0}^{\infty} \frac{(-1)^k}{(2k)!} z^{2k} \quad \text{and} \quad \sin z := \sum_{k=0}^{\infty} \frac{(-1)^k}{(2k+1)!} z^{2k+1}. \tag{1.9}$$

Then Euler's formula can be derived quite quickly by noting that

$$(\mathrm{i}z)^n = \begin{cases} (-1)^k z^{2k} & \text{if } n = 2k \text{ for some } k \in \mathbb{N}_0, \\ \mathrm{i}(-1)^k z^{2k+1} & \text{if } n = 2k+1 \text{ for some } k \in \mathbb{N}_0 \end{cases}$$

and plugging $\mathrm{i}z$ into the definition of $\exp(\,\cdot\,)$ as a power series (1.8). On the other hand, since the cosine and sine functions are usually more familiar in a geometrical context and not as power series, this approach seems a bit weird. One could use the addition theorems for cosine and sine to show that they both solve the linear ordinary differential equation $\ddot{x} + x = 0$ and use this to derive their power series expansions (1.9). More geometrical proofs can also be cooked up, showing (say) that $\mathbb{R} \to \mathbb{R}$, $t \mapsto \exp(\mathrm{i}t)$, travels on the unit circle $\{ z \in \mathbb{C} : |z| = 1 \}$ at unit speed in counter-clockwise direction. Having verified Euler's formula for real values, this can then be extended to all complex numbers using ideas of complex analysis. We shall pursue none of these approaches and instead take Euler's formula on faith. □

Example. What is the absolute value of $\exp(2 + 5\mathrm{i})$?

$$|\exp(2 + 5\mathrm{i})| = |\exp(2)\exp(5\mathrm{i})| = |\exp(2)||\exp(5\mathrm{i})| = |\exp(2)|1 = \exp(2).$$

Remark (A warning)**.** Euler's formula may lead one to believe that

$$\cos(z) = \mathrm{Re}(\exp(\mathrm{i}z)) \quad \text{and} \quad \sin(z) = \mathrm{Im}(\exp(\mathrm{i}z)).$$

This is true for *real* z, but false in general for complex z. For instance, we have

$$\begin{aligned} 2\cos(\pi/2-\mathrm{i}) &= \exp(\mathrm{i}(\pi/2-\mathrm{i})) + \exp(-\mathrm{i}(\pi/2-\mathrm{i})) \\ &= \exp(1+\mathrm{i}\pi/2) + \exp(-1-\mathrm{i}\pi/2) \\ &= \exp(1)\exp(\mathrm{i}\pi/2) + \exp(-1)\exp(-\mathrm{i}\pi/2) \\ &= \exp(1)\mathrm{i} + \exp(-1)(-\mathrm{i}) \\ &= (\exp(1) - 1/\exp(1))\mathrm{i} \in \mathbb{C}\setminus\mathbb{R}. \end{aligned}$$

In particular, this shows that $\cos(\pi/2-\mathrm{i}) \neq \mathrm{Re}(\ldots) \in \mathbb{R}$.

CHAPTER 2

The Laplace transform

In this chapter we discuss a useful gadget for solving linear differential equations. It is especially convenient in situations where one is looking for one solution satisfying certain initial conditions rather than trying to find all solutions to the general equation. Incidentally, the latter can be handled using a different approach based on linear algebra, matrix exponentials and so-called "variation of constants" formulae. Especially in higher dimensions such procedures (although quite general) turn out to be difficult to work with unless left to a computer. For that reason, we stick to the Laplace transform technique and neglect to describe the other approach altogether. (Incidentally, this justification is only partially honest; as it turns out, the necessary step of inverting the Laplace transform can also be quite hard. However, in various sample applications it still turns out to be feasible enough.)

2.1. Motivation and definition

2.1.1. Motivation. Our goal shall be to solve "initial value problems" of the following shape

$$\text{(2.1)}\quad \begin{cases} \text{differential eq.: } a_k x^{(k)} + \ldots + a_2 \ddot{x} + a_1 \dot{x} + a_0 x \stackrel{!}{=} f \\ \qquad\qquad \text{on the interval } (0,\infty) \text{ with continuous extension to } 0, \\ \qquad\qquad \text{right-sided } k\text{—}1\text{-th derivatives existing \& satisfying the} \\ \text{initial conditions: } \begin{cases} x^{(k-1)}(0) \stackrel{!}{=} x_{k-1}, \\ \qquad\vdots \\ \dot{x}(0) \stackrel{!}{=} x_1, \\ x(0) \stackrel{!}{=} x_0. \end{cases} \end{cases}$$

Here $k \in \mathbb{N}$, $a_0, a_1, a_2, \ldots, a_k, x_0, x_1, \ldots, x_{k-1} \in \mathbb{R}$ and $f : [0,\infty) \to \mathbb{R}$ is supposed to be some arbitrary function. Such systems arise naturally, for instance, when studying mechanical systems acted on by an external driving force (modelled by f). This may be contrasted to the mass on a spring considered in § 1.1, whose behaviour after time $t = 0$ we have left undisturbed. In the examples below, we shall restrict ourselves to $k = 2$ for simplicity. We mention also, that the method to be discussed is not limited to imposing initial conditions at $t = 0$, but we do not elaborate this further.

Returning to (2.1), strictly speaking, we should make some assumptions on f at this point as to ensure solubility of the system, but we deliberately do not; the Laplace

transform, which we shall discuss shortly, can still be applied to such situations and often yields an "answer" to the above initial value problem even if no solution exists in the strict sense. As it turns out, such an answer may still satisfy a physicist. (General viewpoint: the initial value problems in question often come from modelling a physical problem. If the initial value problem is ill-posed, this is simply interpreted as saying that the model at hand does not quite reflect all aspects of the real-world situation. A physicist is well aware of this and, consequently, not bothered: in modelling, one routinely makes idealising assumptions. Therefore, a physicist may also be very content to accept 'approximate' or 'almost' solutions, as they may still bear sufficient resemblance to the solution of the real-world problem one actually intends on describing. We shall sketch such a situation in § 2.4 below.

To motivate what we are about to do, just *assume* for the moment that there is some 'magical process', turning functions $x\colon [0,\infty) \to \mathbb{R}$ into functions $\mathscr{L}\{x\}$ (defined on some mysterious subset of the complex numbers which we do not dare to specify here) that satisfies the following properties for any suitable functions $x, y\colon [0,\infty) \to \mathbb{R}$ and any $\lambda, \mu \in \mathbb{R}$:

- $\mathscr{L}\{\lambda x + \mu y\} = \lambda \mathscr{L}\{x\} + \mu \mathscr{L}\{y\}$, (linearity)
- $\mathscr{L}\{t \mapsto 1\} = (s \mapsto 1/s)$,
- $\mathscr{L}\{\dot{x}\}(s) = s\mathscr{L}\{x\}(s) - x(0)$. ('***magical differentiation property***')

Then, tacitly assuming that x is differentiable sufficiently often,

$$\begin{aligned}\mathscr{L}\{\ddot{x}\}(s) &= s\mathscr{L}\{\dot{x}\}(s) - \dot{x}(0) = s(s\mathscr{L}\{x\}(s) - x(0)) - \dot{x}(0)\\ &= s^2\mathscr{L}\{x\}(s) - sx(0) - \dot{x}(0),\end{aligned}$$

and

$$\mathscr{L}\{\dddot{x}\}(s) = \ldots = s^3\mathscr{L}\{x\}(s) - s^2x(0) - s\dot{x}(0) - \ddot{x}(0),$$

and, more generally,

$$\mathscr{L}\{x^{(k)}\}(s) = s^k\mathscr{L}\{x\}(s) - s^{k-1}x(0) - s^{k-2}\dot{x}(0) - \ldots - x^{(k-1)}(0).$$

We now apply this to the following special case of (2.1):

$$\text{(2.2)} \qquad \begin{cases} \text{differential equation: } \ddot{x} - 2\dot{x} - 3x \stackrel{!}{=} f \text{ on } \mathbb{R}_+, \\ \text{initial conditions: } \begin{cases} \dot{x}(0) \stackrel{!}{=} 1, \\ x(0) \stackrel{!}{=} 0. \end{cases} \end{cases}$$

Suppose we have some solution $x\colon [0,\infty) \to \mathbb{R}$ to the above initial value problem. We then compute

$$\begin{aligned}\mathscr{L}\{f\}(s) &= \mathscr{L}\{\ddot{x} + 2\dot{x} - 3x\}(s)\\ &= \mathscr{L}\{\ddot{x}\}(s) - 2\mathscr{L}\{\dot{x}\}(s) - 3\mathscr{L}\{x\}(s)\\ &= (s^2\mathscr{L}\{x\}(s) - sx(0) - \dot{x}(0)) - 2(s\mathscr{L}\{x\}(s) - x(0)) - 3\mathscr{L}\{x\}(s)\\ &= (s^2 - 2s - 3)\mathscr{L}\{x\}(s) + 2x(0) - sx(0) - \dot{x}(0).\end{aligned}$$

Using our initial conditions, we find that

$$\mathscr{L}\{f\}(s) = (s^2 - 2s - 3)\mathscr{L}\{x\}(s) - 1$$

and solving for $\mathscr{L}\{x\}(s)$ gives

$$\mathscr{L}\{x\}(s) = \frac{1 + \mathscr{L}\{f\}(s)}{s^2 - 2s - 3}.$$

Suppose now that $f(t) = 1$ for all t. Then $\mathscr{L}\{f\}(s) = 1/s$. We arrive at

$$\mathscr{L}\{x\}(s) = \frac{1 + 1/s}{s^2 - 2s - 3} = \frac{s+1}{s^3 - 2s^2 - 3s}.$$

Hence, we have a *very explicit* formula for $\mathscr{L}\{x\}(s)$. If we were now able to *invert* the operation of applying $\mathscr{L}$, getting

$$x = \mathscr{L}^{-1}\{\mathscr{L}\{x\}\} = \mathscr{L}^{-1}\left\{s \mapsto \frac{s+1}{s^3 - 2s^2 - 3s}\right\}, \tag{2.3}$$

then we would have solved (2.2).

Remark. One can check that $x\colon t \mapsto \frac{1}{3}e^{3t} - \frac{1}{3}$ solves (2.2). See Example 2.5 below for a solution.

2.1.2. The actual definition. Given a function $f\colon [0,\infty) \to \mathbb{R}$, we look at

$$\mathscr{L}\{f\}(s) \coloneqq \int_0^\infty f(t)e^{-st}\,\mathrm{d}t.$$

Here we tacitly assume that f is "sufficiently nice" so that integration makes sense[1] and we suppose that there is some $s_0 > 0$ such that $|f(t)| \le s_0 e^{s_0 t}$ for every $t \in [0,\infty)$. (One says that f is of ***exponential (growth) order***.) Then the above integral converges for every *complex* number s with $\mathrm{Re}(s) > s_0$. Moreover, we have

$$\lim_{t\to\infty} f(t)e^{-sT} = 0 \tag{2.4}$$

for $\mathrm{Re}(s) > s_0$. The resulting function $\mathscr{L}\{f\}\colon \ldots \to \mathbb{C}$ is called the ***Laplace transform*** of f. (Here we are deliberately vague about the domain of definition of $\mathscr{L}\{f\}$; see Remark 2.1 below.)

Example. $\mathscr{L}\{t \mapsto 1\}(s) = \displaystyle\int_0^\infty 1e^{-st}\,\mathrm{d}t = \frac{1}{-s}e^{-st}\Big|_{t=0}^{\infty} = 0 - \frac{1}{-s}e^{-s0} = \frac{1}{s}.$

Remark 2.1. Observe that the integral defining $\mathscr{L}\{t \mapsto 1\}(s)$ converges only for $s \in \mathbb{C}$ with $\mathrm{Re}(s) > 0$. However, the right hand side, $1/s$, is well-defined for *any* complex number $s \neq 0$.

[1] Actually, we shall assume this throughout, thus making some statements below not strictly correct unless this assumption is added. For seeing that our application of the Laplace transform to various differential equations is actually justified, one could use so-called Grönwall-type inequalities. We shall not venture into this territory, though.

It is quite obvious from the definition of the Laplace transform that it is linear in the following sense: for any two functions $f, g\colon [0,\infty)\to\mathbb{R}$ of exponential order and any $\lambda,\mu\in\mathbb{R}$ the function $\lambda f+\mu g$ is also of exponential order and we have

$$\mathscr{L}\{\lambda f+\mu g\}=\lambda\mathscr{L}\{f\}+\mu\mathscr{L}\{g\}.$$

(At least for all function arguments s with sufficiently large real part. Mind here again that we are scrubbing the domain of definition of the transformed functions under the rug.)

Next, we show how one may derive the "magical differentiation property" that we have encountered in § 2.1.1, namely

$$\mathscr{L}\{\dot f\}(s)=s\mathscr{L}\{f\}(s)-f(0). \tag{2.5}$$

This follows by partial integration. Indeed,

$$\mathscr{L}\{\dot f\}(s)=\int_0^\infty\Big(\frac{\mathrm{d}}{\mathrm{d}t}f(t)\Big)e^{-st}\,\mathrm{d}t=f(t)e^{-st}\Big|_{t=0}^{\infty}-\int_0^\infty f(t)\Big(\frac{\partial}{\partial t}e^{-st}\Big)\,\mathrm{d}t.$$

The evaluation of the first term on the right at ∞ (which should be understood as evaluating at t and then taking $t\to\infty$ using a limiting process) yields zero; see (2.4). Consequently, we find that

$$\mathscr{L}\{\dot f\}(s)=-f(0)e^{-s0}-\int_0^\infty f(t)\big(-se^{-st}\big)\,\mathrm{d}t=s\int_0^\infty f(t)e^{-st}\,\mathrm{d}t-f(0).$$

The last integral here is plainly $\mathscr{L}\{f\}(s)$. Hence, we have (2.5).

2.2. Computing and inverting the Laplace transform

Regarding reversal of the Laplace transform, one has the following general result:

Theorem 2.2 (Laplace inversion formula)**.** *Let $f\colon[0,\infty)\to\mathbb{R}$ be continuous, of bounded variation*[2] *and of exponential order. Suppose that $s_0>0$ is chosen sufficiently large so that $\mathscr{L}\{f\}(s)$ is defined for all s with $\operatorname{Re}(s)>s_0-1$. Then, for $t>0$,*

$$f(t)=\frac{1}{2\pi\mathrm{i}}\int_{s_0-\mathrm{i}\infty}^{s_0+\mathrm{i}\infty}\mathscr{L}\{f\}(s)e^{st}\,\mathrm{d}s:=\frac{1}{2\pi}\lim_{T\to\infty}\int_{-T}^{T}\mathscr{L}\{f\}(s_0+\mathrm{i}\tau)e^{(s_0+\mathrm{i}\tau)t}\,\mathrm{d}\tau.$$

(The formula remains valid for $t=0$ if the left hand side is replaced by $\frac{1}{2}f(t)$.)

Proof (sketch). Let $I=\int_{-T}^{T}\ldots\mathrm{d}\tau$ denote the integral on the very right hand side of the formula claimed in the theorem. Plugging in the definition of $\mathscr{L}\{f\}$, we have

$$I=\int_{-T}^{T}\int_0^\infty f(\theta)e^{-(s_0+\mathrm{i}\tau)\theta}e^{(s_0+\mathrm{i}\tau)t}\,\mathrm{d}\theta\,\mathrm{d}\tau.$$

[2]The reader may replace "bounded variation" with "continuously differentiable with sufficiently quickly decaying derivative" without loosing anything. In fact, the reader may replace all assumptions on f with "let $f\colon[0,\infty)\to\mathbb{R}$ be a *sufficiently nice* function".

Assuming that it is justified to switch the order of integration, we get further

$$I = \int_0^\infty \int_{-T}^{T} \dots \underbrace{\mathrm{d}\tau\,\mathrm{d}\theta}_{\smile} = \int_0^\infty f(\theta) e^{s_0(t-\theta)} \int_{-T}^{T} e^{\mathrm{i}\tau(t-\theta)}\,\mathrm{d}\tau\,\mathrm{d}\theta.$$

The inner-most integral is easily evaluated: for $t \neq \theta$ we have

$$\int_{-T}^{T} e^{\mathrm{i}\tau(t-\theta)}\,\mathrm{d}\tau = \frac{1}{\mathrm{i}(t-\theta)} e^{\mathrm{i}\tau(t-\theta)}\Big|_{\tau=-T}^{T} = \frac{e^{\mathrm{i}T(t-\theta)} - e^{-\mathrm{i}T(t-\theta)}}{\mathrm{i}(t-\theta)} = 2T\,\mathrm{si}(T(t-\theta)),$$

where $\mathrm{si}(x) \coloneqq \sin(x)/x$ for $x \neq 0$ and $\mathrm{si}(0) \coloneqq 1$ (see Figure 14). Hence,

$$I = 2T \int_0^\infty f(\theta) e^{s_0(t-\theta)}\,\mathrm{si}(T(t-\theta))\,\mathrm{d}\theta.$$

Now as T approaches ∞, the integrand $\mathrm{si}(T(t-\theta))$ oscillates faster and faster where $t \neq \theta$, and one may suspect that this leads to massive cancellation (even when weighted[3] by the additional factor $f(\theta)e^{s_0(t-\theta)}$) after integration. For $\theta \approx t$, however, little to no oscillation takes place. The exact evaluation is a bit tricky, but suffice it to say that using the so-called ***Dirichlet integral***

$$\lim_{Y\to\infty} \int_{-Y}^{Y} \frac{\sin x}{x}\,\mathrm{d}x = \pi$$

one can actually show that this oscillatory behaviour "catches" exactly the contribution of $f(\theta)e^{s_0(t-\theta)}$ at $\theta = t$, i.e., $f(t)$ (multiplied by[4] π). □

Remark. Tools for computing integrals such as the ones in Theorem 2.2 are available from a branch of mathematics called "complex analysis". The basic insight is that one can exchange the integral $\int_{s_0-\mathrm{i}\infty}^{s_0+\mathrm{i}\infty}$ for an integral $\int_{s_1-\mathrm{i}\infty}^{s_1+\mathrm{i}\infty}$ at the expense of a controllable correction term, see Figure 15. ("Controllable," that is, if one is familiar with complex analysis.) Observe that if $s_1 \in \mathbb{R}$ is taken to approach[5] $-\infty$, then the new integrand $e^{(s_1+\mathrm{i}\tau)t}$ approaches zero:

$$|e^{(s_1+\mathrm{i}\tau)t}| = |e^{s_1 t} e^{\mathrm{i}\tau t}| = e^{s_1 t}|e^{\mathrm{i}\tau t}| = e^{s_1 t} \xrightarrow{s_1\to-\infty} 0 \quad (\text{if } t > 0).$$

The integral in question is thus seen to equal the "controllable correction term". The magic of complex analysis allows here to evaluate these correction terms (often *without* solving difficult integrals). We cannot possibly go into the details here.

[3] This is one point where the ominous "bounded variation" assumption comes into play.

[4] If $t = 0$, then one only gets half of the Dirichlet integral, explaining the mysterious factor $\frac{1}{2}$ in the statement of the theorem for this case.

[5] This is another reason why it is often necessary to enlarge the domain of definition of the functions one gets from Laplace's transform as hinted at previously in Remark 2.1. If such an enlargement were not used, then the integral exchange just described would not be possible.

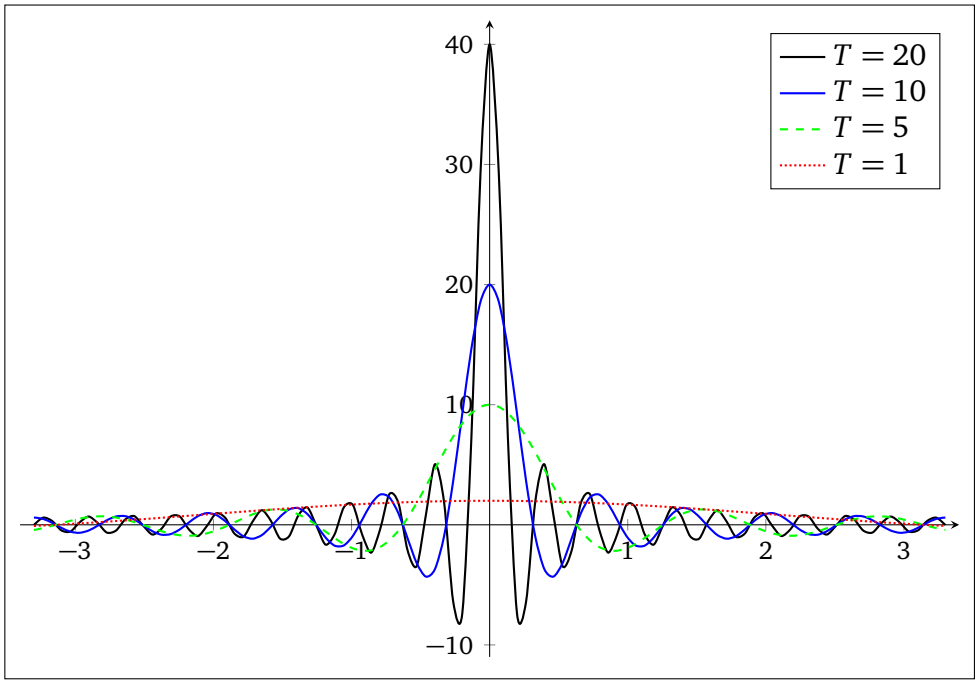


Figure 14. Plot of the function $t \mapsto 2T\,\mathrm{si}(T(t-\theta))$ appearing in the proof of Theorem 2.2 with $\theta = 0$ and various values of T.

Corollary 2.3. *Suppose that $f, g\colon [0,\infty) \to \mathbb{R}$ are continuous, of bounded variation and of exponential order. If the Laplace transforms of f and g coincide, then so do f and g.*

Proof. If $\mathscr{L}\{f\} = \mathscr{L}\{g\}$, then by linearity $\mathscr{L}\{f-g\} = (s \mapsto 0)$. Theorem 2.2 implies $f - g = (t \mapsto 0)$, so that $f = g$, as claimed. □

We have already seen in § 2.1.2 that the Laplace transform actually has the magical properties employed in § 2.1.1. Moreover, Theorem 2.2 guarantees that inverting the Laplace transform is theoretically possible; the worst-case scenario that $\mathscr{L}\{f\} = \mathscr{L}\{g\}$ for two *different* functions does *not occur* (at least not if both functions are sufficiently nice, see Corollary 2.3). We give a (non-exhaustive) list of known Laplace transforms. As we shall see, identifying such "known parts" in a given Laplace transform $\mathscr{L}\{f\}$ can sometimes be used to recover f without having to calculate the integral in Theorem 2.2.

Proposition 2.4. *Let $f, g\colon [0,\infty) \to \mathbb{R}$ be of exponential order and put $F := \mathscr{L}\{f\}$ as well as $G := \mathscr{L}\{g\}$. Let $f * g\colon [0,\infty) \to \mathbb{R}$ be the* ***convolution*** *of f and g defined by*

$$(f * g)(t) := \int_0^t f(\tau) g(t-\tau)\,\mathrm{d}\tau.$$

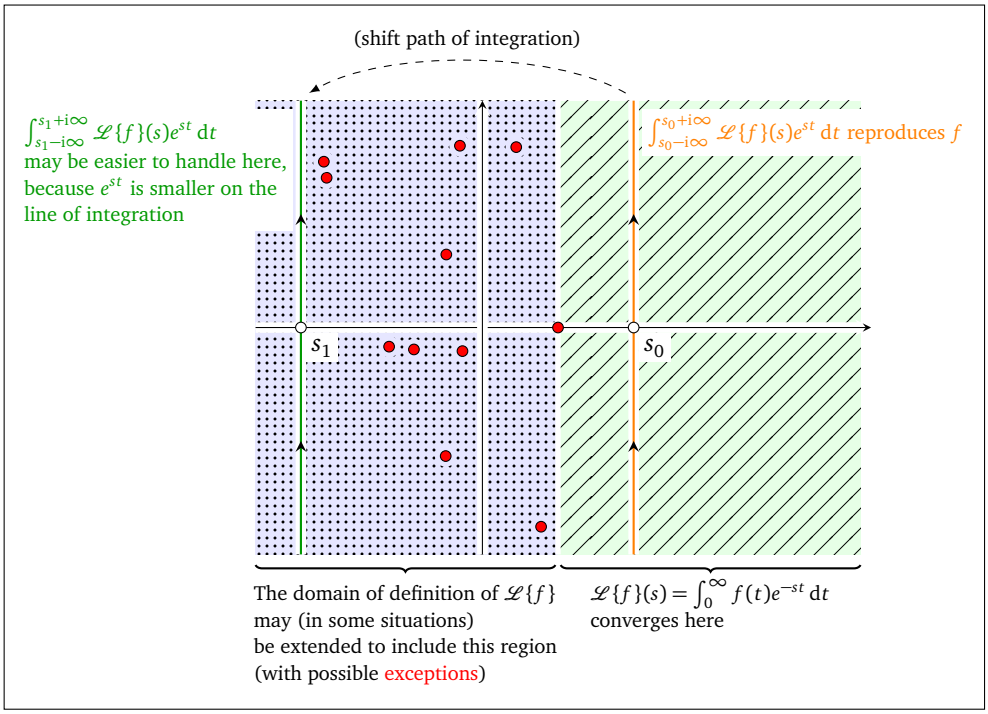


Figure 15. Illustration of the complex analysis approach to inverting the Laplace transform.

$h(t)$	a	t	t^n	e^{at}	$\cos(\omega t)$	$\sin(\omega t)$
$\mathscr{L}\{h\}(s)$	$\frac{a}{s}$	$\frac{1}{s^2}$	$\frac{n!}{s^{n+1}}$	$\frac{1}{s-a}$	$\frac{s}{s^2+\omega^2}$	$\frac{\omega}{s^2+\omega^2}$
$r(t)$	$e^{at}f(t)$	$f(ct)$	$\int_0^t f(\tau)\,d\tau$	$f * g$	$f_{\to c}$	$t^n f(t)$
$\mathscr{L}\{r\}(s)$	$F(s-a)$	$c^{-1}F(s/c)$	$\frac{F(s)}{s}$	$F(s)\cdot G(s)$	$e^{-cs}F(s)$	$(-1)^n \frac{d^n F}{ds^n}(s)$

Table 1. Various functions with their respective Laplace transform.

Moreover, let $f_{\to c}$ denote the right shift of f by c defined by

$$f_{\to c}(t) = \begin{cases} f(t-c) & \text{if } t \geq c, \\ 0 & \text{if } t < c. \end{cases}$$

Then, for $a, \omega \in \mathbb{R}$, $c > 0$, $n \in \mathbb{N}_0$, Table 1 provides a list of various functions with their respective Laplace transform.

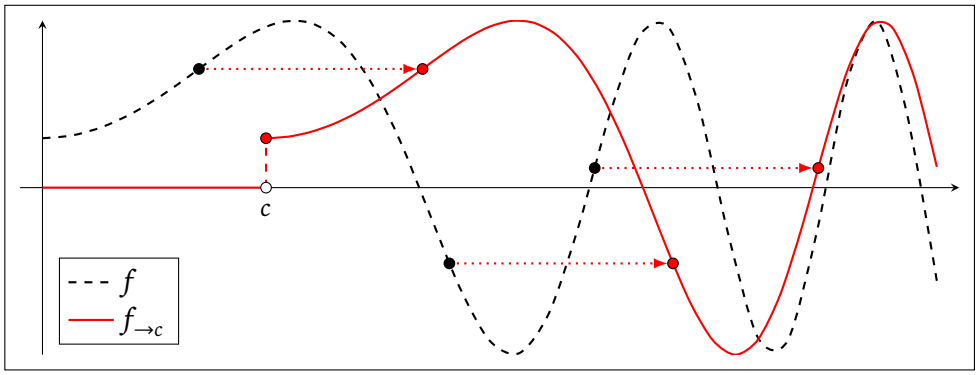


Figure 16. A function f and the right shift $f_{\to c}$ of f by some $c > 0$.

2.3. Examples

We now discuss several examples of how to apply Laplace transform techniques in practice.

2.3.1. Linear differential equations with constant coefficients.

Example 2.5. We now recall that in (2.3) we were faced with the problem of finding a function whose Laplace transform is

$$X\colon s \mapsto \frac{s+1}{s^3-2s^2-3s}.$$

We have already hinted there that $x\colon t \mapsto \frac{1}{3}e^{3t} - \frac{1}{3}$ is such a function, but we shall now use Proposition 2.4 and Corollary 2.3 to arrive at the same conclusion. Indeed, the denominator of the above fraction factors as

$$s^3 - 2s^2 - 3s = (s^2 - 2s - 3)s = (s-3)(s+1)s.$$

Hence,

$$X(s) = \frac{s+1}{s^3-2s^2-3s} = \frac{1}{(s-3)s} = \frac{1}{3(s-3)} - \frac{1}{3s}.$$

For the last step we have computed a partial fraction decomposition (see Proposition 2.8 below and the examples that follow it). Looking at Table 1 and making use of linearity of the Laplace transform, this turns out to be

$$X(s) = \tfrac{1}{3}\mathscr{L}\{t \mapsto e^{3t}\} - \mathscr{L}\{t \mapsto 1/3\} = \mathscr{L}\{t \mapsto \tfrac{1}{3}e^{3t} - \tfrac{1}{3}\}.$$

Hence (by Corollary 2.3), the sought-after solution is $x\colon t \mapsto \frac{1}{3}e^{3t} - \frac{1}{3}$, as claimed. It would be good form to actually verify by hand now that this x does indeed solve (2.2). We leave this as an exercise to the reader.

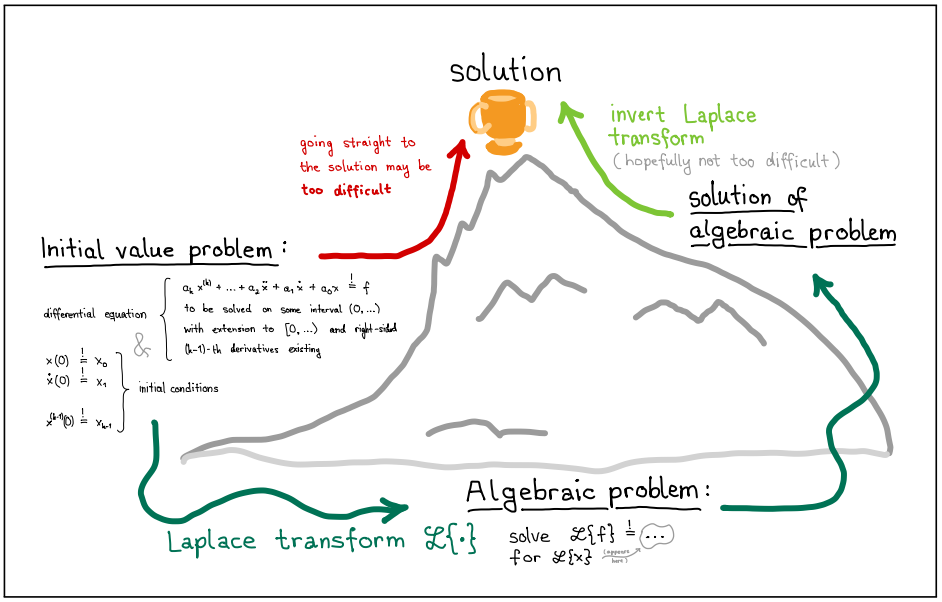


Figure 17. Illustration of the approach to solving initial value problems using the Laplace transform.

Remark (Strategy for solving initial value problems via the Laplace transform)**.** The reader should reflect on the previous example as well as 2.1.1 to realise that the Laplace transform transports the problem of solving the initial value problem (2.1) into an algebraic problem (solving the resulting equation for $\mathscr{L}\{x\}$). Assuming that this problem can be solved and the Laplace transform $\mathscr{L}\{f\}$ of the function f appearing in (2.1) can be computed, one can arrive at a formula for the solution of (2.1) by inverting the Laplace transform. In practice, this last step turns out to be the most complicated one. Here one can use Theorem 2.2 if one is sufficiently versed in complex analysis (which we are not) or hope to find (by other means, e.g., by use of tables such as Table 1 and exploiting linearity) a function whose Laplace transform has the desired shape. This function is then, *a-fortiori*, the sought-after solution of (2.1).

2.3.2. Systems of linear differential equations.

Example 2.6. Consider the following initial value problem:

$$\begin{cases} \text{differential equation:} & \begin{cases} \dot{x}(t) \overset{!}{=} x(t) + y(t), \\ \dot{y}(t) \overset{!}{=} 2x(t) - 2t \text{ for } t > 0, \end{cases} \\ \text{initial conditions:} & \begin{cases} x(0) \overset{!}{=} 1/6, \\ y(0) \overset{!}{=} 1/6. \end{cases} \end{cases}$$

Suppose that x and y are solutions to this. Then, by applying the Laplace transform, we get

$$s\mathscr{L}\{x\}(s)-\tfrac{1}{6}=s\mathscr{L}\{x\}(s)-x(0)=\mathscr{L}\{\dot{x}\}(s)=\mathscr{L}\{x\}(s)+\mathscr{L}\{y\}(s)$$

as well as

$$\begin{aligned}s\mathscr{L}\{y\}(s)-\tfrac{1}{6}&=s\mathscr{L}\{y\}(s)-y(0)=\mathscr{L}\{\dot{y}\}(s)\\&=2\mathscr{L}\{x\}(s)-\mathscr{L}\{t\mapsto t\}(s)=2\mathscr{L}\{x\}(s)-1/s^2.\end{aligned}$$

Hence

$$\begin{cases}s\mathscr{L}\{x\}(s)-\frac{1}{6}=\mathscr{L}\{x\}(s)+\mathscr{L}\{y\}(s),\\ s\mathscr{L}\{y\}(s)-\frac{1}{6}=2\mathscr{L}\{x\}(s)-1/s^2.\end{cases}$$

Solving the second equation for $\mathscr{L}\{y\}$ we find that

$$\mathscr{L}\{y\}(s)=2\mathscr{L}\{x\}(s)/s+1/(6s)-1/s^3. \tag{2.6}$$

Plugging this into the first equation gives

$$(s-1)\mathscr{L}\{x\}(s)=2\mathscr{L}\{x\}(s)/s+1/(6s)-1/s^3+1/6.$$

Rearranging yields

$$\mathscr{L}\{x\}(s)=\frac{1/(6s)-1/s^3+1/6}{s-1-2/s}=\frac{s^3+s^2-6}{6(s^2-s-2)s^2}=\frac{s^3+s^2-6}{6(s-2)(s+1)s^2}.$$

A partial fraction decomposition yields

$$\mathscr{L}\{x\}(s)=\frac{1}{12(s-2)}+\frac{1}{3(s+1)}-\frac{1}{4s}+\frac{1}{2s^2}.$$

In view of Table 1 and Corollary 2.3 we find that

$$x(t)=\frac{e^{2t}}{12}+\frac{e^{-t}}{3}-\frac{1}{4}+\frac{t}{2}.$$

Moreover, plugging the above expression for $\mathscr{L}\{x\}$ into our formula (2.6) for $\mathscr{L}\{y\}$ yields

$$\mathscr{L}\{y\}(s)=\frac{s^3+s^2-6}{3(s-2)(s+1)s^3}+\frac{1}{6s}-\frac{1}{s^3}=\frac{1}{12(s-2)}-\frac{2}{3(s+1)}+\frac{3}{4s}-\frac{1}{2s^2}.$$

Using Table 1 and Corollary 2.3 once more, we find that

$$y(t)=\frac{e^{2t}}{12}-\frac{2e^{-t}}{3}+\frac{3}{4}-\frac{t}{2}.$$

(Alternatively, recalling the original differential equation, we could have computed y as $y=\dot{x}-x$.) It is worth-while to check that the formulas we have derived for x and y do indeed give solutions to our initial value problem. We leave this to the reader.

2.3.3. Linear differential equations with non-constant coefficients.

Example 2.7. Consider the following initial value problem:

$$\begin{cases} \text{differential equation: } t\ddot{x}(t)+(1-t)\dot{x}(t)+2x(t) \overset{!}{=} 0 \text{ for } t>0, \\ \text{initial conditions: } \begin{cases} \dot{x}(0) \overset{!}{=} -2, \\ x(0) \overset{!}{=} 1. \end{cases} \end{cases} \tag{2.7}$$

Note that the differential equation here has *non-constant* coefficients. The Laplace transform can also be used to solve this problem. Indeed, if x is a solution, then

$$\begin{aligned} (s\mapsto 0) &= \mathscr{L}\{0\} = \mathscr{L}\{t\mapsto t\ddot{x}(t)+(1-t)\dot{x}(t)+2x(t)\} \\ &= \mathscr{L}\{t\mapsto t\ddot{x}(t)\}+\mathscr{L}\{\dot{x}\}-\mathscr{L}\{t\mapsto t\dot{x}(t)\}+2\mathscr{L}\{x\}. \end{aligned}$$

Making use of Table 1, we find that

$$(s\mapsto 0) = -\mathscr{L}\{\ddot{x}\}'+\mathscr{L}\{\dot{x}\}+\mathscr{L}\{\dot{x}\}'+2\mathscr{L}\{x\}.$$

Now,

$$\begin{aligned} 0 &= -\frac{\mathrm{d}}{\mathrm{d}s}(s^2\mathscr{L}\{x\}(s)-sx(0)-\dot{x}(0))+(s\mathscr{L}\{x\}(s)-x(0))+ \\ &\quad +\frac{\mathrm{d}}{\mathrm{d}s}(s\mathscr{L}\{x\}(s)-x(0))+2\mathscr{L}\{x\}(s) \\ &= -2s\mathscr{L}\{x\}(s)-s^2\mathscr{L}\{x\}'(s)+x(0)+s\mathscr{L}\{x\}(s)-x(0)+ \\ &\quad +\mathscr{L}\{x\}(s)+s\mathscr{L}\{x\}'(s)+2\mathscr{L}\{x\}(s) \\ &= (1-s)s\mathscr{L}\{x\}'(s)+(3-s)\mathscr{L}\{x\}(s). \end{aligned}$$

Thus, letting $X=\mathscr{L}\{x\}$, we have

$$X'(s)+\frac{3-s}{(1-s)s}X(s)=0. \tag{2.8}$$

Let F be any function. Then $\exp(F(s))$ is strictly positive and multiplying the above equation with this factor does not change its solutions (original solutions remain solutions and no new solutions arise). We choose F as an anti-derivative of $\frac{3-s}{(1-s)s}$, i.e., $F'(s)=\frac{3-s}{(1-s)s}$. We arrive at the equation

$$X'(s)e^{F(s)}+F'(s)X(s)e^{F(s)}=0.$$

The left hand side of this is, by the product rule for differentiation,

$$=\frac{\mathrm{d}}{\mathrm{d}s}\big(X(s)e^{F(s)}\big).$$

Hence,

$$\frac{\mathrm{d}}{\mathrm{d}s}\big(X(s)e^{F(s)}\big)=0,$$

meaning that $X(s)e^{F(s)}$ is constant, getting $X(s) = ce^{-F(s)}$ for some constant $c \in \mathbb{R}$. To find F we do a partial fraction decomposition (see Proposition 2.8 below), getting

$$\frac{3-s}{(1-s)s} = \frac{3}{s} - \frac{2}{s-1}.$$

Integration yields

$$\begin{aligned} F(s) &= \int \frac{3-s}{(1-s)s}\,\mathrm{d}s = 3\int \frac{1}{s}\,\mathrm{d}s - 2\int \frac{1}{s-1}\,\mathrm{d}s \\ &= 3\log s - 2\log(s-1) + \text{const.} = \log\frac{s^3}{(s-1)^2} + \text{const.} \end{aligned}$$

Therefore (ignoring the constant of integration, which would be absorbed into c anyway),

$$X(s) = c\frac{(s-1)^2}{s^3} = c\frac{s^2-2s+1}{s^3} = \left(\frac{1}{s} - \frac{2}{s^2} + \frac{1}{s^3}\right)c.$$

By means of Table 1, we deduce that

$$x(t) = \left(1 - 2t + \tfrac{1}{2}t^2\right)c.$$

Hence, $x(0) = c \overset{!}{=} 1$ (by our initial condition). We could be happy at this point, but after such a lengthy calculation (with some slightly shady steps in-between) it seems worth-while to check that we have ended up with the correct solution. To this end, observe that $\dot{x}(t) = (-2+t)c = t-2$, so that truly $x'(0) = -2$, matching our initial condition on $\dot{x}$. Moreover,

$$t\ddot{x}(t) + (1-t)\dot{x}(t) + 2x = t + (1-t)(t-2) + 2(1 - 2t + \tfrac{1}{2}t^2)$$

and a quick calculation shows that this is indeed $= 0$.

Remark. The cleverly chosen function F in Example 2.7 is called an ***integrating factor*** of the differential equation (2.8). Observe that the initial application of the Laplace transform to the initial value problem (2.7) served to transform the second-order differential equation into a first-order differential equation which we were able to solve directly. Obviously the problem in the example was meticulously chosen as to produce a nice solution at the end and we would have been in trouble if the term $t\ddot{x}(t)$ were replaced by $t^2\ddot{x}(t)$, say. However, the upshot here is that occasionally the Laplace transform can help to simplify a problem.

2.4. Dirac delta distribution

2.4.1. A model problem. Recall the model of a mass on a spring, ignoring friction, from § 1.1 (see Remark 1.1). We imagine leaving the mass at the equilibrium position at time $t = 0$ with zero velocity. If nothing else were to happen, then the solution for this problem would be the constant function $t \mapsto 0$ (the mass does not move away from the equilibrium position at any time).

Now we imagine striking the mass with a hammer at time $t_* > 0$ in such a way that the hammer loses all of its momentum at the impact instantaneously (momentum is mass times velocity; thus, by Newton's second law, its derivative with respect to time is force). Upon abstracting away all constants, we are lead to consider the differential equation

$$\ddot{x} = -x + \boxed{\ ?\ }, \tag{2.9}$$

where the "$\boxed{\ ?\ }$" is meant to model the impulse transmitted from the impact of the hammer. Thus, we should like to have

$$\int_{[t_0,t_1]} \boxed{\ ?\ }\,\mathrm{d}t = \begin{cases} 1 & \text{if } t_* \in [t_0, t_1], \\ 0 & \text{otherwise.} \end{cases}$$

One can see that there is no sensible function which could make this work. However, to get something approximating this, for $\epsilon > 0$, we look at the functions

$$\delta_{\epsilon,t_*} \colon \mathbb{R} \to \mathbb{R}, \quad t \mapsto \frac{1}{\sqrt{2\pi\epsilon}} \exp\left(-\frac{(t-t_*)^2}{2\epsilon}\right).$$

The constant in front of the exponential is chosen such that

$$\int_{-\infty}^{\infty} \delta_{\epsilon,t_*}(t)\,\mathrm{d}t = 1$$

(see Example 7.3 below; such integrals may seem familiar from statistics in the context of normal distributions). Note that the function δ_{ϵ,t_*} has a peak of height $1/\sqrt{2\pi\epsilon}$ ($\to \infty$ as $\epsilon \searrow 0$) at $t = t_*$ and decays exponentially everywhere else.

Upon substituting δ_{ϵ,t_*} for "$\boxed{\ ?\ }$" in (2.9) and rearranging slightly, we now aim to solve the initial value problem

$$(\mathrm{IVP}_\epsilon) \qquad \begin{cases} \text{differential equation: } \ddot{x}_\epsilon + x_\epsilon \stackrel{!}{=} \delta_{\epsilon,t_*} \text{ on } \mathbb{R}_+, \\ \text{initial conditions: } \begin{cases} \dot{x}_\epsilon(0) \stackrel{!}{=} 0, \\ x_\epsilon(0) \stackrel{!}{=} 0. \end{cases} \end{cases}$$

Taking the Laplace transform, we get

$$\mathscr{L}\{\ddot{x}_\epsilon\} + \mathscr{L}\{x_\epsilon\} = \mathscr{L}\{\delta_{\epsilon,t_*}\}. \tag{2.10}$$

Using our initial conditions, we have

$$\mathscr{L}\{\ddot{x}_\epsilon\}(s) = s^2\mathscr{L}\{x_\epsilon\}(s) - sx_\epsilon(0) - \dot{x}_\epsilon(0) = s^2\mathscr{L}\{x_\epsilon\}(s). \tag{2.11}$$

On the other hand, computing

$$\mathscr{L}\{\delta_{\epsilon,t_*}\} = \int_0^\infty \delta_{\epsilon,t_*} e^{-st}\,\mathrm{d}t$$

may be hard. However, at this point we take $\epsilon \searrow 0$ to model the hammer–mass impulse transmission being instantaneous. The idea here is that the initial value

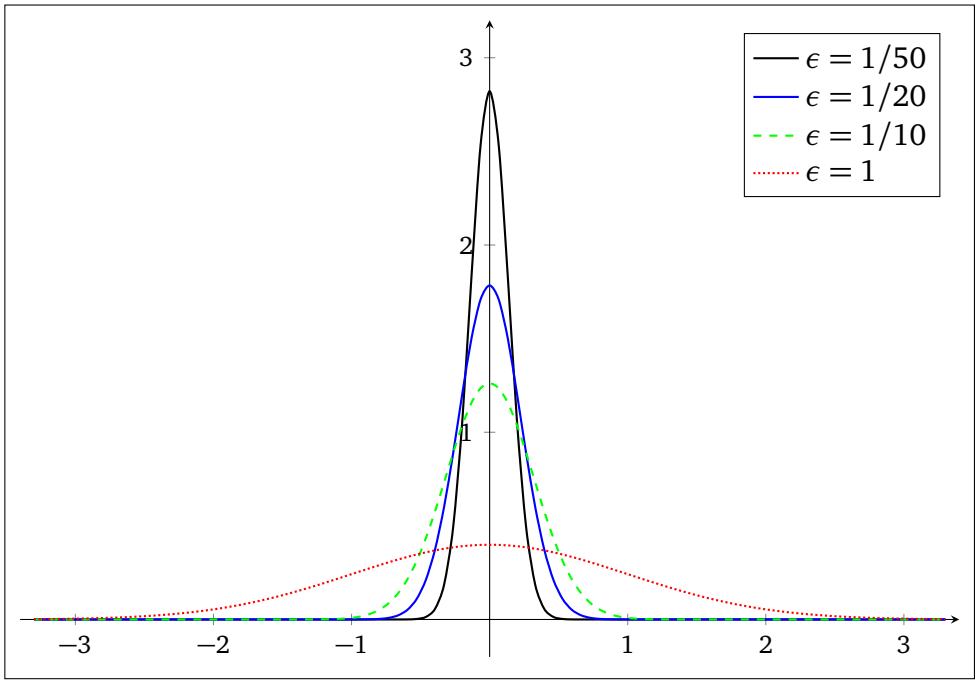


Figure 18. Plot of $\delta_{\epsilon,0}$ with $\epsilon = 1, \frac{1}{10}, \frac{1}{20}, \frac{1}{50}$.

Picture removed due to copyright concerns.
Visit the following website instead:
`https://www.bundesbank.de/en/tasks/cash-management/dm-banknotes-and-coins/-/dm-banknotes-623628`

Figure 19. Picture of the 10 Deutsche Mark bank note of 1991 showing C. F. Gauß and the normal distribution (which also bears his name).

problem (IVP$_\epsilon$) admits a solution x_ϵ for every $\epsilon > 0$ and as $\epsilon \searrow 0$ these solutions *should* converge to the "true" solution x of our physical problem that we are looking for. Similarly, we should expect that the Laplace-transformed solutions $\mathscr{L}\{x_\epsilon\}$ should converge to the Laplace transform $\mathscr{L}\{x\}$ of the desired solution. Then, trying to

invert things, one hopes to recover x. This leads us to look at[6]

$$\lim_{\epsilon \searrow 0} \mathscr{L}\{\delta_{\epsilon,t_*}\}(s) = \lim_{\epsilon \searrow 0} \int_0^\infty \delta_{\epsilon,t_*} e^{-st}\,\mathrm{d}t = e^{-st_*}.$$

(Proving the last equation would not be outrageously involved, but we refrain from doing so nonetheless.) Now, plugging (2.11) into (2.10) and assuming that taking $\epsilon \searrow 0$ lets us replace x_ϵ by x, we arrive at

$$(s^2+1)\mathscr{L}\{x\}(s) \stackrel{?}{=} (s^2+1)\lim_{\epsilon \searrow 0} \mathscr{L}\{x_\epsilon\}(s) \stackrel{?}{=} \lim_{\epsilon \searrow 0} \mathscr{L}\{\delta_{\epsilon,t_*}\}(s) = e^{-st_*}.$$

Hence, we are looking for a function $x\colon [0,\infty) \to \mathbb{R}$ with

$$\mathscr{L}\{x\}(s) = \frac{e^{-st_*}}{s^2+1}.$$

From Proposition 2.4 we conclude that

$$\frac{e^{-st_*}}{s^2+1} = e^{-st_*}\frac{1}{s^2+1} = \mathscr{L}\{\sin_{\to t_*}\}.$$

Therefore, the sought-after solution is

$$x(t) = \sin_{\to t_*}(t) = \begin{cases} \sin(t-t_*) & \text{if } t \geq t_*, \\ 0 & \text{if } t < t_*. \end{cases} \tag{2.12}$$

Note that this solution satisfies the initial conditions $x(0) = x'(0) = 0$ and satisfies the differential equation $\ddot{x}(t)+x(t) = 0$ at every point $t \neq t_*$. It describes the problem we were modelling: for $t < t_*$ we have $x(0) = x'(0) = 0$, i.e., the mass on the spring is at the equilibrium position and has zero velocity. Then, jumping from t very close to t_*, but smaller than t_*, to t very close to t_*, but larger than t_*, the mass accelerates sort of instantaneously to speed (essentially) one and then proceeds to move undisturbed according to how masses on idealised springs move (see Figure 20).

2.4.2. Discussion and sketch of distribution theory. Physicists are usually happy at this point, and we should not blame them too harshly: after all, they know much better than us that the above model is only an idealised situation of reality anyway and the assumption that the impulse is transmitted instantaneously was wrong to begin with. (Modelling the impact of the hammer would involve shock waves travelling through the material etc.) Hence, we should be content the *something* resembling a solution was obtained. Nevertheless, we point out that the x found above is actually not differentiable at t_*! In particular, it *does not* constitute a classical solution to the type of initial value problem we sought to solve.

[6]Recall that we have assumed that $t_* > 0$. For $t_* < 0$ we would get zero here and for $t_* = 0$ a factor $1/2$ would have to be added to the right hand side of the equation, for the integration is cut-off to the left of 0 and then only 'catches half of the mass' placed at 0 by $\delta, \epsilon, 0$. The same phenomenon is responsible for the special behaviour for $t = 0$ in Theorem 2.2.

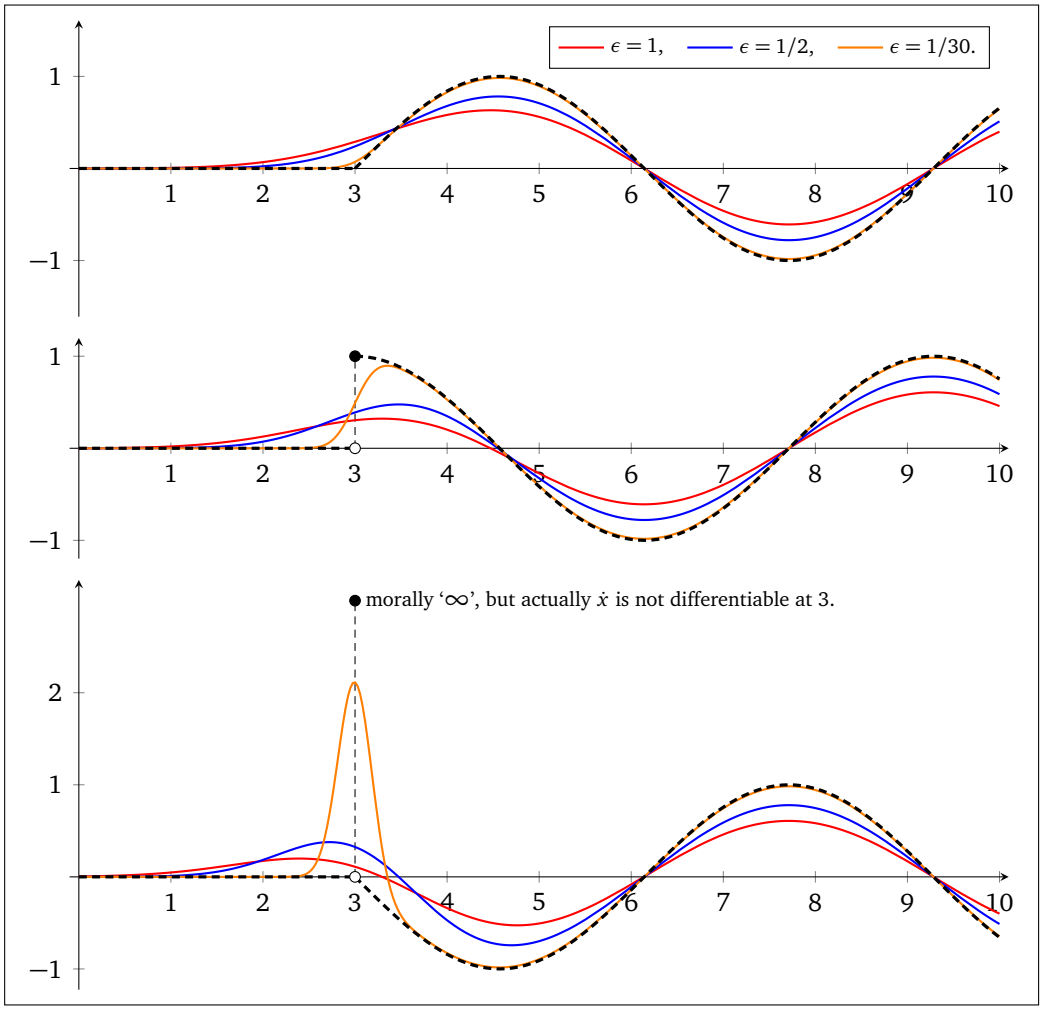


Figure 20. First row: solutions x_ϵ to the initial value problem (IVP$_\epsilon$) with $\epsilon = 1, \frac{1}{2}, \frac{1}{30}$ and $t_* = 3$, and the 'solution' x obtained in (2.12) (dashed graph). Observe that x is not differentiable at t_*. The second and third rows show the first and second derivative of the functions plotted in the first row.

There is a whole theory, called the ***theory of distributions***, due to the French mathematician Laurent Schwartz. This theory was invented to rectify problems such as the above. The general strategy is to enlarge the space of functions in which one is looking for solutions to differential equations of interest. We shall sketch the basic idea in the setting of real-valued functions defined on $\mathbb{R}$. Any suitable function $f : \mathbb{R} \to \mathbb{R}$ gives rise to an integral operator

$$\mathscr{I}\{f\} : g \mapsto \int_{-\infty}^{\infty} f(t)g(t)\,\mathrm{d}t$$

taking functions $g\colon \mathbb{R} \to \mathbb{R}$ to the above integral. For reasons of convergence and for making other constructions work (which we will not discuss, though) one restricts to functions g that vanish outside of a bounded set and are infinitely often differentiable. One can check that the map

$$\{\text{continuous functions } f\colon \mathbb{R} \to \mathbb{R}\} \longrightarrow \{\text{integral operators as above}\}$$

is injective. In that sense we may identify f with its associated integral operator $\mathscr{I}\{f\}$. On the other hand, there are other linear operators acting on functions g as above. For instance, the point evaluation operator

$$\delta\colon g \mapsto g(0)$$

is one such operator and it is *not* of the form $\mathscr{I}\{f\}$. If one sets things up right, then these linear operators[7] can be equipped with a notion of convergence and one gets

$$\lim_{\epsilon \searrow \delta} \delta_{\epsilon,0} = \delta. \tag{2.13}$$

The act of applying such an operator is often written—by abuse of notation!—as an integral as well:

$$g(0) = \delta(g) =: \text{“} \int_{-\infty}^{\infty} \delta(t) g(t)\,\mathrm{d}t \text{”}.$$

The operator δ is called the ***Dirac delta distribution***.

The reader hopefully has gotten the impression that the underlying details are quite delicate. The explanation of this in physics or engineering classes is often via dirty lies, claiming something like the following paragraph:

The Dirac delta function δ is defined to be

$$\delta(t) := \lim_{\epsilon \searrow 0} \frac{1}{\sqrt{2\pi\epsilon}} \exp\left(-\frac{t^2}{2\epsilon}\right) = \begin{cases} \infty & \text{if } t = 0, \\ 0 & \text{if } t \neq 0, \end{cases} \tag{2.14}$$

where “∞” is “so large” such that $\displaystyle\int_{-\infty}^{\infty} \delta(t)\,\mathrm{d}t = 1.$

The above is mathematical ***nonsense***. If you are given such an explanation in class, please *do not be rude* and *do not ask* the lecturer how the above integral is supposed to be defined, or whether

$$\lim_{\epsilon \searrow 0} 2\frac{1}{\sqrt{2\pi\epsilon}} \exp\left(-\frac{t^2}{2\epsilon}\right) = \begin{cases} \infty & \text{if } t = 0, \\ 0 & \text{if } t \neq 0, \end{cases} \tag{2.15}$$

equals $\delta(t)$ or $2\delta(t)$. Is the “∞” in (2.15) different from the one in (2.14)? The lecturer is unlikely to have a good answer to such questions. It should be noted,

[7] Actually, we should speak of “continuous” linear operators here and explain what that means, but we shall not do this.

however, that there is a way in which one can define convergence of operators (on functions) in which the integral operators (distributions) induced by the functions

$$\mathbb{R} \to \mathbb{R}, \quad t \mapsto \frac{1}{\sqrt{2\pi\epsilon}} \exp\left(-\frac{t^2}{2\epsilon}\right),$$

(for varying ϵ) converge to the Dirac delta distribution. In this sense one may write

$$\lim_{\epsilon \searrow 0} \mathscr{I}\left\{ t \mapsto \frac{1}{\sqrt{2\pi\epsilon}} \exp\left(-\frac{t^2}{2\epsilon}\right)\right\} = \delta.$$

One may view this as the 'actual' meaning of (2.14).

2.5. Partial fraction decomposition

We give a quick recapitulation of the well-known technique of ***partial fraction decomposition***. The reader should already be familiar with this from [3]. The utility of partial fraction decomposition lies in it reducing the problem of computing certain integrals involving rational functions (quotients of polynomials) to the computation of integrals involving simpler rational functions. As the Laplace transform is defined using an integral, the relevance of partial fraction decomposition for the purpose of this chapter should be quite obvious. The general result which forms the theoretical foundation for partial fraction decomposition is given as follows:

Proposition 2.8 (Partial fraction decomposition)**.** *Let*

$$\frac{R(X)}{(X-x_1)^{\nu_1}\cdots(X-x_k)^{\nu_k}}$$

be a quotient of two polynomials (with real or complex coefficients) where the roots $x_1,\dots,x_k \in \mathbb{C}$ of the denominator are pairwise distinct and $\nu_1,\dots,\nu_k \in \mathbb{N}$ are exponents. Suppose that the degree of the polynomial in the numerator is strictly smaller than the degree of the denominator. Then there exist complex numbers $A_{r,\nu}$ such that

$$\frac{R(X)}{(X-x_1)^{\nu_1}\cdots(X-x_k)^{\nu_k}} = \sum_{r=1}^{k}\sum_{\nu=1}^{\nu_r} \frac{A_{r,\nu}}{(X-x_r)^{\nu}}. \tag{2.16}$$

Example. We spell out (2.16) in one concrete example, in order to make the notation with the sums somewhat more transparent. By (2.16) we have

$$\frac{1}{(X-8)(X+5)^3} = \frac{A_{1,1}}{X-8} + \frac{A_{2,1}}{(X-5)^1} + \frac{A_{2,2}}{(X-5)^2} + \frac{A_{2,3}}{(X-5)^3},$$

for suitable complex numbers $A_{1,1}, A_{2,1}, A_{2,2}, A_{2,3}$. With a bit of work, one can check that

$$(A_{1,1}, A_{2,1}, A_{2,2}, A_{2,3}) = (1/2197, -1/2197, -1/169, -1/13)$$

does the job. In Example 2.10 below, we show how one can find such numbers, albeit on a computationally less exhausting example.

We start with a list of examples to illustrate what the formula (2.16) actually *says*. In the subsequent example (Example 2.10) we demonstrate how one could go about *finding* the right hand side of (2.16) in practice.

Example 2.9.

- $\dfrac{1}{X(X+1)} = \dfrac{1}{X} - \dfrac{1}{X+1}$,
- $\dfrac{X+2}{X(X+1)} = \dfrac{2}{X} - \dfrac{1}{X+1}$,
- $\dfrac{X+2}{X^2(X+1)} = \dfrac{2}{X^2} + \dfrac{1}{X+1} - \dfrac{1}{X}$,
- $\dfrac{X^6+X^5+X+2}{X^2(X+1)} = \dfrac{X^3 \cdot X^2(X+1)+(X+2)}{X^2(X+1)} = X^3 + \dfrac{2}{X^2} + \dfrac{1}{X+1} - \dfrac{1}{X}$.

Example 2.10. The equations in the previous example may be *checked* easily by hand. However, they can also be *obtained* by hand. For instance, to arrive at the expansion

$$\frac{X+2}{X(X+1)} = \frac{2}{X} - \frac{1}{X+1},$$

we note that Proposition 2.8 ensures that there *are* numbers $A, B \in \mathbb{C}$ such that

$$\frac{X+2}{X(X+1)} = \frac{A}{X} + \frac{B}{X+1},$$

and it remains to find them. To this end, multiplying by the denominator of the right hand side, we arrive at

$$X+2 = A(X+1) + BX.$$

Plugging in -1 for X yields the equation $-1+2 = A \cdot 0 + B \cdot (-1)$ and plugging in 0 for X yields the equation $0+2 = A \cdot (0+1) = B \cdot 0$. This immediately shows that $(A, B) = (2, -1)$, as expected.

The general recipe for finding a partial fraction decomposition of $P(X)/Q(X)$ is as follows:

(1) Do a polynomial division to write $P(X) = P_*(X)Q(X)+R(X)$ with polynomials $P_*, R \in \mathbb{C}[X]$ such that R has degree strictly less than Q. Then

$$\frac{P(X)}{Q(X)} = P_*(X) + \frac{R(X)}{Q(X)}$$

and it remains to find the partial fraction decomposition of the second term on the right hand side.

(2) Factor the polynomial Q in the way $Q = c(X - x_1)^{\nu_1} \cdots (X - x_k)^{\nu_k}$ with some non-zero $c \in \mathbb{C}$ and distinct $x_1, \ldots, x_k \in \mathbb{C}$. (This step may be hard depending on the polynomial.)

(3) Write out the general form

$$\frac{R(X)}{(X-x_1)^{\nu_1}\cdots(X-x_k)^{\nu_k}} = \sum_{r=1}^{k}\sum_{\nu=1}^{\nu_r}\frac{A_{r,\nu}}{(X-x_r)^{\nu}}$$

of the desired partial fraction decomposition, multiply through by the denominator of the right hand side and plug in sufficiently many distinct values for X as to obtain enough independent linear(!) equations and solve these for the unknowns $A_{r,\nu}$. (Plugging in $x_1,\ldots,x_k$ for X is particularly convenient, because this makes certain parts of the equations vanish. If the exponents $\nu_\bullet$ are not all 1, however, more particular choices for X will be necessary to fully determine the $A_{r,\nu}$.)

CHAPTER 3

Linear algebra

Linear algebra may initially seem like a lot of language with the express purpose of solving systems of linear equations. This only scratches the surface of what is ultimately a theory closely intertwined with the geometry of space and it is ultimately only the algebraic approach that makes solving problems feasible. For that reason, calculus—the mathematics of change—is based around the concept of *linearisation*. Although analysis in one dimension can certainly be understood without much background in linear algebra, the rising complexity of calculus in higher dimensions necessitates a firm understanding of the relevant "linearised" situation. (Note that calculus was invented originally for studying physics.)

Having readers in mind who have already had some exposure to linear algebra mostly from an algorithmic perspective, our exposition focuses more strongly on the geometrical flavour of the subject. We shall mostly restrict ourselves to $n = 1, 2, 3$ for concreteness, but everything we discuss here as appropriate (and often straightforward) generalisations to higher dimensions. Even more so: many of the present concepts even generalise to an infinite-dimensional setting, although in such cases the algebraic aspects must be subsidised by topological–analytic considerations in order to built a viable theory. We shall get a glimpse at this later when we deal with Fourier series in Chapter 4.

Moreover, our approach will mainly be via pre-chosen coordinates. An expert would rightfully say that this is utterly the wrong way for understanding the theory and truly appreciating the geometrical nature thereof, but it seems adequate for the purpose of getting to some important concepts like determinants and dot products quickly. Unfortunately, this approach does make our later discussion of eigenvectors arguably rather awkward.

3.1. Vectors, linear maps and matrices

3.1.1. Vectors. Vectors are used to describe points in space or (the physicist's view: *directions* in space). Formally, any element of $\mathbb{R}^n$ (or $\mathbb{C}^n$) with $n \in \mathbb{N}_0$ is called a ***vector***. A vector $\vec{v} \in \mathbb{R}^3$ is given by its coordinates

$$\vec{v} = \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix}.$$

For notational convenience we also often write $\vec{v} = (v_1, v_2, v_3)$. When speaking of a vector $\vec{v} \in \mathbb{R}^n$ it shall be understood implicitly that there are associated numbers $v_1, \dots, v_n \in \mathbb{R}$ (the coordinates of $\vec{v}$) as above. The same applies when using letters other than v, of course. $\vec{0} = (0, \dots, 0) \in \mathbb{R}^n$ is the ***zero vector*** in $\mathbb{R}^n$. Furthermore, for $j = 1, \dots, n$, we let $\vec{e}_j$ denote the ***j-th standard unit vector***. This is the vector in $\mathbb{R}^n$ whose coordinates are all zero, apart from the j-th coordinate, which equals one. Observe that both $\vec{0}$ and $\vec{e}_j$ implicitly depend on n. To make this clear, we spell it out for small n:

Examples.

- $n = 1$: $\vec{0} = \begin{pmatrix}0\end{pmatrix}$, $\vec{e}_1 = \begin{pmatrix}1\end{pmatrix}$.
- $n = 2$: $\vec{0} = \begin{pmatrix}0\\0\end{pmatrix}$, $\vec{e}_1 = \begin{pmatrix}1\\0\end{pmatrix}$, $\vec{e}_2 = \begin{pmatrix}0\\1\end{pmatrix}$.
- $n = 2$: $\vec{0} = \begin{pmatrix}0\\0\\0\end{pmatrix}$, $\vec{e}_1 = \begin{pmatrix}1\\0\\0\end{pmatrix}$, $\vec{e}_2 = \begin{pmatrix}0\\1\\0\end{pmatrix}$, $\vec{e}_3 = \begin{pmatrix}0\\0\\1\end{pmatrix}$.

Vectors $\vec{v}, \vec{w} \in \mathbb{R}^3$ can be ***added and multiplied*** with real (or complex) numbers λ (referred to as ***scalars*** in this context):

$$\vec{v} + \vec{w} = \begin{pmatrix}v_1\\v_2\\v_3\end{pmatrix} + \begin{pmatrix}w_1\\w_2\\w_3\end{pmatrix} := \begin{pmatrix}v_1 + w_1\\v_2 + w_2\\v_3 + w_3\end{pmatrix}, \quad \lambda\vec{v} := \begin{pmatrix}\lambda v_1\\\lambda v_2\\\lambda v_3\end{pmatrix}.$$

(Similar definitions apply in dimensions other than $n = 3$.) One has the usual rules for vector operations: for $\vec{v}, \vec{w}, \vec{z} \in \mathbb{R}^n$ and scalars $\lambda, \mu \in \mathbb{R}$ one has

- $\left.\begin{array}{l}(\vec{v} + \vec{w}) + \vec{z} = \vec{v} + (\vec{w} + \vec{z}),\\(\lambda \cdot \mu)\vec{v} = \lambda \cdot (\mu\vec{v}),\end{array}\right\}$ (associative laws)
- $\vec{v} + \vec{w} = \vec{w} + \vec{v}$, (commutative law)
- $\left.\begin{array}{l}(\lambda + \mu)\vec{v} = \lambda\vec{v} + \mu\vec{v},\\\lambda(\vec{v} + \vec{w}) = \lambda\vec{v} + \lambda\vec{w},\end{array}\right\}$ (distributive laws)
- $\vec{v} + \vec{0} = \vec{0} + \vec{v} = \vec{v}$ and $\lambda\vec{0} = \vec{0}$.

For any vectors $\vec{v}, \vec{w} \in \mathbb{R}^n$, vectors of the form $\lambda\vec{v} + \mu\vec{w}$ (with $\lambda, \mu \in \mathbb{R}$) are called ***linear combinations*** of $\vec{v}$ and $\vec{w}$. This concept generalises in the obvious way to sums of more than two vectors.

Remark. Mathematicians like to build the theory of linear algebra with objects more general than $\mathbb{R}^n$. We give a toy explanation here for why this is done. (The real reason is that experience has shown that linear algebra, when developed in sufficiently general terms, sheds light on a huge amount of phenomena, with the added benefit of being able to exploit tools, initially developed for something else, to solve new problems.) Imagine trying to describe to someone how you paint something on a canvas. Clearly we may think of the canvas as sitting in three-dimensional space and points on it can be described by vectors with three coordinates (subject to some arbitrarily chosen origin $\vec{0}$). However, the canvas, on which we would like to paint,

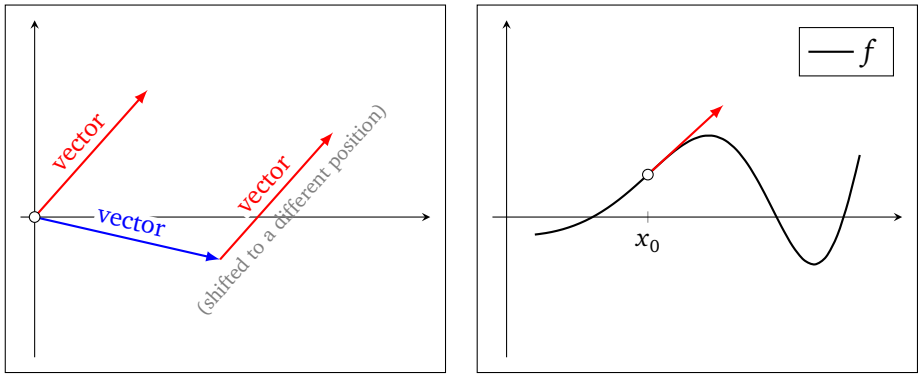
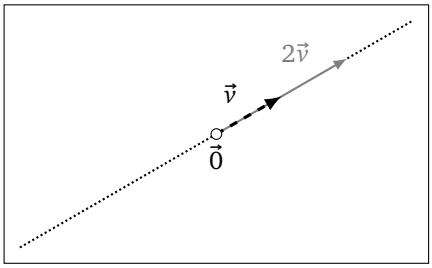


Figure 21. Illustration regarding how we imagine vectors: generally, as points in space and arrows anchored at zero and pointing towards them, but often we draw vectors anchored at different points. For instance, we visualise the derivative of a function $f : \mathbb{R} \to \mathbb{R}$ at a point x_0 by drawing the vector $(1, f'(x_0))$ attached to the point $(x_0, f(x_0))$.

also sits on a plane ('flat surface'; not an 'air plane'). Hence, we should be able to do with just two coordinates instead of three. A two-dimensional version of this situation is illustrated below:



Our above example shows that even in three-dimensional space we may encounter situations where we want to study lower-dimensional objects, but may wish to treat them 'in their own right'. Consequently, one may be inclined to consider some more general notion of 'space' which includes all spaces $\mathbb{R}^n$, but also situations involving our canvas. One such notion is that of a ***vector space***. We shall not, however, pursue this here, and refer to more specialised texts on linear algebra instead.

3.1.2. Linear maps and matrices. Maps $f : \mathbb{R}^n \to \mathbb{R}^m$ that respect vector addition and scalar multiplication are called ***linear***, namely for such a map to be called linear, one requires

$$f(\vec{v} + \vec{w}) = f(\vec{v}) + f(\vec{w}) \quad \text{and} \quad f(\lambda\vec{v}) = \lambda f(\vec{v})$$

to hold for any vectors $\vec{v}, \vec{w} \in \mathbb{R}^n$ and any scalar $\lambda \in \mathbb{R}$.

Example.

- For any fixed $c \in \mathbb{R}$ and any $n \in \mathbb{N}$, the map $\mathbb{R}^n \to \mathbb{R}^n$, $\vec{v} \mapsto c\vec{v}$, is linear.

- The map $f\colon \mathbb{R}^1 \to \mathbb{R}^1$, $v \mapsto v^2$ is not(!) linear. ($f(1+1) = f(2) = 4 \neq 2 = f(1) + f(1)$.)
- For any differentiable function $f\colon \mathbb{R} \to \mathbb{R}$ and any point $x_0 \in \mathbb{R}$, the map $v \mapsto f'(x_0)v$ is linear. We have the approximation

$$f(x_0 + \underbrace{v}_{\text{distortion}}) \approx \overbrace{f(x_0)}^{\text{point evaluation}} + \underbrace{f'(x_0)v}_{\text{linearisation evaluated at distortion}}.$$

  This and especially its higher-dimensional generalisations are our primary reason for considering linear maps. (Analysis $\approx$ "studying functions via linearisation.")
- The map $\mathbb{R}^3 \to \mathbb{R}^2$, $\begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix} \mapsto \begin{pmatrix} v_1 \\ v_1 - v_3 \end{pmatrix}$, is linear.
- The map $g\colon \mathbb{R}^3 \to \mathbb{R}^2$, $\begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix} \mapsto \begin{pmatrix} v_1 + 42 \\ v_1 - v_3 \end{pmatrix}$, is not linear. (Any linear map $f\colon \mathbb{R}^n \to \mathbb{R}^m$ must map the zero vector $\vec{0} \in \mathbb{R}^n$ to the zero vector $\vec{0} \in \mathbb{R}^m$, because of $f(\vec{0}) = f(0\,\vec{0}) = 0f(\vec{0}) = \vec{0}$. However, $g(0,0,0) = (42,0) \neq \vec{0}$.)

Linear maps are nice, because they are easy to work with. This is due to them being quite "rigid". As an example for this, note that a linear map $f\colon \mathbb{R}^3 \to \mathbb{R}^2$ is already determined by its values on the three(!)[1] vectors

$$\vec{e}_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \quad \vec{e}_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \quad \vec{e}_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}.$$

Indeed,

$$\vec{v} = \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix} = v_1\vec{e}_1 + v_2\vec{e}_2 + v_3\vec{e}_3,$$

and, by linearity,

$$f(\vec{v}) = v_1 f(\vec{e}_1) + v_2 f(\vec{e}_2) + v_3 f(\vec{e}_3).$$

To go even further, note that $f(\vec{e}_1)$ is a vector in $\mathbb{R}^2$ and can be expanded as a sum of scalar multiples of $\vec{e}_1, \vec{e}_2 \in \mathbb{R}^2$. (Note that in the first part of the previous sentence $\vec{e}_1$ has a different meaning than at its end: the former refers to the vector $\vec{e}_1 = (1,0,0)$ whereas the latter is $\vec{e}_1 = (1,0)$.) Thus, we need two real numbers

[1] The surprising point here is that the number in question is *finite*. Indeed, to determine an arbitrary map $f\colon \mathbb{R}^3 \to \mathbb{R}^2$ (not assumed to be linear!) we need to know *all* values $f(\vec{v})$ for all (infinitely many!) vectors $\vec{v}$. Hence, for linear maps, much less information on their values suffices, to determine them.

to completely describe $f(\vec{e}_1)$ and similarly for $f(\vec{e}_2)$ and $f(\vec{e}_3)$. In particular f is completely described by 3×2 numbers. If

$$f\left(\begin{pmatrix}1\\0\\0\end{pmatrix}\right) = \begin{pmatrix}a_{11}\\a_{21}\end{pmatrix}, \quad f\left(\begin{pmatrix}0\\1\\0\end{pmatrix}\right) = \begin{pmatrix}a_{12}\\a_{22}\end{pmatrix}, \quad f\left(\begin{pmatrix}0\\0\\1\end{pmatrix}\right) = \begin{pmatrix}a_{13}\\a_{23}\end{pmatrix}$$

we consider the "rectangular array"

$$A := \begin{pmatrix}a_{11} & a_{12} & a_{13}\\ a_{21} & a_{22} & a_{23}\end{pmatrix} \in \mathbb{R}^{2\times 3}.$$

We shall refer to such rectangular arrays in $\mathbb{R}^{m\times n}$ as ***matrices*** (singular: ***matrix***). If the matrix is denoted by a letter, A say, then its ***entries*** are usually denoted by a_{ij} (same letter in lower case with indices for rows and columns) or, when more convenient, by A_{ij}.

Now let us generalise the above representation process. Suppose that $f\colon \mathbb{R}^n \to \mathbb{R}^m$ is linear. We build the matrix

$$A = \begin{pmatrix} | & \dots & | \\ f(\vec{e}_1) & \dots & f(\vec{e}_n) \\ | & \dots & | \end{pmatrix} \in \mathbb{R}^{m\times n}. \tag{3.1}$$

We say that this is the ***matrix representing f*** (or ***matrix associated with f***).[2]

Example. Consider the linear map

$$f\colon \mathbb{R}^3 \to \mathbb{R}^2, \quad \begin{pmatrix}v_1\\v_2\\v_3\end{pmatrix} \mapsto \begin{pmatrix}v_1\\v_1 - v_3\end{pmatrix} \tag{3.2}$$

Then

$$f\left(\begin{pmatrix}1\\0\\0\end{pmatrix}\right) = \begin{pmatrix}1\\1\end{pmatrix}, \quad f\left(\begin{pmatrix}0\\1\\0\end{pmatrix}\right) = \begin{pmatrix}0\\0\end{pmatrix}, \quad f\left(\begin{pmatrix}0\\0\\1\end{pmatrix}\right) = \begin{pmatrix}0\\-1\end{pmatrix}.$$

Hence f is represented by the matrix $\begin{pmatrix}1 & 0 & 0\\ 1 & 0 & -1\end{pmatrix}$.

Returning to the matrix (3.1) representing f, we seek to *recover* f. That is, given the matrix A from (3.1), we define

$$A\vec{v} := f(\vec{v})$$

and ask how to compute the right hand side if we only know A. To this end, observe that

$$A\vec{v} = f(\vec{v}) = f(v_1\vec{e}_1 + \ldots + v_n\vec{e}_n) = v_1 f(\vec{e}_1) + \ldots + v_n f(\vec{e}_n).$$

[2] At this point, linear algebra texts usually take the very justified view that the choice of vectors $\vec{e}_1, \ldots, \vec{e}_n$ was slightly arbitrary. This yields to the concept of a ***basis*** of a vector space, but we avoid these notions here.

The values $f(\vec{e}_j)$ are the *columns* of A. If A is written in the form

$$A = \begin{pmatrix} a_{11} & \dots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \dots & a_{mn} \end{pmatrix},$$

then

$$A\vec{v} = v_1 \begin{pmatrix} a_{11} \\ \vdots \\ a_{m1} \end{pmatrix} + \dots + v_n \begin{pmatrix} a_{1n} \\ \vdots \\ a_{mn} \end{pmatrix} = \begin{pmatrix} v_1 a_{11} + \dots + v_n a_{1n} \\ \vdots \\ v_1 a_{m1} + \dots + v_n a_{mn} \end{pmatrix}.$$

This is known as ***matrix–vector multiplication***.

Example. Consider the map $f\colon \mathbb{R}^3 \to \mathbb{R}^2$ from the previous example (see (3.2)). It is represented by the matrix

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

Let us see, that our definition of matrix–vector multiplication really recovers f. We have

$$A\vec{v} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & -1 \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix} = \begin{pmatrix} 1v_1 + 0v_2 + 0v_3 \\ 1v_1 + 0v_2 + (-1)v_3 \end{pmatrix} = \begin{pmatrix} v_1 \\ v_1 - v_3 \end{pmatrix}.$$

The right hand side evidently equals $f(\vec{v})$, as desired (see (3.2)).

Our discussion from above shows that knowing f is sufficient for knowing A and, conversely, knowing A is sufficient for knowing f. This yields a 1:1-correspondence between linear maps $\mathbb{R}^n \to \mathbb{R}^m$ and matrices $\mathbb{R}^{m\times n}$. For this reason, one also thinks of matrices as *being* linear maps and *vice versa*.

3.1.3. Composition of linear maps, matrix–matrix multiplication. Suppose now that we are given two linear maps

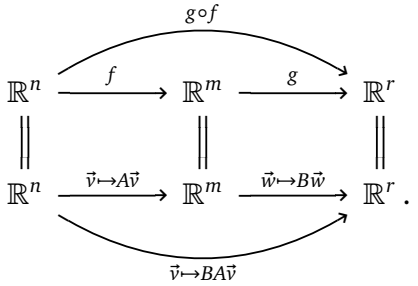
$$\mathbb{R}^n \xrightarrow{\ f\ } \mathbb{R}^m \xrightarrow{\ g\ } \mathbb{R}^r.$$

Then their composition

$$g \circ f\colon \mathbb{R}^n \to \mathbb{R}^r, \quad \vec{v} \mapsto g(f(\vec{v})),$$

is also linear. Suppose that f is represented by $A \in \mathbb{R}^{m\times n}$, g is represented by $B \in \mathbb{R}^{r\times m}$ and $g \circ f$ is represented by $C \in \mathbb{R}^{r\times n}$. We *define* the ***matrix–matrix product*** $BA := B \cdot A := C$. The situation may be visualised in the following commutative

diagram:



As A and B determine f and g respectively, we ought to be able to compute C from just knowing A and B. Note that the j-th column $C_{\bullet j}$ of C is given by

$$C_{\bullet j} = (g \circ f)(\vec{e}_j) = g(f(\vec{e}_j)) = g(A_{\bullet j}),$$

where $A_{\bullet j}$ denotes the j-th column of A. Now

$$A_{\bullet j} = A_{1j}\vec{e}_1 + \ldots + A_{mj}\vec{e}_m,$$

so that

$$C_{\bullet j} = g(A_{1j}\vec{e}_1 + \ldots + A_{mj}\vec{e}_m) = A_{1j}g(\vec{e}_1) + \ldots + A_{mj}g(\vec{e}_m) = A_{1j}B_{\bullet 1} + \ldots + A_{mj}B_{\bullet m}.$$

Consequently, the s-th entry of the j-th column of C is given by

$$C_{sj} = A_{1j}B_{s1} + \ldots + A_{mj}B_{sm} = \sum B_{si}A_{ij}, \quad \text{where} \quad \sum = \sum_{i=1}^{m}.$$

Therefore, we have

$$\begin{pmatrix} B_{11} & \ldots & B_{1m} \\ \vdots & \ddots & \vdots \\ B_{r1} & \ldots & B_{rm} \end{pmatrix} \begin{pmatrix} A_{11} & \ldots & A_{1n} \\ \vdots & \ddots & \vdots \\ A_{m1} & \ldots & A_{mn} \end{pmatrix} = \begin{pmatrix} \sum B_{1i}A_{i1} & \ldots & \sum B_{1i}A_{in} \\ \vdots & \ddots & \vdots \\ \sum B_{ri}A_{i1} & \ldots & \sum B_{ri}A_{in} \end{pmatrix}.$$

This formula can be remembered as 'row × column' via the scheme depicted in Figure 22.

Example. We have

$$\begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 7 & 0 \\ 8 & 0 \end{pmatrix} = \begin{pmatrix} 38 & 1 \\ 83 & 4 \end{pmatrix}.$$

Here, for instance, the entry 38 is computed as $1 \cdot 0 + 2 \cdot 7 + 3 \cdot 8 = 38$.

Next, suppose that we have an additional linear map $h\colon \mathbb{R}^r \to \mathbb{R}^u$. The situation is thus

$$\mathbb{R}^n \xrightarrow{\ f\ } \mathbb{R}^m \xrightarrow{\ g\ } \mathbb{R}^r \xrightarrow{\ h\ } \mathbb{R}^u.$$

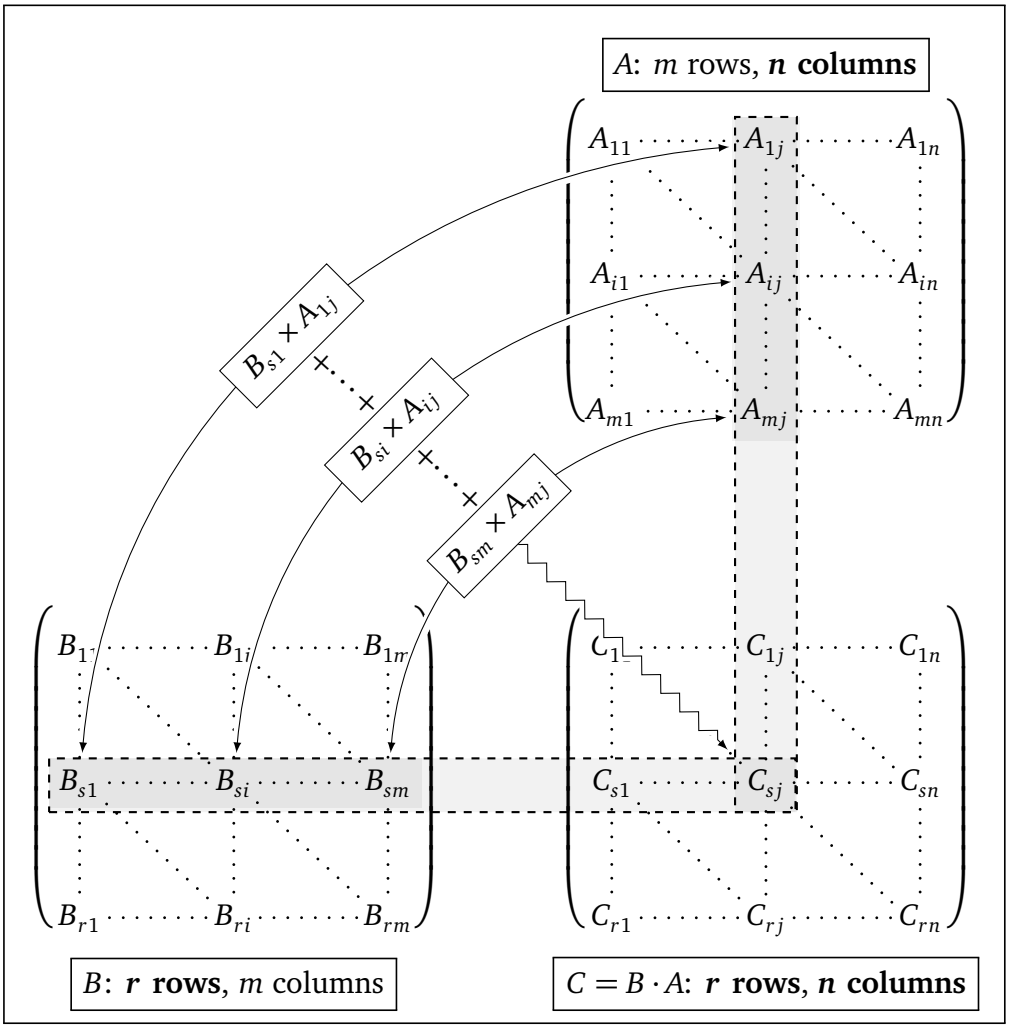
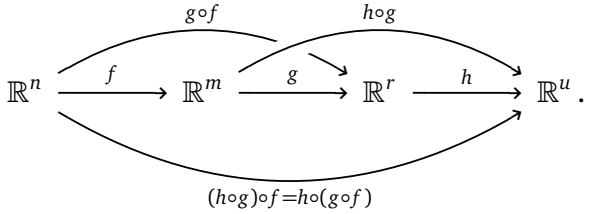


Figure 22. Illustration of matrix multiplication: row × column.
(Adapted from Alain Matthes, `https://www.TeXample.net`.)

There are now two ways of building up a triple composition with the maps f, g and h, namely $(h \circ g) \circ f$ and $h \circ (g \circ f)$. Both ways yield the same result, for they yield the map sending $\vec{v} \in \mathbb{R}^n$ to $h(g(f(\vec{v})))$:



$$A = \begin{pmatrix} 1 & 4 & 7 \\ 2 & 5 & 8 \\ 3 & 6 & 9 \end{pmatrix}, \qquad A^{-1} = \begin{pmatrix} 6 & 9 & 3 \\ 8 & 5 & 2 \\ 7 & 4 & 1 \end{pmatrix}?$$

Figure 23. How to invert a matrix. (Yes, this is a joke. In fact, the matrix A shown here is not even invertible. Can you see why?)

Turning to the level of matrices, and assuming that h is represented by a matrix U, we have

$$(UB)A = U(BA),$$

because the corresponding associativity holds on the level of linear maps.

3.1.4. Addition of linear maps, matrix–matrix sum. Next, consider the situation where f and g are both linear maps from $\mathbb{R}^n \to \mathbb{R}^m$, represented by matrices A and B respectively. The sum $f + g$ of f and g, given by

$$f + g \colon \mathbb{R}^n \to \mathbb{R}^m, \quad \vec{v} \mapsto f(\vec{v}) + g(\vec{v}),$$

is again linear. The matrix representing it is denoted by $A+B$. Similar considerations as with the matrix–matrix product above yield the following explicit formula for $A+B$:

$$A+B = \begin{pmatrix} A_{11} & \dots & A_{1n} \\ \vdots & \ddots & \vdots \\ A_{m1} & \dots & A_{mn} \end{pmatrix} + \begin{pmatrix} B_{11} & \dots & B_{1n} \\ \vdots & \ddots & \vdots \\ B_{m1} & \dots & B_{mn} \end{pmatrix} = \begin{pmatrix} A_{11}+B_{11} & \dots & A_{1n}+B_{1n} \\ \vdots & \ddots & \vdots \\ A_{m1}+B_{m1} & \dots & A_{mn}+B_{mn} \end{pmatrix}.$$

3.1.5. Scaling of linear maps, scaling of matrices. Moreover, for any $\lambda \in \mathbb{R}$, also the map

$$\lambda f \colon \mathbb{R}^n \to \mathbb{R}^m, \quad \vec{v} \mapsto \lambda f(\vec{v}),$$

is linear. The matrix representing it, denoted by λA, can be seen to have the form

$$\lambda A = \lambda \begin{pmatrix} A_{11} & \dots & A_{1n} \\ \vdots & \ddots & \vdots \\ A_{m1} & \dots & A_{mn} \end{pmatrix} = \begin{pmatrix} \lambda A_{11} & \dots & \lambda A_{1n} \\ \vdots & \ddots & \vdots \\ \lambda A_{m1} & \dots & \lambda A_{mn} \end{pmatrix}.$$

The operations of taking the sum of two linear maps or rescaling a linear map obey the same laws (associativity, commutativity, distributivity etc.) as vector addition and scalar multiplication (see § 3.1.1).

3.1.6. Inverse of a linear map, inverse of a matrix. If the linear map $f\colon \mathbb{R}^n \to \mathbb{R}^n$ is bijective, then its inverse function $f^{-1}\colon \mathbb{R}^n \to \mathbb{R}^n$ is also linear. If $A \in \mathbb{R}^{n\times n}$ is representing f, then the matrix in $\mathbb{R}^{n\times n}$ representing f^{-1} is denoted by A^{-1}. It satisfies

$$AA^{-1} = A^{-1}A = \mathbf{1}_n := \begin{pmatrix} 1 & & \\ & \ddots & \\ & & 1 \end{pmatrix},$$

where $\mathbf{1}_n$ is called the unit matrix in $\mathbb{R}^{n\times n}$, having all 1's on the "diagonal" and all other entries being zero, for instance,

$$\mathbf{1}_1 = (1), \quad \mathbf{1}_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad \mathbf{1}_3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Conversely, if any matrix $B \in \mathbb{R}^{n\times n}$ satisfies $AB = \mathbf{1}_n$ or $BA = \mathbf{1}_n$, then A is invertible and $B = A^{-1}$.

3.2. Determinants

What is the 1-dimensional volume (length!) spanned by a vector $\vec{v} = (v_1) \in \mathbb{R}^1$? Well, it is $|v_1|$. When viewing $\vec{v}$ as a 1×1-matrix, we let $\det(v_1) := v_1$. Next, we shall study higher-dimensional versions of this.

3.2.1. Two dimensions. Any two vectors

$$\vec{v} = \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} \quad \text{and} \quad \vec{w} = \begin{pmatrix} w_1 \\ w_2 \end{pmatrix}$$

in $\mathbb{R}^2$ span a parallelogram

$$\square \begin{pmatrix} v_1 & w_1 \\ v_2 & w_2 \end{pmatrix} := \square(\vec{v}, \vec{w}) := \{\, \lambda\vec{v} + \mu\vec{w} : 0 \le \lambda, \mu \le 1 \,\}.$$

(Here we allow for the degenerate case when $\vec{v}$ and $\vec{w}$ both lie on a line through the origin; in this case the aforementioned "parallelogram" takes the shape of a line segment.)

What is the area of $\square(\vec{v}, \vec{w})$? Obviously, for any $\lambda \in \mathbb{R}$,

$$\square(\vec{v}, \vec{w}) \quad \text{and} \quad \square(\vec{v}, \vec{w} - \lambda\vec{v})$$

have the same area (Cavalieri's principle; see Figure 24). We assume that v_1 is non-zero and apply this with $\lambda = w_1/v_1$, getting

$$\text{area}\,\square \begin{pmatrix} v_1 & w_1 \\ v_2 & w_2 \end{pmatrix} \overset{\cdot(-\frac{w_1}{v_1})\ +}{=} \text{area}\,\square \begin{pmatrix} v_1 & 0 \\ v_2 & w_2 - \frac{w_1}{v_1}v_2 \end{pmatrix} = \left| v_1 \left(w_2 - \frac{w_1}{v_1} v_2 \right) \right| = |v_1 w_2 - w_1 v_2|.$$

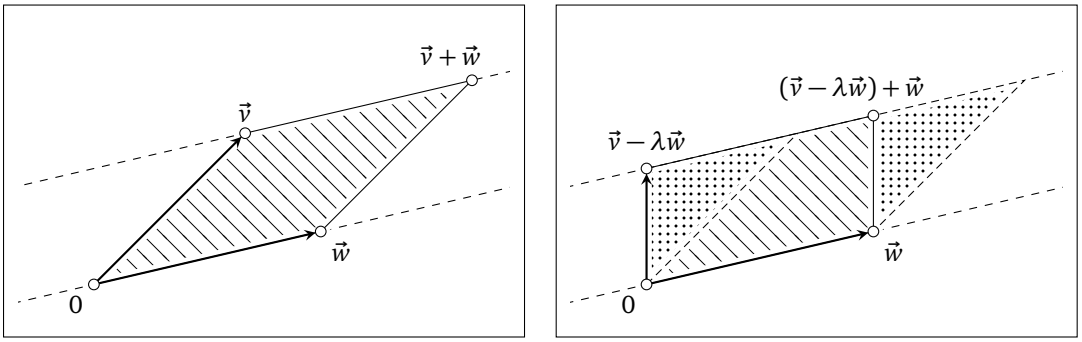


Figure 24. Illustration of Cavalieri's principle which we use for computing the area of $\measuredangle(\vec{v},\vec{w})$.

$$\det\begin{pmatrix} v_1 & w_1 \\ v_2 & w_2 \end{pmatrix} = v_1w_2 - v_2w_1.$$

Figure 25. Mnemonic for the formula (3.3) for determinants of 2×2-matrices.

It can be checked that we would have obtained the same formula if $v_2 \neq 0$ (then taking $\lambda = w_2/v_2$) and if both v_1 and v_2 are zero, then

$$\operatorname{area}\measuredangle(\vec{v},\vec{w}) = 0 = |v_1w_2 - w_1v_2|$$

for trivial reasons. We now let

$$\det\begin{pmatrix} v_1 & w_1 \\ v_2 & w_2 \end{pmatrix} := v_1w_2 - w_1v_2. \tag{3.3}$$

Then, by the above,

$$\left|\det\begin{pmatrix} v_1 & w_1 \\ v_2 & w_2 \end{pmatrix}\right| = \operatorname{area} P(\vec{v},\vec{w}).$$

Examples.

$$\det\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = 1,\ \det\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} = -1,\ \det\begin{pmatrix} 1 & 3 \\ 2 & 4 \end{pmatrix} = 1\cdot 4 - 3\cdot 2 = 4 - 6 = -2.$$

3.2.2. Three dimensions. One can do the same for three vectors in $\vec{v},\vec{w},\vec{z} \in \mathbb{R}^3$. Using (hopefully) obvious notation, and assuming that in the following divisions one

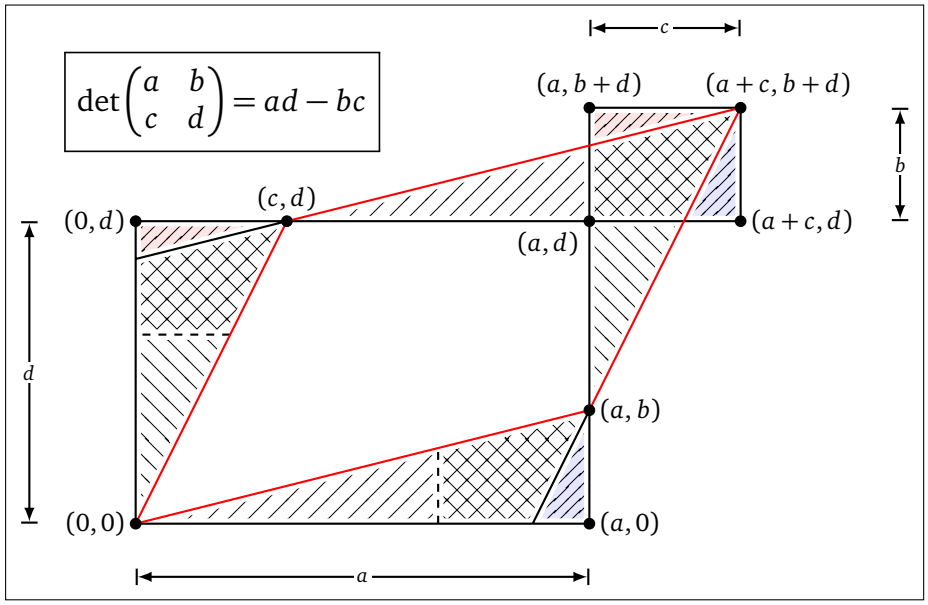


Figure 26. Illustration of the geometric meaning of the formula for 2×2-determinants.

never divides by zero, we find that

$$\begin{aligned}
\operatorname{vol}\square\begin{pmatrix} v_1 & w_1 & z_1 \\ v_2 & w_2 & z_2 \\ v_3 & w_3 & z_3 \end{pmatrix} &= \operatorname{vol}\square\begin{pmatrix} v_1 & 0 & 0 \\ v_2 & w_2 - \frac{w_1}{v_1}v_2 & z_2 - \frac{z_1}{v_1}v_2 \\ v_3 & w_3 - \frac{w_1}{v_1}v_3 & z_3 - \frac{z_1}{v_1}v_3 \end{pmatrix} \\
&= \operatorname{vol}\square\begin{pmatrix} v_1 & 0 & 0 \\ v_2 & w_2 - \frac{w_1}{v_1}v_2 & 0 \\ v_3 & w_3 - \frac{w_1}{v_1}v_3 & z_3 - \frac{z_1}{v_1}v_3 - \frac{z_2 - \frac{z_1}{v_1}v_2}{w_2 - \frac{w_1}{v_1}v_2}\left(w_3 - \frac{w_1}{v_1}v_3\right) \end{pmatrix} \\
&= \left| v_1 \left(w_2 - \frac{w_1}{v_1}v_2 \right) \left(z_3 - \frac{z_1}{v_1}v_3 - \frac{z_2 - \frac{z_1}{v_1}v_2}{w_2 - \frac{w_1}{v_1}v_2}\left(w_3 - \frac{w_1}{v_1}v_3\right) \right) \right|.
\end{aligned}$$

After multiplying out (it is advisable to multiply the middle term with the last term first in order to clear the most complicated denominators there, then carry out the multiplication by v_1 and collect all terms), we find that the volume of the parallelepiped $\square(\vec{v}, \vec{w}, \vec{z}) = \{\, \lambda\vec{v} + \mu\vec{w} + \eta\vec{z} : 0 \le \lambda, \mu, \eta \le 1 \,\}$ spanned by $\vec{v}$, $\vec{w}$ and $\vec{z}$ equals

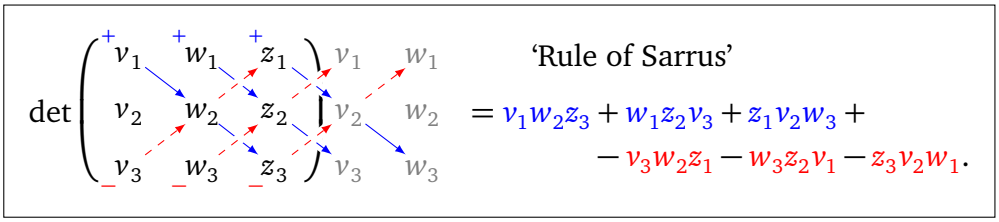


Figure 27. Mnemonic for the formula (3.4) for determinants of 3×3-matrices. This particular mnemonic is known by the name "***rule of Sarrus***". Unfortunately, such a simple formula does not exist for determinants of 4×4-matrices (it would involve only $2 \cdot 4 = 8$ terms, yet the actual formula consists of $4! = 24$ terms).

$\operatorname{vol}\measuredangle(\vec{v},\vec{w},\vec{z}) = |\det(\vec{v},\vec{w},\vec{z})|$, where

$$\det\begin{pmatrix} | & | & | \\ \vec{v} & \vec{w} & \vec{z} \\ | & | & | \end{pmatrix} := \det\begin{pmatrix} v_1 & w_1 & z_1 \\ v_2 & w_2 & z_2 \\ v_3 & w_3 & z_3 \end{pmatrix} := v_1w_2z_3 + v_3w_1z_2 + v_2w_3z_1 + - v_3w_2z_1 - v_1w_3z_2 - v_2w_1z_3. \tag{3.4}$$

Again a (tedious) case analysis can be carried out to see that this formula (for whose derivation we have assumed $v_1 \neq 0$, for instance) works in general as well.

Example. In the following example we use (3.4) to compute the determinant of a 3×3-matrix which has many zero entries. (This makes applying said formula particularly easy, as many terms just turn out to be zero.)

$$\det\begin{pmatrix} 1 & 0 & 2 \\ 0 & 2 & 0 \\ 2 & 0 & 3 \end{pmatrix} = 1 \cdot 2 \cdot 3 - 2 \cdot 2 \cdot 2 = 6 - 8 = -2.$$

Example. For the next computation we use the same procedure that we have used to derive the formula for the determinant in the first place: subtracting a suitable multiple of one column from another. (This does not change the value of the determinant and may lead to a matrix whose determinant is easier to compute.)

$$\det\begin{pmatrix} 1 & 3 & 2 \\ 2 & 2 & 0 \\ 2 & 0 & 3 \end{pmatrix} = \det\begin{pmatrix} -2 & 3 & 2 \\ 0 & 2 & 0 \\ 2 & 0 & 3 \end{pmatrix} = \det\begin{pmatrix} -2 & 3 & 0 \\ 0 & 2 & 0 \\ 2 & 0 & 5 \end{pmatrix} = \det\begin{pmatrix} -2 & 3 & 0 \\ 0 & 2 & 0 \\ 2 & 0 & 5 \end{pmatrix}$$
$$= (-2) \cdot 2 \cdot 5 = -20.$$

3.2.3. Higher dimensions. Determinants exist for arbitrary square matrices $A \in \mathbb{R}^{n\times n}$ and we would like to mention that there are general formulae for these as well. In principal, determinants of matrices $A \in \mathbb{R}^{n\times n}$ with arbitrary n do appear later in the notes, but the reader will never actually need to compute any of these.

Therefore, we shall not bother to actually define $\det A$ for $n > 3$ or provide formulas for this.

We only remark that the previously expounded procedure of subtracting suitable multiples of one column of a matrix from another column can be used to compute determinants also in higher dimensions:

$$\overset{+ \quad \cdot(-1)}{} \qquad D := \det\begin{pmatrix} 1 & 1 & 3 & 2 \\ 2 & 1 & 2 & 0 \\ 0 & 1 & 0 & 0 \\ 2 & 1 & 0 & 0 \end{pmatrix} = \det\begin{pmatrix} -1 & 1 & 3 & 2 \\ 0 & 1 & 2 & 0 \\ -2 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}.$$

Moreover, one can interchange two columns at the cost of flipping the sign:

$$D = \det\begin{pmatrix} -1 & 1 & 3 & 2 \\ 0 & 1 & 2 & 0 \\ -2 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix} = -\det\begin{pmatrix} -1 & 2 & 3 & 1 \\ 0 & 0 & 2 & 1 \\ -2 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix} = +\det\begin{pmatrix} 3 & 2 & -1 & 1 \\ 2 & 0 & 0 & 1 \\ 0 & 0 & -2 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

The result is

$$D = -\det\begin{pmatrix} 2 & 3 & -1 & 1 \\ 0 & 2 & 0 & 1 \\ 0 & 0 & -2 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

This can be justified as follows:

$$\det\begin{pmatrix} | & | & \dots \\ \vec{a} & \vec{b} & \dots \\ | & | & \dots \end{pmatrix} \overset{+}{=} \det\begin{pmatrix} | & | & \dots \\ \vec{a} & \vec{b}+\vec{a} & \dots \\ | & | & \dots \end{pmatrix} \overset{+\ \cdot(-1)}{=} \det\begin{pmatrix} | & | & \dots \\ -\vec{b} & \vec{b}+\vec{a} & \dots \\ | & | & \dots \end{pmatrix}$$

$$\overset{+}{=} \det\begin{pmatrix} | & | & \dots \\ -\vec{b} & \vec{a} & \dots \\ | & | & \dots \end{pmatrix} = -\det\begin{pmatrix} | & | & \dots \\ \vec{b} & \vec{a} & \dots \\ | & | & \dots \end{pmatrix}.$$

Moreover, determinants of "upper triangular" matrices are easy to compute:

$$\det\begin{pmatrix} a_1 & * & \dots & \dots & * \\ 0 & a_2 & * & & \vdots \\ 0 & 0 & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & * \\ 0 & \dots & \dots & 0 & a_n \end{pmatrix} = a_1 a_2 \cdots a_n.$$

(The entries marked with "$*$" can be filled in with arbitrary numbers.) In particular, $D = -(2 \cdot 2 \cdot (-2) \cdot 1) = 8$.

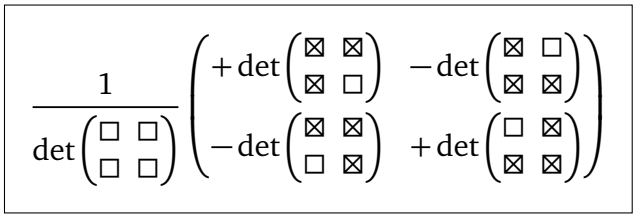


Figure 28. Schematic illustration of (3.6) from Cramer's rule (Proposition 3.2) for computing the inverse of an invertible 2×2-matrix.

3.2.4. Laplace and Cramer. One has the following useful formula.

Lemma 3.1 (Laplace's expansion)**.** *One has*

$$\det\begin{pmatrix} v_1 & w_1 & z_1\\ v_2 & w_2 & z_2\\ v_3 & w_3 & z_3\end{pmatrix} = v_1\det\begin{pmatrix}\boxtimes & \boxtimes & \boxtimes\\ \boxtimes & w_2 & z_2\\ \boxtimes & w_3 & z_3\end{pmatrix} - v_2\det\begin{pmatrix}\boxtimes & w_1 & z_1\\ \boxtimes & \boxtimes & \boxtimes\\ \boxtimes & w_3 & z_3\end{pmatrix} + v_3\det\begin{pmatrix}\boxtimes & w_1 & z_1\\ \boxtimes & w_2 & z_2\\ \boxtimes & \boxtimes & \boxtimes\end{pmatrix}$$
$$= v_1\det\begin{pmatrix} w_2 & z_2\\ w_3 & z_3\end{pmatrix} - v_2\det\begin{pmatrix} w_1 & z_1\\ w_3 & z_3\end{pmatrix} + v_3\det\begin{pmatrix} w_1 & z_1\\ w_2 & z_2\end{pmatrix},$$

where the "⊠" symbols indicate rows and columns that got removed from the original matrix. Hence, on the right hand side one has to calculate determinants of 2×2*-matrices.*

Proof. This could be proved using geometric reasoning taking into account the area considerations that have lead us to the determinant in the first place, or by a plain calculation. We shall do neither and omit the proof of this fact altogether. □

The above is called "expansion with respect to the first column". Similar expansions with respect to other columns and also with respect to rows exist (also for determinants of arbitrary $n \times n$ matrices). One must, however, pay attention to the signs in the formula. (We refrain from stating the general case.)

Determinants can be used to compute the inverse of a square matrix. The underlying principle is that determinants facilitate the construction of linear maps $\mathbb{R}^n \to \mathbb{R}$ that have a certain desired behaviour on n vectors in $\mathbb{R}^n$; this is ultimately an instance of what is called "duality" in linear algebra. However, we do not intend to develop such ideas in this course. Consequently, the subsequent formulae will seem more daunting than they would need to.

Proposition 3.2 (Cramer's rule)**.** *A matrix $A \in \mathbb{R}^{3\times 3}$ is invertible if and only if* $\det A \neq 0$*. In this case one has*

$$A^{-1} = \frac{1}{\det A}\begin{pmatrix} +\det\begin{pmatrix} a_{22} & a_{23}\\ a_{32} & a_{33}\end{pmatrix} & -\det\begin{pmatrix} a_{12} & a_{13}\\ a_{32} & a_{33}\end{pmatrix} & +\det\begin{pmatrix} a_{12} & a_{13}\\ a_{22} & a_{23}\end{pmatrix}\\ -\det\begin{pmatrix} a_{21} & a_{23}\\ a_{31} & a_{33}\end{pmatrix} & +\det\begin{pmatrix} a_{11} & a_{13}\\ a_{31} & a_{33}\end{pmatrix} & -\det\begin{pmatrix} a_{11} & a_{13}\\ a_{21} & a_{23}\end{pmatrix}\\ +\det\begin{pmatrix} a_{21} & a_{22}\\ a_{31} & a_{32}\end{pmatrix} & -\det\begin{pmatrix} a_{11} & a_{12}\\ a_{31} & a_{32}\end{pmatrix} & +\det\begin{pmatrix} a_{11} & a_{12}\\ a_{21} & a_{22}\end{pmatrix}\end{pmatrix}. \tag{3.5}$$

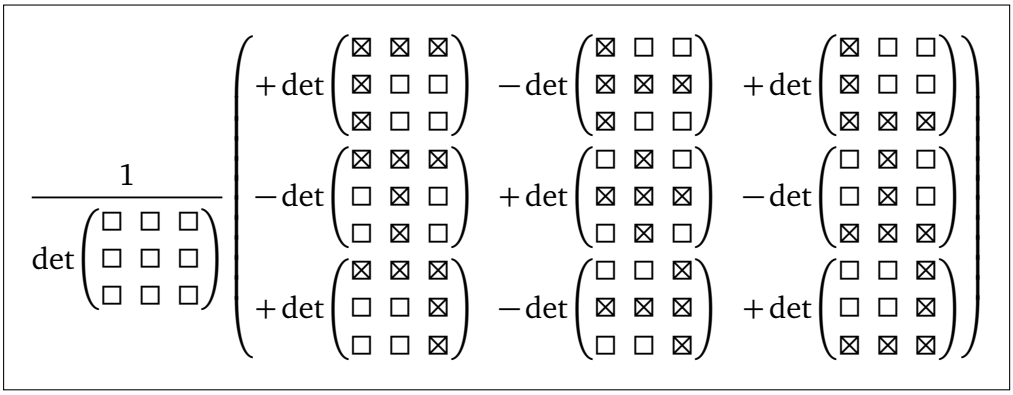


Figure 29. Schematic illustration of (3.5) from Cramer's rule (Proposition 3.2) for computing the inverse of an invertible 3×3-matrix.

A matrix $B \in \mathbb{R}^{2\times 2}$ is invertible if and only if $\det B \neq 0$. *In this case one has*

$$B^{-1} = \frac{1}{\det B}\begin{pmatrix} b_{22} & -b_{12} \\ -b_{21} & b_{11} \end{pmatrix}. \tag{3.6}$$

Proof (sketch). The entries of the matrix on the right hand side of the first formula are set up such that after multiplying by A, one can use Laplace's expansion (backwards). For instance, for computing the $(1,2)$-entry of the aforementioned matrix, one gets

$$\begin{aligned} &\frac{1}{\det A}\left(+\det\begin{pmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{pmatrix} \; -\det\begin{pmatrix} a_{12} & a_{13} \\ a_{32} & a_{33} \end{pmatrix} \; +\det\begin{pmatrix} a_{12} & a_{13} \\ a_{22} & a_{23} \end{pmatrix}\right)\begin{pmatrix} a_{12} \\ a_{22} \\ a_{32} \end{pmatrix} \\ &\quad = \frac{1}{\det A}\left(a_{12}\det\begin{pmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{pmatrix} - a_{22}\det\begin{pmatrix} a_{12} & a_{13} \\ a_{32} & a_{33} \end{pmatrix} + a_{32}\det\begin{pmatrix} a_{12} & a_{13} \\ a_{22} & a_{23} \end{pmatrix}\right) \\ &\quad = \frac{1}{\det A}\det\begin{pmatrix} a_{12} & a_{12} & a_{13} \\ a_{22} & a_{22} & a_{23} \\ a_{32} & a_{32} & a_{33} \end{pmatrix} = 0, \end{aligned}$$

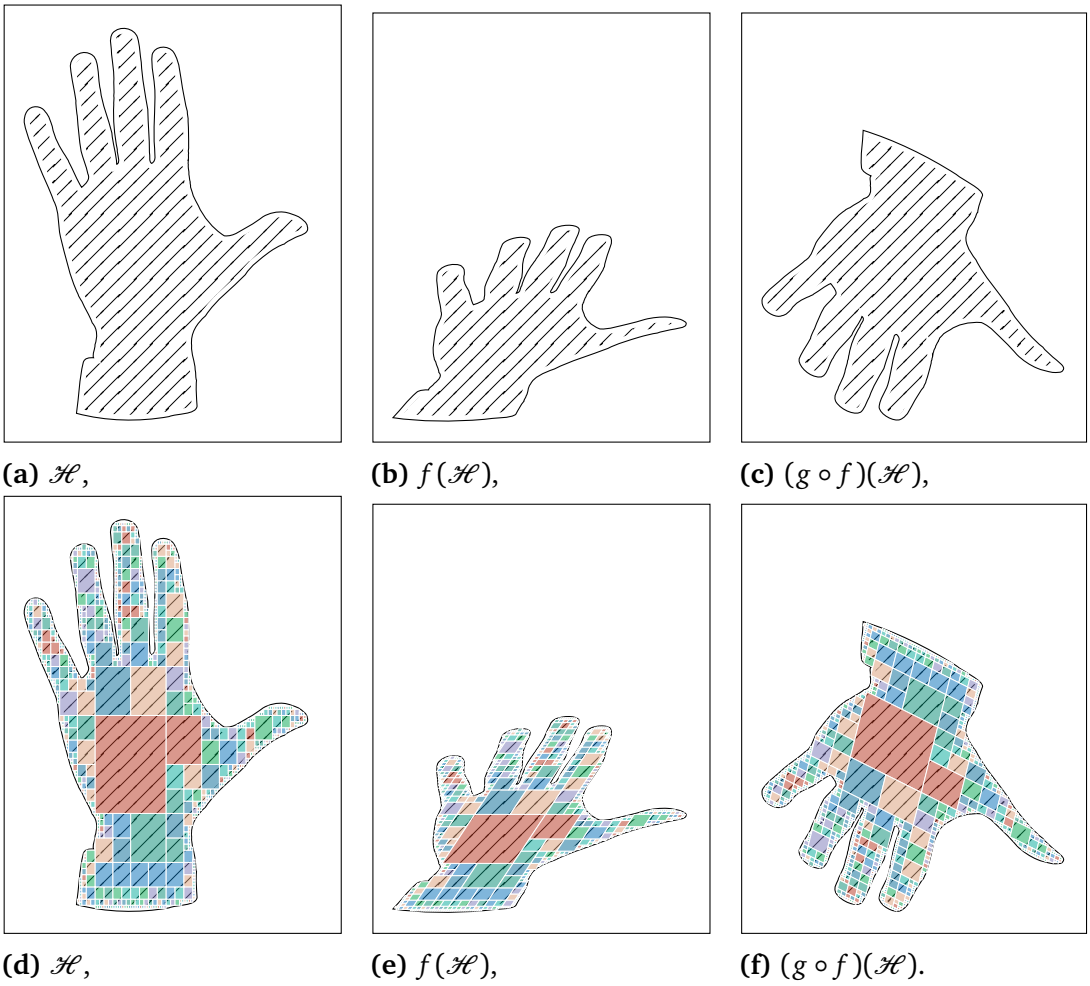
because one has two equal columns in the matrix whose determinant is computed in the last line. Similarly, when computing diagonal entries, one arrives at $\frac{1}{\det A}\det A = 1$ after applying Laplace's expansion. □

Example. To illustrate (3.6), observe that

$$\begin{pmatrix} 1 & 2 \\ 0 & 4 \end{pmatrix}^{-1} = \frac{1}{1\cdot 4 - 0\cdot 2}\begin{pmatrix} 4 & -2 \\ -0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & -1/2 \\ 0 & 1/4 \end{pmatrix}.$$

Indeed, one can check that

$$\begin{pmatrix} 1 & -1/2 \\ 0 & 1/4 \end{pmatrix}\begin{pmatrix} 1 & 2 \\ 0 & 4 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 2 \\ 0 & 4 \end{pmatrix}\begin{pmatrix} 1 & -1/2 \\ 0 & 1/4 \end{pmatrix}.$$



(a) $\mathscr{H}$, **(b)** $f(\mathscr{H})$, **(c)** $(g\circ f)(\mathscr{H})$,

(d) $\mathscr{H}$, **(e)** $f(\mathscr{H})$, **(f)** $(g\circ f)(\mathscr{H})$.

Figure 30. Illustration of a two-dimensional object $\mathscr{H}$ (shadow of a hand) being subjected to the action of linear maps. Here $f, g\colon \mathbb{R}^2 \to \mathbb{R}^2$ are linear with

$$f(\vec{v}) = \begin{pmatrix} 1 & 1/3 \\ 0 & 1/2 \end{pmatrix}\vec{v}, \quad \text{and} \quad g(\vec{w}) = \begin{pmatrix} 1 & -1 \\ -1/2 & -1 \end{pmatrix}\vec{w}.$$

The second row includes a partition of (part of) $\mathscr{H}$ into rectangles. The rectangles get mapped onto parallelograms by the linear maps. From this one can glean how the area of $\mathscr{H}$ gets distorted.

3.2.5. Determinant of a linear map. The (absolute value of the) determinant of a matrix $A \in \mathbb{R}^{3\times 3}$ can also be viewed as a measure for how the linear map $f\colon \mathbb{R}^3 \to \mathbb{R}^3$, $\vec{v} \mapsto A\vec{v}$, represented by A stretches volumes. Indeed, f maps the standard unit vectors to the columns of A respectively; in particular, the standard unit cube (volume 1) spanned by the standard unit vectors gets mapped to the parallelepiped spanned by

the columns of A (volume $|\det A|$). The fact that, on applying f, cubes (and rectangles) have their volumes distorted by a factor $|\det A|$ (where the sign of the determinant determines if f respects or reverses orientation) carries over to more general shapes whose volumes we (naively) imagine to be measured by exhausting them using cubes (see Figure 30 for a two-dimensional illustration of this). We define

$$\det f := \det A.$$

If we have a second matrix $B \in \mathbb{R}^{3\times 3}$, then the linear map $g\colon \mathbb{R}^3 \to \mathbb{R}^3$, $\vec{v} \mapsto B\vec{v}$, represented by B distorts volumes by $|\det B|$. Since their composition $f \circ g$ distorts volumes by a factor of $|\det AB|$, the next result should not come as a surprise:

Proposition 3.3 (Determinant multiplication formula)**.** *For any two matrices $A, B \in \mathbb{R}^{3\times 3}$, we have*

$$\det(AB) = (\det A)(\det B).$$

(This formula also holds for 2×2-matrices and, in fact, given the correct general definition of determinants, for arbitrary square matrices.)

We record another useful formula, which may seem quite magical. For this we need to introduce some notation. For any matrix

$$A = \begin{pmatrix} a_{11} & \dots & a_{1n} \\ a_{21} & \dots & a_{2n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \dots & a_{mn} \end{pmatrix} \in \mathbb{R}^{m\times n},$$

we define its ***transpose*** $A^\intercal$ to be the matrix

$$A^\intercal := \begin{pmatrix} a_{11} & a_{21} & \dots & a_{m1} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1n} & a_{2n} & \dots & a_{mn} \end{pmatrix} \in \mathbb{R}^{n\times m}.$$

(Flipping the matrix across the diagonal.)

Example (Transposing matrices)**.**

$$\begin{pmatrix} 1 \\ 2 \end{pmatrix} \underset{B^\intercal \leftarrow\!\shortmid B}{\overset{A \mapsto A^\intercal}{\longleftrightarrow}} \begin{pmatrix} 1 & 2 \end{pmatrix}, \quad \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \underset{B^\intercal \leftarrow\!\shortmid B}{\overset{A \mapsto A^\intercal}{\longleftrightarrow}} \begin{pmatrix} 1 & 3 \\ 2 & 4 \end{pmatrix}, \quad \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} \underset{B^\intercal \leftarrow\!\shortmid B}{\overset{A \mapsto A^\intercal}{\longleftrightarrow}} \begin{pmatrix} 1 & 2 & 3 \end{pmatrix},$$

$$\begin{pmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{pmatrix} \underset{B^\intercal \leftarrow\!\shortmid B}{\overset{A \mapsto A^\intercal}{\longleftrightarrow}} \begin{pmatrix} 1 & 3 & 5 \\ 2 & 4 & 6 \end{pmatrix}, \quad \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix} \underset{B^\intercal \leftarrow\!\shortmid B}{\overset{A \mapsto A^\intercal}{\longleftrightarrow}} \begin{pmatrix} 1 & 4 & 7 \\ 2 & 5 & 8 \\ 3 & 6 & 9 \end{pmatrix}.$$

Proposition 3.4. *The linear map induced by $A \in \mathbb{R}^{m\times n}$ distorts m-dimensional volumes by a factor of*

$$\sqrt{\det(A^\intercal A)}.$$

Proof. For a *square* matrix $A \in \mathbb{R}^{n\times n}$ we have $\det(A^\intercal) = \det A$ (as one can check easily for $n \leq 3$ by looking at the explicit formulas derived above for determinants), so that, by Proposition 3.3,

$$\sqrt{\det(A^\intercal A)} = \sqrt{\det(A^\intercal)\det A} = \sqrt{(\det A)^2} = |\det A|.$$

For the non-square case one can, for instance, appeal to the so-called "QR decomposition" of A, but we refrain from doing so and skip this part of the proof completely. (Note that this is utter cheating as the proposition is mainly of interest in cases where A is non-square so that "$\det A$" cannot be used to determine volume distortion.) □

The expression $\det(A^\intercal A)$ is called ***Gram determinant***.[3]

Example (Computing a Gram determinant)**.** Let $f\colon \mathbb{R}^2 \to \mathbb{R}^3$ be the linear map induced by

$$A = \begin{pmatrix} 3 & 0 \\ 0 & 0 \\ 0 & 5 \end{pmatrix},$$

i.e., $f(v_1, v_2) = (3v_1, 0, 5v_2)$. Then

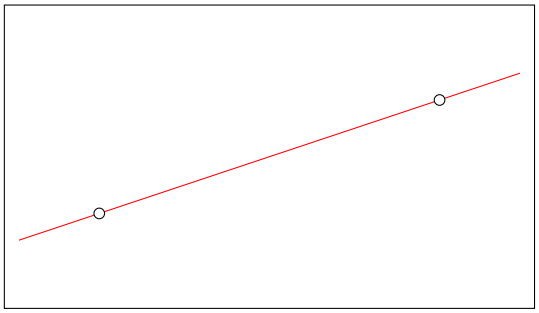
$$A^\intercal = \begin{pmatrix} 3 & 0 & 0 \\ 0 & 0 & 5 \end{pmatrix} \quad \text{and} \quad A^\intercal A = \begin{pmatrix} 3 & 0 & 0 \\ 0 & 0 & 5 \end{pmatrix}\begin{pmatrix} 3 & 0 \\ 0 & 0 \\ 0 & 5 \end{pmatrix} = \begin{pmatrix} 9 & 0 \\ 0 & 25 \end{pmatrix}.$$

Hence, $\sqrt{\det(A^\intercal A)} = \sqrt{9\cdot 25} = 3\cdot 5 = 15$.

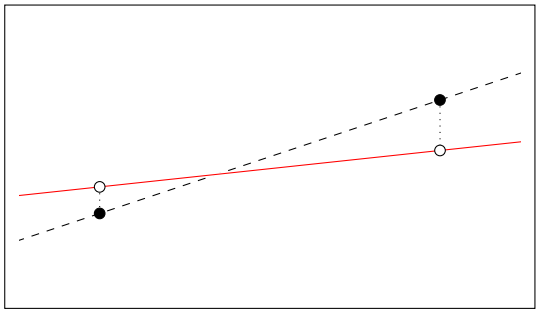
3.2.6. Application: fitting curves to data. Here we mention a cute application of linear algebra (also without providing proofs) which may be useful to the reader. Imagine having performed some physical experiment and having recorded the measured data. We model this by imagining there to be some function $f\colon \mathbb{R} \to \mathbb{R}$ describing the state of some apparatus and being given samples $f(t_i)$ at discrete times $t_1 < t_2 < \ldots < t_m$.

Suppose that we know from physics that the graph of f is a line, i.e., $f(t) = \alpha t + \beta$ for some $\alpha, \beta \in \mathbb{R}$ and all times t. Note that two points determine a line (Figure 31 (a)). We now imagine our data being subject to error (after all, we cannot expect our measurements to be absolutely precise). Then we may find completely wrong values for α and β (Figure 31 (b)). However, if more measurements are performed, we may hope to be able to improve our guess for α and β (Figure 31 (c)). Unfortunately, there may be no line at all which connects all measured points. However, linear algebra can be used to find a satisfactory substitute. We sketch a more general procedure first (see Figure 32 for an example of this in action). Suppose one is given n functions

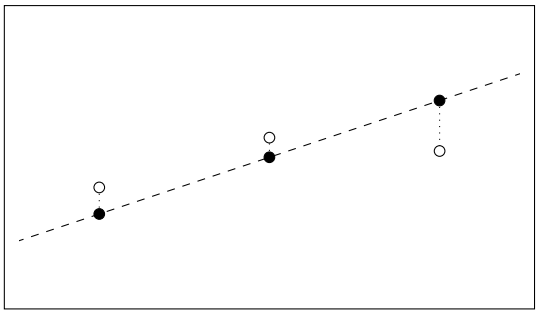
[3] Actually, we are making a conceptual error here in that we should have written $\det(A^*A)$ with the Hermitian adjoint A^* of A. However, in the present context, the matrices A^* and $A^\intercal$ coincide.



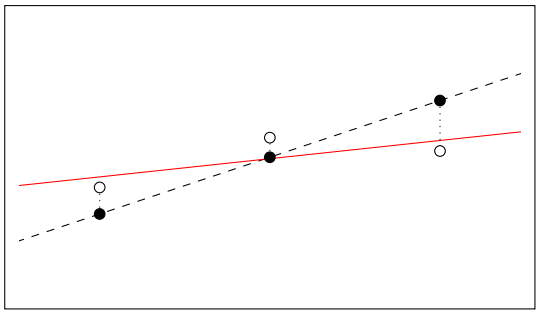
(a) Two points determine a line.



(b) A line through two measured points.



(c) Three measurements. It may happen that no line goes through all of them.



(d) Fitting a line to three data points.

Figure 31. Fitting functions to data.

$f_1, \dots, f_n \colon \mathbb{R} \to \mathbb{R}$ and $m \geq n$ measurements $(t_1, y_1), (t_2, y_2), \dots, (t_m, y_m) \in \mathbb{R}^2$, then the solution $\vec{x} \in \mathbb{R}^n$ of the system

$$A^\intercal A\vec{x} \stackrel{!}{=} A^\intercal \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix} \quad \text{with} \quad A := \begin{pmatrix} f_1(t_1) & \dots & f_n(t_1) \\ f_1(t_2) & \dots & f_n(t_2) \\ \vdots & & \vdots \\ f_1(t_m) & \dots & f_n(t_m) \end{pmatrix} \in \mathbb{R}^{m \times n}$$

minimises the sum of the squared errors

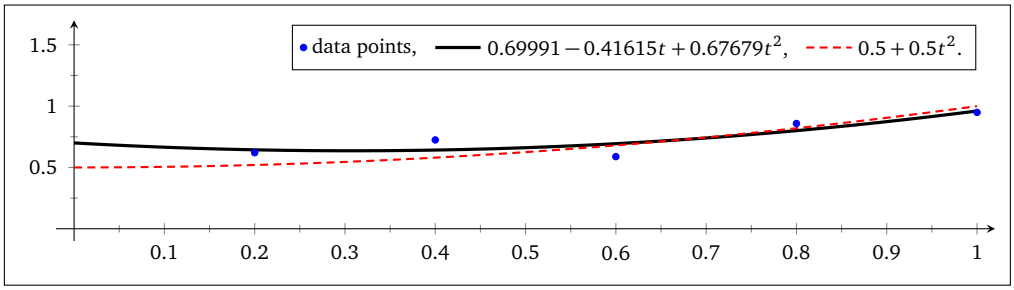
$$\sum_{i=1}^{m} (y_i - (x_1 f_1(t_i) + \dots + x_n f_n(t_i)))^2.$$

Example. We wish to find the best approximation (in the sense of minimising the sum of squared errors) by a line to the three points

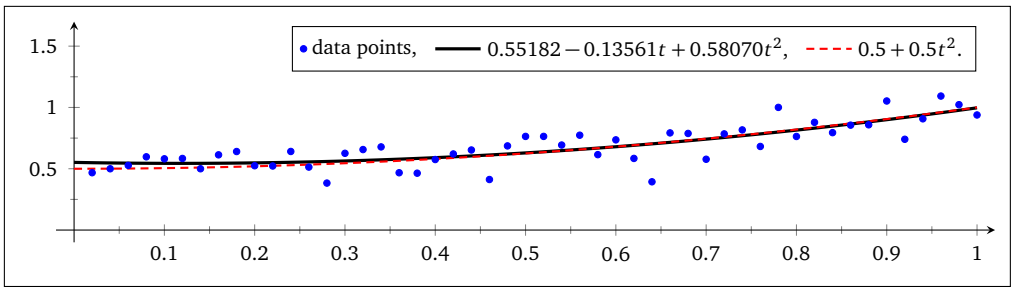
$$(0, 0.234), \quad (1.5, 0.673), \quad (3, 0.555).$$

Thus, we pick $f_1(t) = 1$, $f_2(t) = t$ in the above and solve

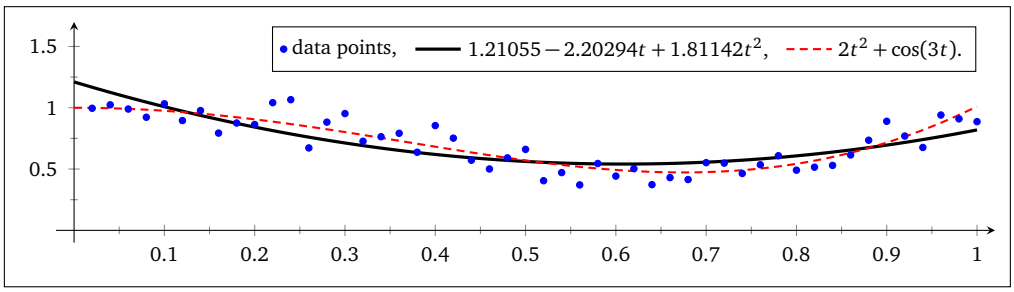
$$A^\intercal A \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \stackrel{!}{=} A^\intercal \begin{pmatrix} 0.234 \\ 0.673 \\ 0.555 \end{pmatrix} \quad \text{with} \quad A := \begin{pmatrix} 1 & 0 \\ 1 & 1.5 \\ 1 & 3 \end{pmatrix}$$



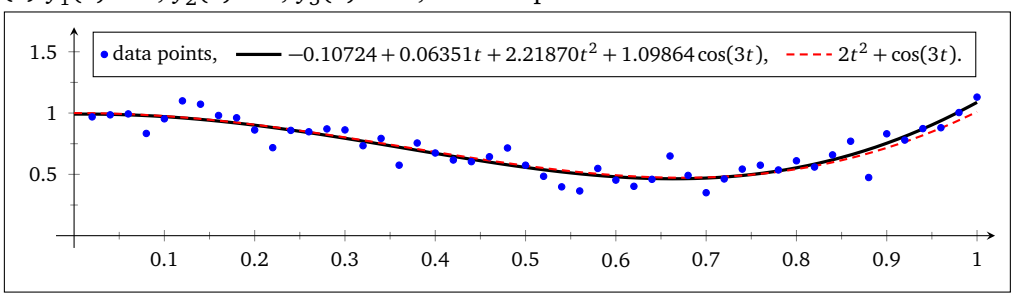
(a) $f_1(t) = 1$, $f_2(t) = t$, $f_3(t) = t^2$, 5 data points.



(b) $f_1(t) = 1$, $f_2(t) = t$, $f_3(t) = t^2$, 50 data points.



(c) $f_1(t) = 1$, $f_2(t) = t$, $f_3(t) = t^2$, 50 data points.



(d) $f_1(t) = 1$, $f_2(t) = t$, $f_3(t) = t^2$, $f_4(t) = \cos(3t)$, 50 data points.

Figure 32. Fitting functions to data. The points are obtained by adding pseudo-random (normally distributed) noise to the dashed graph and the black graph shows the best approximation reconstructed from the data.

for $\vec{x} \in \mathbb{R}^2$. Upon carrying this out—which we do not do here—this yields $\vec{x} = (0.3268\overline{3}, 0.107)$ and we are lead to

$$t \mapsto 0.3268\overline{3} + 0.107t.$$

The graph of this function can be seen in Figure 31 (d).

3.2.7. Orientation. Lastly, we close this section with a discussion of the meaning of the *sign of the determinant*. So far, in our geometrical interpretation of the determinant as a measure for length/area/volume distortion, we have always used the absolute value of the determinant, thus forgetting the sign information. However, as it turns out, said sign also admits a geometrical interpretation: using one of your hands[4] point your thumb, index and middle finger towards the standard unit vectors $\vec{e}_1, \vec{e}_2, \vec{e}_3$ respectively.[5] Suppose now that you are given an invertible matrix $A \in \mathbb{R}^{3\times 3}$. Then you will be able to rotate your chosen hand in space and point your thumb, index and middle finger in the directions of the columns of A (in their given order) if and only if $\det A > 0$. We record a vague formulation of this in the next proposition and hope that our use of the word "orientation" is sufficiently self-explanatory.

Proposition 3.5. *The sign of the determinant* $\det A$ *of an invertible matrix* $A \in \mathbb{R}^{3\times 3}$ *determines the orientation of the columns of A in the way just described; the linear map* $\mathbb{R}^3 \to \mathbb{R}^3$ *induced by A preserves orientation of and only if* $\det A > 0$ *and reverses orientation if and only if* $\det A < 0$.

Proof (sketch). If done correctly,[6] the operation of rotating and moving your fingers to match the directions of the columns of A traces over a bunch of invertible matrices in $\mathbb{R}^{3\times 3}$ in a "continuous" fashion over time $t \in [0,1]$ (say). Taking determinants, we get a continuous map $[0,1] \to \mathbb{R} \setminus \{0\}$ starting with $\det(\mathbf{1}_3) = 1$ and ending at $\det A$. Now if $\det A < 0$ this map would have to take the value 0 at some intermediate time $t \in (0,1)$. However, this is not possible. Therefore, we must have $\det A > 0$. This proves that translating the standard unit vectors to match the orientation of the columns of A via moving your hand (in the way described above) is possible only if $\det A > 0$.

Showing that $\det A > 0$ is actually sufficient for making such moves possible is slightly more involved but can be carried out by extracting continuous moves from the subtraction procedure we have used in the beginning to derive formulas for determinants of $\mathbb{R}^{2\times 2}$ and $\mathbb{R}^{3\times 3}$ matrices. We omit the details. □

[4]The usual choice is using the right hand, but this really is a matter of taste at this point.

[5]Whether or not this is possible without incurring any injuries to your fingers depends on how you draw your coordinate system and which choice of hand (left or right) you are using. Either way, suppose that you made your choices in such a way that it fits.

[6]Here "correctly" means not artificially extending your thumb, index and middle fingers in such a way that they all lie in a plane.

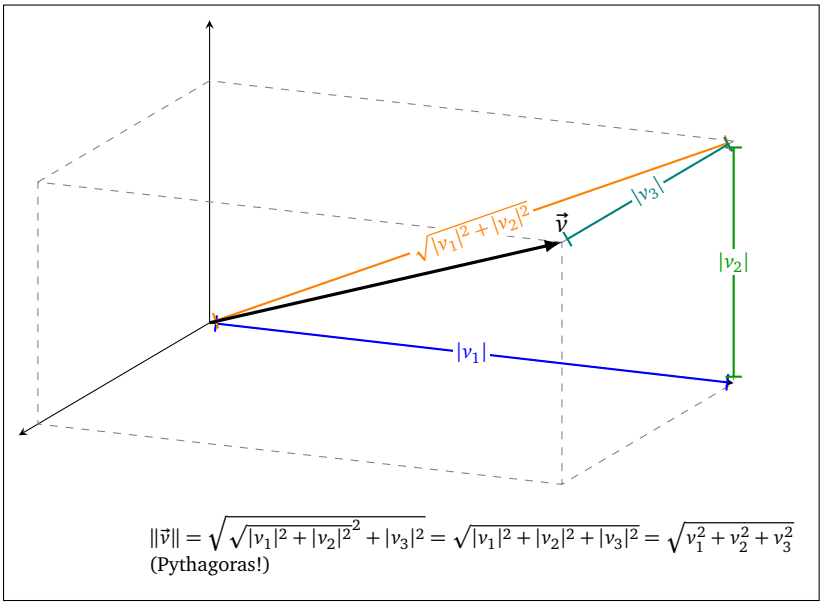


Figure 33. Illustration of the definition of the length of a vector.

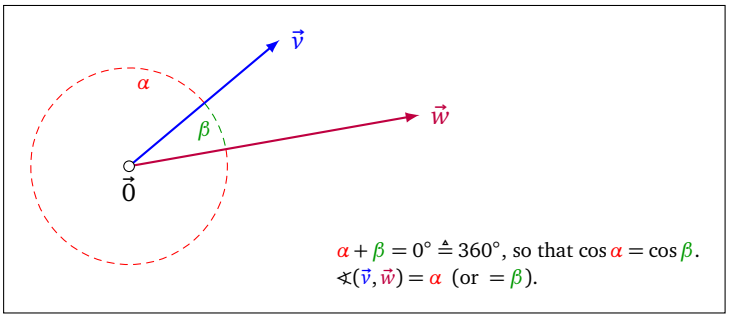


Figure 34. The angle between two vectors. Observe that $\cos\sphericalangle(\vec{v},\vec{w})$ is well-defined regardless of which of the two possible angles α or β one considers to be $\sphericalangle(\vec{v},\vec{w})$. Moreover, a similar argument also shows that $\sin\alpha = -\sin\beta$ here, so that $|\sin\sphericalangle(\vec{v},\vec{w})|$ is well-defined too.

3.3. Dot product

The ***length*** (or ***norm***) of a vector $\vec{v} = (v_1, \ldots, v_n) \in \mathbb{R}^n$ is given by

$$\|\vec{v}\| = \sqrt{v_1^2 + \ldots + v_n^2}.$$

This can be justified by appealing to the Pythagorean theorem or (more abstractly) by viewing $\vec{v}$ as a matrix in $\mathbb{R}^{n\times 1}$ and computing the 1-dimensional volume (length) distortion furnished by the induced linear map $\mathbb{R}^1 \to \mathbb{R}^n$, $\lambda \mapsto \lambda\vec{v}$, by means of Proposition 3.4.

The ***scalar product*** or ***dot product*** of two vectors $\vec{v}, \vec{w} \in \mathbb{R}^n$ is defined by

$$\vec{v} \cdot \vec{w} := \|\vec{v}\|\|\vec{w}\| \cos \sphericalangle(\vec{v}, \vec{w}),$$

where $\sphericalangle(\vec{v}, \vec{w})$ denotes the (oriented[7]) angle between $\vec{v}$ and $\vec{w}$. (Special case: if either of the vectors is zero, then the $\sphericalangle(\vec{v}, \vec{w})$ is not really defined, but in that case $\|\vec{v}\|\|\vec{w}\| = 0$ anyway, so we just agree that $\vec{v} \cdot \vec{w} = 0$ in that case.)

Lemma 3.6. *The dot product is linear in both arguments, i.e., for any vectors $\vec{v}, \vec{w}, \vec{z} \in \mathbb{R}^n$ and any scalar $\lambda \in \mathbb{R}$, we have*

- $(\lambda\vec{v}) \cdot \vec{w} = \lambda(\vec{v} \cdot \vec{w}) = \vec{v} \cdot (\lambda\vec{w})$,
- $\vec{v} \cdot (\vec{w} + \vec{z}) = \vec{v} \cdot \vec{w} + \vec{v} \cdot \vec{z}$ *and* $(\vec{v} + \vec{w}) \cdot \vec{z} = \vec{v} \cdot \vec{z} + \vec{w} \cdot \vec{z}$.

Proof. The formula $(\lambda\vec{v}) \cdot \vec{w} = \lambda(\vec{v} \cdot \vec{w})$ is obvious from the definition: replacing $\vec{v}$ by $\lambda\vec{v}$, because $\|\vec{v}\|$ is changed by a factor of $|\lambda|$ and the cosine of the angle compensates for the loss of any sign (if $\lambda < 0$, then $\sphericalangle(\vec{v}, \vec{w})$ differs from $\sphericalangle(\lambda\vec{v}, \vec{w})$ by $\pm\pi = \pm 180°$). The proof of $\lambda(\vec{v} \cdot \vec{w}) = \vec{v} \cdot (\lambda\vec{w})$ is similar.

Of the last two remaining formulae we only show $\vec{v} \cdot (\vec{w} + \vec{z}) = \vec{v} \cdot \vec{w} + \vec{v} \cdot \vec{z}$, the proof of the second being similar. Moreover, by rescaling $\vec{v}$, we may assume that $\|\vec{v}\| = 1$. Since the definition of $\vec{v} \cdot \vec{w}$ is invariant under rotations, we may assume additionally that $\vec{v} = (1, 0, \ldots)$. It is clear from two-dimensional geometry that $\|\vec{w}\| \cos \sphericalangle(\vec{v}, \vec{w})$ is the (oriented) length of the orthogonal projection of $\vec{w}$ onto the line $\mathbb{R}\vec{v}$. Therefore,

$$\|\vec{w}\| \cos \sphericalangle(\vec{v}, \vec{w}) = w_1, \quad \|\vec{z}\| \cos \sphericalangle(\vec{v}, \vec{z}) = z_1 \quad \text{and} \quad \|\vec{w} + \vec{z}\| \cos \sphericalangle(\vec{v}, \vec{w} + \vec{z}) = w_1 + z_1.$$

Consequently, $\vec{v} \cdot (\vec{w} + \vec{z}) = w_1 + z_1 = \vec{v} \cdot \vec{w} + \vec{v} \cdot \vec{z}$, and the lemma is proved. □

The key observation is that Lemma 3.6 allows us to compute the dot product of two vectors quite easily:

Proposition 3.7. *For any vectors $\vec{v}, \vec{w} \in \mathbb{R}^n$, we have $\vec{v} \cdot \vec{w} = v_1 w_1 + \ldots + v_n w_n$.*

Proof. Write $\vec{v} = v_1\vec{e}_1 + \ldots + v_n\vec{e}_n = \sum_i v_i\vec{e}_i$ and, similarly, $\vec{w} = \sum_j w_j\vec{e}_j$. (We use different indices here, because both sums will show up next to each other in a moment.) Then, by linearity

$$\vec{v} \cdot \vec{w} = \Big(\sum\nolimits_i v_i\vec{e}_i\Big) \cdot \Big(\sum\nolimits_j w_j\vec{e}_j\Big) = \sum\nolimits_i v_i\Big(\vec{e}_i \cdot \Big(\sum\nolimits_j w_j\vec{e}_j\Big)\Big) = \sum\nolimits_i v_i \sum\nolimits_j w_j(\vec{e}_i \cdot \vec{e}_j).$$

However, $\vec{e}_i \cdot \vec{e}_j = 0$ whenever $i \neq j$ and $= 1$ for $i = j$. Hence, in the above double sum, only the 'diagonal' terms with $i = j$ survive. This shows that $\vec{v} \cdot \vec{w} = \sum_i v_i w_i$, as claimed. □

Examples.

(1) $\begin{pmatrix} 2 \\ 0 \end{pmatrix} \cdot \begin{pmatrix} 3 \\ 4 \end{pmatrix} = 2 \cdot 3 + 0 \cdot 4 = 6.$

[7] Actually, orientation does not matter here; see Figure 34.

(2) $\begin{pmatrix}1\\0\end{pmatrix}\cdot\begin{pmatrix}0\\1\end{pmatrix} = 1\cdot 0 + 0\cdot 1 = 0$. ($\vec{e}_1\cdot\vec{e}_2 = 0$, or, more generally, $\vec{e}_i\cdot\vec{e}_j = 0$ for $i \neq j$.)

We record some useful consequences, which are immediate from the definition of the dot product, but turn out to be computationally feasible only because Proposition 3.7 furnishes an easy way of computing the dot product (the original definition via the angle is rather opaque).

Corollary 3.8. *For any vectors $\vec{v}, \vec{w} \in \mathbb{R}^n$, we have*

$$\cos \measuredangle(\vec{v}, \vec{w}) = \frac{\vec{v}\cdot\vec{w}}{\|\vec{v}\|\|\vec{w}\|}.$$

In particular, the vectors $\vec{v}$ and $\vec{w}$ are perpendicular (orthogonal) to each other if and only if $\vec{v}\cdot\vec{w} = 0$. (Notation: $\vec{v} \perp \vec{w}$; special case: the zero vector $\vec{0}$ is perpendicular to every vector.)

Examples. Consider the vectors $\vec{v} = \begin{pmatrix}2\\0\end{pmatrix}$ and $\vec{w} = \begin{pmatrix}3\\4\end{pmatrix}$. Then

$$\cos \measuredangle(\vec{v}, \vec{w}) = \frac{\vec{v}\cdot\vec{w}}{\|\vec{v}\|\|\vec{w}\|} = \frac{6}{2\sqrt{3^2+4^2}} = \frac{3}{5}.$$

We saw in Figure 34 that there is some ambiguity in defining the angle $\measuredangle(\vec{v}, \vec{w})$ between $\vec{v}$ and $\vec{w}$. Of the (up to) two choices for which $\measuredangle(\vec{v}, \vec{w})$ is in $[0, 2\pi]$ (that is, in $[0^\circ, 360^\circ]$), there is one choice for which $\measuredangle(\vec{v}, \vec{w})$ lies in $[0, \pi]$. To determine this particular choice of $\measuredangle(\vec{v}, \vec{w})$, we can use the arcus cosine function arccos, which is the inverse function of $\cos|_{[0,\pi]}\colon [0, \pi] \to [-1, 1]$. Then

$$\arccos(\cos \measuredangle(\vec{v}, \vec{w})) = \arccos\frac{3}{5} = 0.9272952\ldots \approx 53.1301^\circ.$$

(Here the last two values were obtained using a calculator.)

3.4. Cross product in three dimensions

Let $\vec{v}$ and $\vec{w}$ be two vectors in $\mathbb{R}^3$. Then we define the ***cross product*** (or ***vector product***) $\vec{v}\times\vec{w}$ of $\vec{v}$ and $\vec{w}$ via the formula

$$\begin{aligned}\vec{v}\times\vec{w} &:= \det\begin{pmatrix}\vec{e}_1 & | & |\\ \vec{e}_2 & \vec{v} & \vec{w}\\ \vec{e}_3 & | & |\end{pmatrix} = \det\begin{pmatrix}\vec{e}_1 & v_1 & w_1\\ \vec{e}_2 & v_2 & w_2\\ \vec{e}_3 & v_3 & w_3\end{pmatrix}\\ &:= \begin{pmatrix}1\\0\\0\end{pmatrix}\det\begin{pmatrix}\boxtimes & \boxtimes & \boxtimes\\ \boxtimes & v_2 & w_2\\ \boxtimes & v_3 & w_3\end{pmatrix} - \begin{pmatrix}0\\1\\0\end{pmatrix}\det\begin{pmatrix}\boxtimes & v_1 & w_1\\ \boxtimes & \boxtimes & \boxtimes\\ \boxtimes & v_3 & w_3\end{pmatrix} + \begin{pmatrix}0\\0\\1\end{pmatrix}\det\begin{pmatrix}\boxtimes & v_1 & w_1\\ \boxtimes & v_2 & w_2\\ \boxtimes & \boxtimes & \boxtimes\end{pmatrix},\end{aligned}$$

where the determinants in the first line ought to be computed by formally(!) applying Laplace's expansion with respect to the first (vector-valued) column (see Lemma 3.1). More explicitly,

$$\begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix} \times \begin{pmatrix} w_1 \\ w_2 \\ w_3 \end{pmatrix} = \begin{pmatrix} v_2 w_3 - v_3 w_2 \\ v_3 w_1 - v_1 w_3 \\ v_1 w_2 - v_2 w_1 \end{pmatrix}.$$

The reason for the above definition is the following fundamental formula. Let $\vec{n}$ be any other vector in $\mathbb{R}^3$. Then

$$\begin{aligned} \vec{n} \cdot (\vec{v} \times \vec{w}) &= n_1 \det \begin{pmatrix} v_2 & w_2 \\ v_3 & w_3 \end{pmatrix} - n_2 \det \begin{pmatrix} v_1 & w_1 \\ v_3 & w_3 \end{pmatrix} + n_3 \det \begin{pmatrix} v_1 & w_1 \\ v_2 & w_2 \end{pmatrix} \\ &= \det \begin{pmatrix} n_1 & v_1 & w_1 \\ n_2 & v_2 & w_2 \\ n_3 & v_3 & w_3 \end{pmatrix} = \det \begin{pmatrix} | & | & | \\ \vec{n} & \vec{v} & \vec{w} \\ | & | & | \end{pmatrix}. \end{aligned} \tag{3.7}$$

Note that (3.7) is zero for any $\vec{n} \in \operatorname{span}\{\vec{v}, \vec{w}\}$. Therefore, $(\vec{v} \times \vec{w}) \perp \operatorname{span}\{\vec{v}, \vec{w}\}$, i.e., $\vec{v} \times \vec{w}$ is perpendicular to the plane spanned by $\vec{v}$ and $\vec{w}$.

Examples. A reader who wants to compute some examples to get used to the definition of the cross products may try to verify the following:

$$\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \times \begin{pmatrix} 2 \\ 1 \\ 0 \end{pmatrix} = \begin{pmatrix} -1 \\ 2 \\ -1 \end{pmatrix}, \quad \begin{pmatrix} 2 \\ 1 \\ 0 \end{pmatrix} \times \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \\ -2 \\ 1 \end{pmatrix}, \quad \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} \times \begin{pmatrix} 2 \\ 3 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}.$$

The next lemma shows how the cross product interacts with linear maps.

Lemma 3.9. *For any matrix $A \in \mathbb{R}^{3\times 3}$ and any vectors $\vec{v}, \vec{w} \in \mathbb{R}^3$ one has*

$$A^{\intercal}((A\vec{v}) \times (A\vec{w})) = (\det A)(\vec{v} \times \vec{w}).$$

Proof. It suffices to show that the claimed equality holds after looking at the dot product with every vector $\vec{x} \in \mathbb{R}^3$ (take $\vec{x}$ to be the standard unit vectors). Hence, we want to show that

$$\vec{x} \cdot A^{\intercal}((A\vec{v}) \times (A\vec{w})) = \vec{x} \cdot (\det A)(\vec{v} \times \vec{w}).$$

Upon applying (3.7) as well as the well-known identity $(B\vec{y}) \cdot \vec{z} = \vec{y} \cdot (B^{\intercal}\vec{z})$, we find that the left hand side of the above equals

$$\begin{aligned} &(A\vec{x}) \cdot ((A\vec{v}) \times (A\vec{w})) \\ &\quad = \det \begin{pmatrix} | & | & | \\ A\vec{x} & A\vec{v} & A\vec{w} \\ | & | & | \end{pmatrix} = \det \left(A \begin{pmatrix} | & | & | \\ \vec{x} & \vec{v} & \vec{w} \\ | & | & | \end{pmatrix} \right) = (\det A) \det \begin{pmatrix} | & | & | \\ \vec{x} & \vec{v} & \vec{w} \\ | & | & | \end{pmatrix} \\ &\quad = (\det A)(\vec{x} \cdot (\vec{v} \times \vec{w})). \end{aligned}$$

This proves the lemma. □

Figure 35. Picture of the Swiss 200 franc bank note of 2018 showing, amongst other things, a right-handed coordinate system.

Let $\mathrm{SO}(3) = \{A \in \mathbb{R}^{3\times 3} : A^\intercal A = \mathbf{1}_3, \det A = 1\}$. The matrices $A \in \mathrm{SO}(3)$ are precisely those whose columns are normalised to have unit length, are orthogonal to each other and ordered in the correct orientation (compare Proposition 3.5). These matrices are precisely those whose induced linear maps are *rotations* in space.

Corollary 3.10. *For any matrix $A \in \mathrm{SO}(3)$ and any vectors $\vec{v}, \vec{w} \in \mathbb{R}^3$ one has*

$$(A\vec{v}) \times (A\vec{w}) = A(\vec{v} \times \vec{w}).$$

Because of

$$\begin{pmatrix}1\\0\\0\end{pmatrix} \times \begin{pmatrix}0\\1\\0\end{pmatrix} = \begin{pmatrix}0\\0\\1\end{pmatrix},$$

Lemma 3.9 justifies the "*right hand rule*" for determining the spacial orientation of $\vec{v} \times \vec{w}$: using the thumb, index finger and middle finger of your right hand to form a coordinate cross, suppose that your coordinates are chosen such that your thumb points in the $\vec{e}_1$ direction, your index finger points in the $\vec{e}_2$ direction and your middle finger points in the $\vec{e}_3$ direction. Now rotate your right hand such that your thumb and index finger lie in the plane spanned by $\vec{v}$ and $\vec{w}$. Then your middle finger points in the direction of $\vec{v} \times \vec{w}$.

We have

$$\begin{pmatrix}v_1\\v_2\\0\end{pmatrix} \times \begin{pmatrix}w_1\\w_2\\0\end{pmatrix} = \begin{pmatrix}0\\0\\1\end{pmatrix} \det \begin{pmatrix}v_1 & w_1\\v_2 & w_2\end{pmatrix}.$$

and this equals the area of the parallelogram spanned by the vectors $(v_1, v_2, 0)$ and $(w_1, w_2, 0)$. Since for any two vectors $\vec{v}, \vec{w} \in \mathbb{R}^3$ there is some rotation matrix $A \in \mathrm{SO}(3)$ such that

$$A\vec{v}, A\vec{w} \in \mathbb{R}^2 \times \{0\},$$

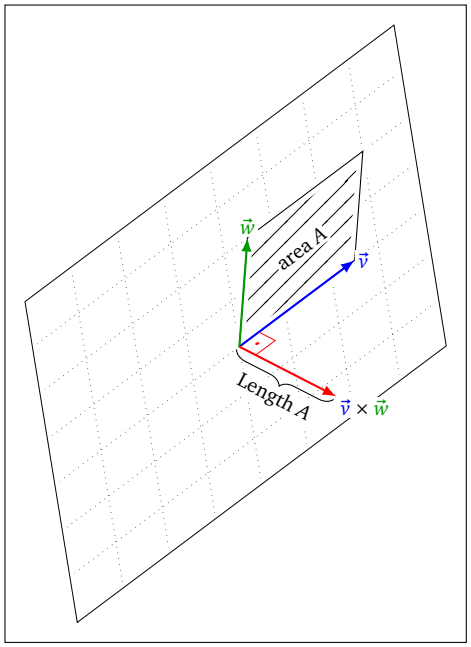


Figure 36. Illustration of the cross product $\vec{v} \times \vec{w}$ of two vectors $\vec{v}$ and $\vec{w}$. It stands perpendicularly on the plane spanned by $\vec{v}$ and $\vec{w}$ (see (3.7)) and its length is the area of the parallelogram spanned by $\vec{v}$ and $\vec{w}$ (Proposition 3.11).

the above computation shows the following:

Proposition 3.11. $\|\vec{v} \times \vec{w}\| =$ *area of the parallelogram spanned by* $\vec{v}$ *and* $\vec{w}$.

Proof (alternative). For readers who are unhappy with the above deduction of Proposition 3.11 using rotations, we sketch another argument, albeit skipping some tedious calculations. Indeed, if the reader is happy with Proposition 3.4 (which, being honest here, we did not actually prove in the generality in which we intend to apply it at this point), one can also note that the area of the parallelogram spanned by $\vec{v}$ and $\vec{w}$ is given by

$$\sqrt{\det\left(\begin{pmatrix} | & | \\ \vec{v} & \vec{w} \\ | & | \end{pmatrix}^{\mathsf{T}} \begin{pmatrix} | & | \\ \vec{v} & \vec{w} \\ | & | \end{pmatrix}\right)}.$$

A sort of tedious calculation shows that this is equal to the square root of

$$v_1^2w_2^2 + v_1^2w_3^2 + v_2^2w_1^2 + v_2^2w_3^2 + v_3^2w_1^2 + v_3^2w_2^2 - 2v_1v_2w_1w_2 - 2v_1v_3w_1w_3 - 2v_2v_3w_2w_3.$$

Hence, to prove Proposition 3.11, one needs only check that $\|\vec{v} \times \vec{w}\|$ gives the same expression. We do not do this, but a reader who wants to will have no difficulty doing so. □

Basic geometry tells us that the area of the parallelogram spanned by $\vec{v}$ and $\vec{w}$ can also be computed from their length and the angle between them. Combining this with Proposition 3.11 yields the following result.

Corollary 3.12. $\|\vec{v} \times \vec{w}\| = \|\vec{v}\| \, \|\vec{w}\| \, |\sin \measuredangle(\vec{v}, \vec{w})|$.

We remind the reader of the observation made in Figure 34, that the value $|\sin \measuredangle(\vec{v}, \vec{w})|$ does not depend on which of the two possible choices for the angle $\measuredangle(\vec{v}, \vec{w})$ one picks.

Example. The two vectors $\vec{v} = (2, 0, 0)$ and $\vec{w} = (1, 1, 0)$ span a parallelogram in $\mathbb{R}^3$ with area

$$\|\vec{v} \times \vec{w}\| = \left\| \begin{pmatrix} 2 \\ 0 \\ 0 \end{pmatrix} \times \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} \right\| = \left\| \begin{pmatrix} 0 \\ 0 \\ 2 \end{pmatrix} \right\| = \sqrt{0^2 + 0^2 + 2^2} = 2.$$

This is geometrically obvious, because the sought-after area is also the area of the parallelogram spanned by $\vec{v}$ and $\vec{w} - \frac{1}{2}\vec{v} = (0, 1, 0)$, i.e., the area of the rectangle spanned by $2\vec{e}_1$ and $\vec{e}_2$, which is 2.

We close this section with a three-dimensional variant of Cramer's rule (Proposition 3.2) phrased using the cross product. This particular variant seems to be especially popular in solid state physics to compute the "reciprocal lattice" associated to the lattice spanned by a set $\{\vec{v}, \vec{w}, \vec{z}\}$ of "primitive lattice vectors".

Theorem 3.13 (Cramer's rule, variant)**.** *One has the following formula for inverting a* 3×3*-matrix provided that the denominator in the fraction on the right hand side if non-zero:*

$$\begin{pmatrix} | & | & | \\ \vec{v} & \vec{w} & \vec{z} \\ | & | & | \end{pmatrix}^{-1} = \frac{1}{\vec{v} \cdot (\vec{w} \times \vec{z})} \begin{pmatrix} | & | & | \\ (\vec{w} \times \vec{z}) & -(\vec{v} \times \vec{z}) & (\vec{v} \times \vec{w}) \\ | & | & | \end{pmatrix}^{\mathsf{T}}.$$

Proof. This is simply Proposition 3.2 in disguise (as one sees after unpacking the definition of the cross product here). □

3.5. Eigenvalues and eigenvectors

Consider the linear map $f : \mathbb{R}^3 \to \mathbb{R}^3$ given by the matrix

$$A = \begin{pmatrix} \frac{342}{61} & -\frac{8}{61} & -\frac{265}{61} \\ \frac{165}{61} & \frac{116}{61} & -\frac{275}{61} \\ \frac{165}{61} & -\frac{6}{61} & -\frac{92}{61} \end{pmatrix}. \tag{3.8}$$

It maps $\vec{e}_1$ to the vector $(\frac{342}{61}, \frac{165}{61}, \frac{165}{61})$, for instance. This looks reasonably complicated. When calculating $f(\vec{v})$ of an arbitrary vector $\vec{v} = (v_1, v_2, v_3)$, we have to add all the columns of A scaled by the v_is. This may cause some headache.

It would seem much easier if the action of f on $\vec{v}$ were more transparent. The easiest would probably be f acting via scalar multiplication:[8] $f(\vec{v}) = \lambda\vec{v}$ for some $\lambda \in \mathbb{R}$. We have seen already (take $\vec{v} = \vec{e}_1$) that this is not always the case. However, one can check that the vectors

$$\vec{b}_1 = \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix}, \quad \vec{b}_2 = \begin{pmatrix} 2 \\ 55 \\ 0 \end{pmatrix}, \quad \vec{b}_3 = \begin{pmatrix} 5 \\ 0 \\ 3 \end{pmatrix} \tag{3.9}$$

satisfy[9]

$$f(\vec{b}_1) = 1\vec{b}_1, \quad f(\vec{b}_2) = 2\vec{b}_2 \quad \text{and} \quad f(\vec{b}_3) = 3\vec{b}_3.$$

As we know, any vector $\vec{v} \in \mathbb{R}^3$ can be written as

$$\vec{v} = v_1\vec{e}_1 + v_2\vec{e}_2 + v_3\vec{e}_3. \tag{3.10}$$

However, incidentally, any such $\vec{v}$ can also be written as

$$\vec{v} = \hat{v}_1\vec{b}_1 + \hat{v}_2\vec{b}_2 + \hat{v}_3\vec{b}_3 =: \begin{pmatrix} \hat{v}_1 \\ \hat{v}_2 \\ \hat{v}_3 \end{pmatrix}_{\mathscr{B}}, \tag{3.11}$$

where $\mathscr{B} := (\vec{b}_1, \vec{b}_2, \vec{b}_3)$. Therefore, by linearity,

$$f\left(\begin{pmatrix} \hat{v}_1 \\ \hat{v}_2 \\ \hat{v}_3 \end{pmatrix}_{\mathscr{B}}\right) = \hat{v}_1 f(\vec{b}_1) + \hat{v}_2 f(\vec{b}_2) + \hat{v}_3 f(\vec{b}_3) = \hat{v}_1 1\vec{b}_1 + \hat{v}_2 2\vec{b}_2 + \hat{v}_3 3\vec{b}_3 = \begin{pmatrix} 1\hat{v}_1 \\ 2\hat{v}_2 \\ 3\hat{v}_3 \end{pmatrix}_{\mathscr{B}}.$$

Hence, suddenly computing the action of f on vectors has become *much* easier, at least when representing vectors in the form (3.11) rather than in (3.10). Note that given a vector $\vec{v}$ in the form (3.11) it is still possible to find its representation in the form (3.10) by using (3.9). Conversely, it is possible to go in the other direction if one has expressions for

$$\vec{e}_j = \hat{e}_{j1}\vec{b}_1 + \hat{e}_{j2}\vec{b}_2 + \hat{e}_{j3}\vec{b}_3 \quad (\text{for } j = 1, 2, 3).$$

These can be obtained by (say) solving an appropriate system of linear equations. In this way one obtains, for instance,

$$\vec{e}_1 = -\tfrac{165}{122}\vec{b}_1 + \tfrac{3}{61}\vec{b}_2 + \tfrac{55}{122}\vec{b}_3 = \begin{pmatrix} -\frac{165}{122} \\ \frac{3}{61} \\ \frac{55}{122} \end{pmatrix}_{\mathscr{B}}.$$

Admittedly, this may look equally ugly as $f(\vec{e}_1) = (\frac{342}{61}, \frac{165}{61}, \frac{165}{61})$ which was what has lead us to consider the vectors (3.9) to begin with. So, has one gained anything from all this mess?—Yes, indeed! The hope here is that choosing another coordinate

[8] Okay, the easiest would rather be f mapping $\vec{v}$ to zero, but this alone would be far too restrictive to be useful and this case is in fact also contained in our discussion (set $\lambda = 0$).

[9] Obviously this looks like magic at this point. An explanation on how one would actually go about finding such vectors in the first place follows later.

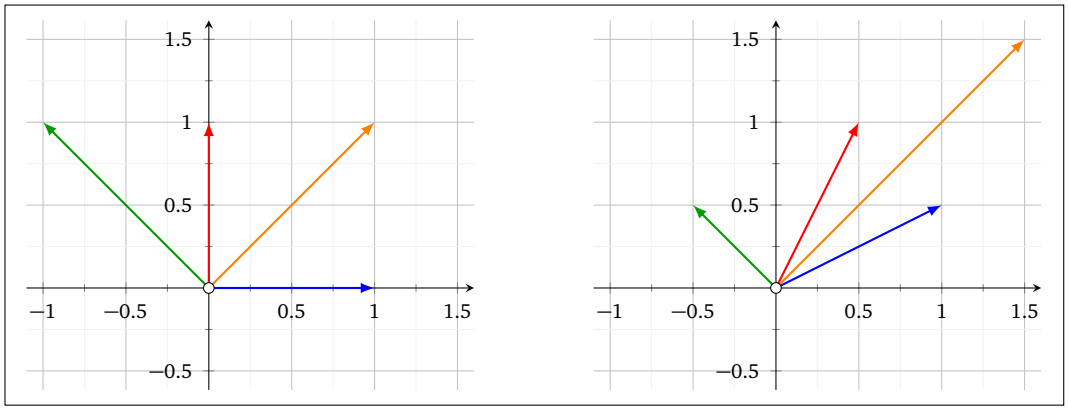


Figure 37. Vectors $\vec{v}$ (left side) being mapped to $A\vec{v}$ (right side), where $A = \begin{pmatrix} 1 & 1/2 \\ 1/2 & 1 \end{pmatrix}$. The standard unit vectors (blue and red on the left hand side) are mapped to somewhat more complicated vectors. On the other hand, the vectors $(1,1)$ and $(-1,1)$ (orange and green on the right hand side) are mapped to multiples of themselves (scaled by $3/2$ and $1/2$ respectively).

system (furnished by $\vec{b}_1$, $\vec{b}_2$ and $\vec{b}_3$ in place of the standard unit vectors $\vec{e}_1$, $\vec{e}_2$ and $\vec{e}_3$) might make any calculations one wants to accomplish sufficiently easy so that the relevant insights can be gained in these coordinates and later (if the need arises) this insight can be transported back to the original coordinate system.

Remark 3.14 (Switching between coordinate systems)**.** Given a vector $\vec{v}$ written as in (3.10), how does one compute the coefficients $\hat{v}_1$, $\hat{v}_2$ and $\hat{v}_2$? Because of

$$\begin{pmatrix} | & | & | \\ \vec{b}_1 & \vec{b}_2 & \vec{b}_3 \\ | & | & | \end{pmatrix} \begin{pmatrix} \hat{v}_1 \\ \hat{v}_2 \\ \hat{v}_3 \end{pmatrix} = \hat{v}_1\vec{b}_1 + \hat{v}_2\vec{b}_2 + \hat{v}_3\vec{b}_3 = \begin{pmatrix} \hat{v}_1 \\ \hat{v}_2 \\ \hat{v}_3 \end{pmatrix}_{\mathscr{B}} = \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix},$$

we find that

$$\begin{pmatrix} \hat{v}_1 \\ \hat{v}_2 \\ \hat{v}_3 \end{pmatrix} = \begin{pmatrix} | & | & | \\ \vec{b}_1 & \vec{b}_2 & \vec{b}_3 \\ | & | & | \end{pmatrix}^{-1} \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix}.$$

The obvious generalisation of this to an n-dimensional setting is also valid: when one wants to represent arbitrary vectors $\vec{v} \in \mathbb{R}^n$ with respect to a new coordinate system furnished by vectors $\vec{b}_1, \ldots, \vec{b}_m \in \mathbb{R}^n$, then one needs $m = n$ (if $m < n$, then not all vectors in $\mathbb{R}^n$ can be represented and if $m > n$ such representations could not

be unique[10]) and the matrix given by

$$\begin{pmatrix} | & \dots & | \\ \vec{b}_1 & \dots & \vec{b}_n \\ | & \dots & | \end{pmatrix} \in \mathbb{R}^{n\times n}$$

needs to be invertible. If this is the case, then the coordinates of a vector $\vec{v} = v_1\vec{e}_1 + \ldots + v_n\vec{e}_n$ with respect to $\vec{b}_1, \ldots, \vec{b}_n$ may be obtained via the formula

$$\begin{pmatrix} \hat{v}_1 \\ \vdots \\ \hat{v}_n \end{pmatrix} = \begin{pmatrix} | & \dots & | \\ \vec{b}_1 & \dots & \vec{b}_n \\ | & \dots & | \end{pmatrix}^{-1} \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix},$$

i.e., $\vec{v} = \hat{v}_1\vec{b}_1 + \ldots + \hat{v}_n\vec{b}_n$.

Now let $A \in \mathbb{R}^{n\times n}$ be an arbitrary (square) matrix and $f\colon \mathbb{R}^n \to \mathbb{R}^n$ the matrix induced by A. Then a scalar $\lambda \in \mathbb{R}$ is called ***eigenvalue*** of A (or f) if there exists a non-zero vector $\vec{b} \in \mathbb{R}^n \setminus \{\vec{0}\}$ such that $A\vec{b} = \lambda\vec{b}$ (i.e., $f(\vec{b}) = \lambda\vec{b}$). Such a vector $\vec{b}$ is then called ***eigenvector*** to the eigenvalue λ of A (or f). Observe that Eigenvectors are never unique: if $\vec{b}$ is an eigenvector to the eigenvalue λ of A, then so is $\mu\vec{b}$ for any $\mu \in \mathbb{R} \setminus \{0\}$. It turns out that an $n\times n$-matrix admits at most n eigenvalues. Indeed, one can check that

$$\begin{aligned}(\exists \vec{b} \in \mathbb{R}^n \setminus \{\vec{0}\}\colon A\vec{b} = \lambda\vec{b}) &\iff \exists \vec{b} \in \mathbb{R}^n \setminus \{\vec{0}\}\colon (\lambda\mathbf{1}_n - A)\vec{b} = \vec{0} \\ &\iff \det(\lambda\mathbf{1}_n - A) = 0.\end{aligned}$$

Now $\chi_f := \chi_A := \det(X\mathbf{1}_n - A)$ turns out to be a polynomial in the variable X of degree n and λ happens to be an eigenvalue of A if and only if λ is a root of that polynomial. χ_A is called the ***characteristic polynomial*** of A. (Note that we prefer to write X for a variable of a polynomial here and reserve the letter λ for its roots. For the most part, this is a matter of personal taste.)

Example. The matrix $A = \begin{pmatrix} 1 & 4 \\ 2 & 3 \end{pmatrix}$ has the characteristic polynomial

$$\det\begin{pmatrix} X-1 & -4 \\ -2 & X-3 \end{pmatrix} = (X-1)(X-3) - (-4)(-2) = X^2 - 4X - 5 = (X-5)(X+1).$$

Hence, the eigenvalues of A are 5 and -1. To find the eigenvectors to the eigenvalue 5, we need to determine the non-zero solutions to the following system of linear equations:

$$(5\cdot\mathbf{1}_2 - A)\vec{b} = \begin{pmatrix} 5-1 & -4 \\ -2 & 5-3 \end{pmatrix}\begin{pmatrix} b_1 \\ b_2 \end{pmatrix} \overset{!}{=} \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

[10]Both stated facts are hopefully intuitive enough given the correct gut feeling that $\mathbb{R}^n$ should be "n-dimensional", although we do not give formal definitions of this here and consequently do not present proofs of this.

It is easy to see that the solutions are $\vec{b} = (b_1, b_1)$ (plus requiring $b_1 \neq 0$ if one is looking only for non-zero solutions). A similar consideration shows that the eigenvectors to the eigenvalue -1 of A are precisely $\vec{b} = (-2b_2, b_2)$, $b_2 \neq 0$. Let

$$\vec{b}_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \quad \text{and} \quad \vec{b}_2 = \begin{pmatrix} -2 \\ 1 \end{pmatrix}.$$

Note that every vector $\vec{v} \in \mathbb{R}^2$ may be written as

$$\vec{v} = \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \frac{v_1 + 2v_2}{3}\vec{b}_1 + \frac{-v_1 + v_2}{3}\vec{b}_2 = \begin{pmatrix} \frac{v_1+2v_2}{3} \\ \frac{-v_1+v_2}{3} \end{pmatrix}_{(\vec{b}_1, \vec{b}_2)}.$$

Example. The matrix $A = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ has the characteristic polynomial

$$\det \begin{pmatrix} X-1 & -1 \\ 0 & X-1 \end{pmatrix} = (X-1)^2.$$

Hence, its only eigenvalue is 1. To compute its eigenvectors we must look for non-trivial solutions to the following system of linear equations:

$$(1 \cdot \mathbf{1}_2 - A)\vec{b} = \begin{pmatrix} 1-1 & -1 \\ 0 & 1-1 \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} \overset{!}{=} \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

These are $\vec{b} = (b_1, 0)$ (plus requiring $b_1 \neq 0$ if one is looking only for non-zero solutions). Note that if we had changed A to the matrix $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = \mathbf{1}_2$, then every (non-zero) vector would have been an eigenvector.

Example. The matrix $A = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$ has the characteristic polynomial

$$\det \begin{pmatrix} X & 1 \\ -1 & X \end{pmatrix} = X^2 + 1.$$

Hence, it has no *real* eigenvalues. However, when looking for complex eigenvalues (which, strictly speaking, we have not defined, because we were always assuming to be working over the real numbers) one finds that $X^2 + 1 = (X - \mathrm{i})(X + \mathrm{i})$ has the roots $\pm\mathrm{i}$. One can then check that $(\mathrm{i}, 1)$ and $(-\mathrm{i}, 1)$ are eigenvectors of A to the eigenvalues i and $-\mathrm{i}$ respectively. One has

$$\begin{pmatrix} \mathrm{i} & -\mathrm{i} \\ 1 & 1 \end{pmatrix}^{-1} A \begin{pmatrix} \mathrm{i} & -\mathrm{i} \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} \mathrm{i} & 0 \\ 0 & -\mathrm{i} \end{pmatrix}.$$

As already in Chapter 2, complex numbers prove useful (provided, one is sufficiently motivated to want A to admit eigenvalues).

Example. By computing its characteristic polynomial, one can check—which we will not do—that the matrix A given in (3.8) has the eigenvalues 1, 2 and 3.

A matrix $A \in \mathbb{R}^{n\times n}$ is called ***diagonalisable*** if there are n eigenvectors forming a coordinate system in the sense of Remark 3.14. This is equivalent to there being an invertible matrix $B \in \mathbb{R}^{n\times n}$ such that $B^{-1}AB$ equals a ***diagonal matrix***, i.e.,

$$B^{-1}AB = \begin{pmatrix} \lambda_1 & 0 & \cdots & \cdots & 0 \\ 0 & \lambda_2 & 0 & & \vdots \\ \vdots & 0 & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ 0 & \cdots & \cdots & 0 & \lambda_n \end{pmatrix}.$$

In this case the diagonal entries $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of A and each eigenvalue λ appears as often as the order of the root λ of χ_A dictates. (That is, $\chi_A = (X-1)^2$ implies that the eigenvalue $\lambda = 1$ appears exactly twice.) The columns of B are then necessarily eigenvectors of A.

Not every matrix is diagonalisable; for instance,

$$A = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$$

is not. Indeed, we have seen previously that all eigenvectors of A are of the form $\vec{b} = (b_1, 0)$, $b_1 \neq 0$. There is no possibility of writing any two such eigenvectors into the columns of a 2×2-matrix and get an invertible matrix, for

$$\det \begin{pmatrix} b_1 & b_1' \\ 0 & 0 \end{pmatrix} = b_1 \cdot 0 - b_1' \cdot 0$$

is always zero.

One can show that eigenvectors of a matrix belonging to different eigenvalues are always "sufficiently independent." This gives the following result, whose proof we omit. (The proof would be quite short, though.)

Proposition 3.15. *If a matrix $A \in \mathbb{R}^{n\times n}$ admits n pairwise distinct eigenvalues, then A is diagonalisable.*

Note that the converse of the above proposition does not hold; for instance, the 2×2 identity matrix $\mathbf{1}_2$ is diagonalisable (in fact, it is already of diagonal shape), yet it has only one eigenvalue, namely 1.

3.6. Solving linear equations

Systems of linear equations show up in many situations (see, for instance, § 3.2.6 or § 3.5).

3.6.1. The naïve method. As a prototype of what we wish to study here, we consider two examples (with a naïve approach).

Example 3.16. Consider the following system of linear equations:

$$\begin{cases} 7X_1 + 3X_2 \stackrel{!}{=} 1, \\ 2X_1 + 1X_2 \stackrel{!}{=} 1. \end{cases} \tag{3.12}$$

This is a system in two variables with two equations. By subtracting the second equation from the first three times, we derive the new equation

$$(7 - 3 \cdot 2)X_1 + (3 - 3 \cdot 2)X_2 = 1 - 3 \cdot 1,$$

that is, $X_1 = -2$. Upon plugging this into the second equation, we obtain

$$1 = 2X_1 + 1X_2 = 2(-2) + 1X_2 = -4 + X_2.$$

By adding 4 to both sides, we find that $X_2 = 5$. Hence, the (unique) solution to (3.12) is $(x_1, x_2) = (-2, 5)$.[11]

Example 3.17. Consider the following system of linear equations:

$$\begin{cases} 7X_1 + 3X_2 + 8X_3 \stackrel{!}{=} 1, \\ 2X_1 + 1X_2 + 0X_3 \stackrel{!}{=} 1. \end{cases} \tag{3.13}$$

As in Example 3.16, we start by subtracting the second equation from the first three times. This time we get

$$(7 - 3 \cdot 2)X_1 + (3 - 3 \cdot 2)X_2 + (8 - 3 \cdot 0)X_3 = 1 - 3 \cdot 1,$$

that is, $X_1 + 8X_3 = -2$. Hence, this time, we do not quite have the value of X_1 pinned down, but still can do so easily: by subtracting $8X_3$, we get the equation $X_1 = -2 - 8X_3$. As before, we plug in this value for X_1 into the second equation. We obtain

$$1 = 2X_1 + 1X_2 = 2(-2 - 8X_3) + 1X_2 = -4 + X_2 - 16X_3.$$

Solving this for X_2, we obtain $X_2 = 5 + 16X_3$. Hence, we have found the following solutions to (3.13):

$$(x_1, x_2, x_3) = (-2 - 8x_3, 5 + 16x_3, x_3). \tag{3.14}$$

Here the equality of the third entries on both sides ('$x_3 = x_3$') looks rather redundant and, indeed, it is. So what *is* x_3 now?—Well, anything one chooses, really. For instance, choosing $x_3 = 0$ and computing x_1 and x_2 according to the above equations, we recover the solution $(x_1, x_2) = (-2, 5)$ found Example 3.16. This should not be surprising, as (3.13) degenerates into (3.12) if one sets x_3 to zero. However, letting $x_3 = 1$ in (3.14), we get $(x_1, x_2, x_3) = (-10, 21, 1)$, which also constitutes a solution to (3.13).

[11] Here we follow the (vague!) convention *variables* in equations to be solved are denoted using capital letters, and particular solutions are denoted using the corresponding lower-case letters. There is certainly some room for interpretation in the above convention. A reader who is confused by this is welcome to just replace all capital letters by their lower-case counter-part (or vice-versa).

3.6.2. Different views on systems of linear equations. Next, we point out a connection to linear maps. Indeed, consider the left hand side of (3.13), for instance. If one substitutes numbers for the three variables X_1, X_2 and X_3, then the left hand side of (3.13) produces two numbers. They might not both equal 1. (In fact, neither of the two need be equal to 1, in general.) However, in any case, one gets a *map*

$$f\colon \mathbb{R}^3 \to \mathbb{R}^2, \quad (x_1,x_2,x_3) \mapsto \begin{array}{l}\text{value of the left hand side of (3.13) when}\\ \text{plugging in } (x_1,x_2,x_3) \text{ for } (X_1,X_2,X_3).\end{array}$$

It is easily checked that this map f is *linear*. The *equation* (3.13) can be re-written as

$$f(X_1,X_2,X_3) \stackrel{!}{=} (1,1). \tag{3.15}$$

Therefore, linear equations reduce the study of certain associated linear maps. The matrix A representing f is plainly given by

$$A = \begin{pmatrix} 7 & 3 & 8 \\ 2 & 1 & 0 \end{pmatrix}$$

and we may re-write (3.15) as

$$\begin{pmatrix} 7 & 3 & 8 \\ 2 & 1 & 0 \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} \stackrel{!}{=} \begin{pmatrix} 1 \\ 1 \end{pmatrix}. \tag{3.16}$$

To sum up, we have found two further (equivalent) ways of thinking about the linear equation (3.13): an interpretation via finding certain values of a linear map (3.15), and an interpretation with matrices (3.16). Albeit all of these viewpoints yield the same solutions, morally, they are different:

(1) In the initial guise of (3.13), we deal with a collection of equations. We feel comfortable with picking out an individual equation, manipulating it, substituting it into another of the given equations etc., as we have done in Example 3.17.
(2) In the guise (3.15) we see a linear map. Solving our equations appears to be intimately connected with that linear map and our understanding of such maps may inform us on how to solve it.
(3) In the guise of (3.16), we see a matrix. We know that matrices are *tools* used for carrying out *calculations*. We might feel that there ought to be some sort of *procedure*—an *algorithm*—for solving (3.16).

Point (2) is ultimately the one that offers access to the most powerful tools (which have counter-parts in (1) and (3), but rightfully belong to (2)). As we have not developed the theory of linear algebra sufficiently deeply, we cannot pursue this much further here. Nevertheless, to give at least a flavour of this, let us mention only that 'dimension theory' (which we have not covered) immediately tells us that (3.15) *either* has *no solutions at all* or *infinitely many solutions*. (The situation in Example 3.16 of finding exactly one solution cannot occur here.) In Example 3.17 we have seen that the latter is the case. The corresponding counter-part in (1) is that

any system of linear equations with more variables than equations admits the above property (either the equations contradict each other somehow, so that there is no solution at all, or the equations can be satisfied, but are insufficient to determine all variables as to have only one solution, thus admitting an infinitude of solutions).

We now stop the rather philosophical discussion from above and turn to the procedure predicted in point (3) above. To this end, we consider the general problem of solving a system of n linear equations in m variables:

$$\begin{cases} a_{11}X_1 + a_{12}X_2 + \ldots + a_{1m}X_m \stackrel{!}{=} b_1, \\ \vdots \quad\quad \vdots \quad\quad \ddots \quad\quad \vdots \quad\quad \vdots \\ a_{n1}X_1 + a_{n2}X_2 + \ldots + a_{nm}X_m \stackrel{!}{=} b_n, \end{cases} \tag{3.17}$$

(Here terms labelled with a_{ij} and b_i ($1 \leq i \leq n$, $1 \leq j \leq m$) are supposed to be real or complex numbers.) The equivalent matrix form of this equation is

$$\begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \ldots & a_{nm} \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_m \end{pmatrix} \stackrel{!}{=} \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix}. \tag{3.18}$$

For obvious reasons, one calls the $n \times m$-matrix on the left hand side of (3.18) the ***coefficient matrix*** of the system (3.17). To save space, one simply writes

$$\left(\begin{array}{cccc|c} a_{11} & a_{12} & \ldots & a_{1m} & b_1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ a_{n1} & a_{n2} & \ldots & a_{nm} & b_n \end{array}\right) \tag{3.19}$$

The matrix above is called ***augmented coefficient matrix***.

3.6.3. Solving a system in row echelon form. If the coefficients of the matrix (3.19) were such that the matrix takes a certain special form, then we can easily read-off the solutions to our equation. Writing down a formal version of what we mean would be utterly incomprehensible due to many indices, so we just give an example.

Example 3.18. Consider the equation encoded in

$$\left(\begin{array}{cccccc|c} 1 & a_{12} & a_{13} & a_{14} & a_{15} & a_{16} & b_1 \\ & 1 & a_{23} & a_{24} & a_{25} & a_{26} & b_2 \\ & & & 1 & a_{35} & a_{36} & b_3 \\ & & & & & & b_4 \\ & & & & & & b_5 \end{array}\right), \tag{3.20}$$

where empty entries ought to be interpreted as zero entries. (This particular shape, resembling a staircase, is called ***(reduced) row echelon form***;[12] individual steps

[12]The word 'reduced' means that we require the entries on the stairs to be normalised to equal 1.

in the staircase may be arbitrarily wide and high in general.) The two last rows correspond to equations with all zero coefficients in front of the variables and right hand side equal to b_4 or b_5 respectively. These equations are soluble if and only if $b_4 = b_5$. Assume that this is the case. Then it is possible to find all solutions of the equation in question by the following procedure. The variables X_5 and X_6 are 'free', meaning that their value can be set to arbitrary numbers. Suppose that such values were chosen. (If one is interested in finding *all* solutions, then one must consider the following for all choices of X_5 and X_6.) Now look at the equation encoded by the third row of our matrix. It is

$$0X_1 + 0X_2 + 0X_3 + 1X_4 + a_{35}X_5 + a_{36}X_6 \stackrel{!}{=} b_3.$$

By rearranging, this can be written equivalently as

$$X_4 \stackrel{!}{=} b_3 - a_{35}X_5 - a_{36}X_6.$$

Note that, having chosen values for X_5 and X_6 before, this completely determines the value of X_4. Next, we look at the equation encoded by the second row of our matrix:

$$0X_1 + 1X_2 + a_{23}X_3 + a_{24}X_4 + a_{25}X_5 + a_{26}X_6 \stackrel{!}{=} b_2.$$

Once more, we rearrange, getting

$$X_2 \stackrel{!}{=} b_2 - a_{23}X_3 - a_{24}X_4 - a_{25}X_5 - a_{26}X_6.$$

Here the value of X_2 is determined if we know the values of X_3, X_4, X_5 and X_6. By the above, this is already the case for $X_4, \ldots, X_6$, but X_3 is not fixed; it is again 'free'. Hence, pick an arbitrary value for X_3, so that the value of X_2 is determined. (Again, if one wants to find *all* solutions, then one must consider the following for all possible values of X_3.) Lastly, look at the equation encoded by the first row of our matrix:

$$1X_1 + a_{25}X_2 + a_{35}X_3 + a_{45}X_4 + a_{15}X_5 + a_{16}X_6 \stackrel{!}{=} b_1.$$

By rearranging,

$$X_1 \stackrel{!}{=} b_1 - a_{25}X_2 - a_{35}X_3 - a_{45}X_4 - a_{15}X_5 - a_{16}X_6.$$

3.6.4. Transforming into row echelon form; Gauß's algorithm. Returning to the general equation (3.18) encoded by the matrix (3.19), we shall now describe a procedure for producing a matrix in row echelon form which encodes a system of linear equations with precisely the same solutions. This last system can then subsequently be solved by working as in Example 3.18. The procedure for producing this row echelon form is known as ***Gauß's algorithm***. It uses the following manipulations, aptly dubbed ***row operations***:

(1) ***Interchanging of two rows***:

$$\begin{pmatrix} \vdots & \vdots & & \vdots & \vdots \\ a_{i1} & a_{i2} & \dots & a_{im} & b_i \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ a_{j1} & a_{j2} & \dots & a_{jm} & b_j \\ \vdots & \vdots & & \vdots & \vdots \end{pmatrix} \begin{array}{l}\longleftarrow\\ \\ \\ \\ \longleftarrow\end{array} \rightsquigarrow \begin{pmatrix} \vdots & \vdots & & \vdots & \vdots \\ a_{j1} & a_{j2} & \dots & a_{jm} & b_j \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ a_{i1} & a_{i2} & \dots & a_{im} & b_i \\ \vdots & \vdots & & \vdots & \vdots \end{pmatrix}.$$

When viewing the matrix as a system of linear equation, this corresponds to interchanging the i-th and j-th equation with each other. Clearly this does not affect the set of solutions to the system.

(2) ***Rescaling a row by a non-zero number λ***:

$$\begin{pmatrix} \vdots & \vdots & & \vdots & \vdots \\ a_{i1} & a_{i2} & \dots & a_{im} & b_i \\ \vdots & \vdots & & \vdots & \vdots \end{pmatrix} \mid \cdot\lambda \rightsquigarrow \begin{pmatrix} \vdots & \vdots & & \vdots & \vdots \\ \lambda a_{i1} & \lambda a_{i2} & \dots & \lambda a_{im} & \lambda b_i \\ \vdots & \vdots & & \vdots & \vdots \end{pmatrix}.$$

In terms of corresponding equation this simply multiplies one equation with λ. Since this operation is invertible (multiply by λ^{-1} to reverse it), it does not change the set of solutions either.

(3) ***Adding one row to another***:

$$\begin{pmatrix} \vdots & & \vdots & \vdots \\ a_{i1} & \dots & a_{im} & b_i \\ \vdots & \ddots & \vdots & \vdots \\ a_{j1} & \dots & a_{jm} & b_j \\ \vdots & & \vdots & \vdots \end{pmatrix} \begin{array}{l}\\ \\ \\ \longleftarrow{+}\end{array} \rightsquigarrow \begin{pmatrix} \vdots & & \vdots & \vdots \\ a_{i1} & \dots & a_{im} & b_i \\ \vdots & \ddots & \vdots & \vdots \\ a_{j1}+a_{i1} & \dots & a_{jm}+a_{im} & b_j+b_i \\ \vdots & & \vdots & \vdots \end{pmatrix}.$$

In terms of equations, this means replacing the j-th equation by

$$(a_{j1}+a_{i1})X_1+\ldots+(a_{jm}+a_{im})X_m \stackrel{!}{=} (b_j+b_i). \tag{3.21}$$

Clearly any solution to the original system satisfies both equations

$$\begin{cases} a_{i1}X_1+\ldots+a_{im}X_m \stackrel{!}{=} b_i, \\ a_{j1}X_1+\ldots+a_{jm}X_m \stackrel{!}{=} b_j, \end{cases}$$

also, therefore, also their sum (3.21). Hence, solutions to the original system also solve the new system. On the other hand, rescaling the i-th equation by -1, adding this new equation to our new j-th equation (3.21), and subsequently rescaling the i-th equation by -1 once more, reproduces the original system of equations, we see that our row addition process also does not introduce new solutions.

For illustration's sake, let us give numerical examples of the above:

Examples (Row operations)**.**

(1) Interchanging of two rows:

$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 1 & 1 & 1 \\ 0 & 2 & 0 & 0 \end{pmatrix} \begin{matrix} \leftarrow \\ \\ \leftarrow \end{matrix} \rightsquigarrow \begin{pmatrix} 0 & 2 & 0 & 0 \\ 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 \end{pmatrix}.$$

(2) Rescaling a row by a non-zero number λ:

$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 1 & 1 & 1 \\ 0 & 2 & 0 & 0 \end{pmatrix} \begin{matrix} \\ | \cdot 8 \\ \\ \end{matrix} \rightsquigarrow \begin{pmatrix} 1 & 2 & 3 & 4 \\ 8 & 8 & 8 & 8 \\ 0 & 2 & 0 & 0 \end{pmatrix}.$$

(3) Adding one row to another:

$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 1 & 1 & 1 \\ 0 & 2 & 0 & 0 \end{pmatrix} \begin{matrix} \\ \\ \leftarrow + \end{matrix} \rightsquigarrow \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 3 & 4 & 5 \\ 0 & 2 & 0 & 0 \end{pmatrix}.$$

The steps (2) and (3) can be combined to add any non-zero multiple of one row to another. For instance,

$$\begin{pmatrix} 1 & 2 \\ 2 & 5 \end{pmatrix} \begin{matrix} | \cdot (-2) \\ \leftarrow \end{matrix} \begin{matrix} \\ + \end{matrix} \begin{matrix} | \cdot (-\frac{1}{2}) \\ \\ \end{matrix} \rightsquigarrow \begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix},$$

which we shall shorten as

$$\begin{pmatrix} 1 & 2 \\ 2 & 5 \end{pmatrix} \begin{matrix} \cdot (-2) \\ \leftarrow \end{matrix} \begin{matrix} \\ + \end{matrix} \rightsquigarrow \begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix},$$

can be obtained as the following sequence of row operations:

$$\begin{pmatrix} 1 & 2 \\ 2 & 5 \end{pmatrix} \begin{matrix} | \cdot (-2) \\ \\ \end{matrix} \rightsquigarrow \begin{pmatrix} -2 & -4 \\ 2 & 5 \end{pmatrix} \begin{matrix} \\ \leftarrow + \end{matrix} \rightsquigarrow \begin{pmatrix} -2 & -4 \\ 0 & 1 \end{pmatrix} \begin{matrix} | \cdot (-\frac{1}{2}) \\ \\ \end{matrix} \rightsquigarrow \begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix}.$$

We shall write

$$A \rightsquigarrow B$$

if the matrix B arises from the matrix A after application of (potentially multiple) steps of the above shape.

We are now ready to describe ***Gauß's algorithm***. We do so, using an example, because the general formulation is somewhat awkward to write out (lots of indices and lots of dots). We wish to transform

$$A = \begin{pmatrix} 0 & 1 & 2 & 8 & 4 \\ 2 & 2 & 2 & 2 & 0 \\ 1 & 3 & 1 & 5 & 4 \end{pmatrix}$$

into row echelon form. The algorithm works by generating the sought-after row echelon form one row at a time, starting with the first row. It takes (at most) as many steps as the number of rows of A. In order to uphold the progress made during earlier steps, in step k the algorithm only considers the rows $k, k+1, \ldots$ of A.

For the first step, we look at the first *column* and pick any row with a non-zero entry. In our example, let us pick row number two:

$$A = \begin{pmatrix} 0 & 1 & 2 & 8 & 4 \\ \boxed{2} & 2 & 2 & 2 & 0 \\ 1 & 3 & 1 & 5 & 4 \end{pmatrix}.$$

(If we cannot find a non-zero entry in the first column, we move on to the second column and so on. If no non-zero entry can be found at all, then the algorithm terminates.) The boxed element is often referred to as ***pivot element***. Now we swap the first row and our chosen second row:

$$A = \begin{pmatrix} 0 & 1 & 2 & 8 & 4 \\ \boxed{2} & 2 & 2 & 2 & 0 \\ 1 & 3 & 1 & 5 & 4 \end{pmatrix} \begin{matrix} \leftarrow \\ \leftarrow \end{matrix} \rightsquigarrow \begin{pmatrix} \boxed{2} & 2 & 2 & 2 & 0 \\ 0 & 1 & 2 & 8 & 4 \\ 1 & 3 & 1 & 5 & 4 \end{pmatrix}.$$

We now multiply our new first two by the multiplicative inverse of the boxed entry, namely $1/2$. This yields

$$A \rightsquigarrow \begin{pmatrix} \boxed{2} & 2 & 2 & 2 & 0 \\ 0 & 1 & 2 & 8 & 4 \\ 1 & 3 & 1 & 5 & 4 \end{pmatrix} \begin{matrix} \\ \\ \end{matrix} \begin{matrix} | \cdot \frac{1}{2} \\ \\ \end{matrix} \rightsquigarrow \begin{pmatrix} \boxed{1} & 1 & 1 & 1 & 0 \\ 0 & 1 & 2 & 8 & 4 \\ 1 & 3 & 1 & 5 & 4 \end{pmatrix}.$$

Now we look at the first column and add suitable multiples of our first row to the other rows as to turn all entries below the box to zero. In our example, we thus add -1 times the first row to the third row, getting

$$A \rightsquigarrow \begin{pmatrix} \boxed{1} & 1 & 1 & 1 & 0 \\ 0 & 1 & 2 & 8 & 4 \\ 1 & 3 & 1 & 5 & 4 \end{pmatrix} \begin{matrix} \\ \\ \leftarrow \end{matrix} \begin{matrix} \\ \\ {\scriptstyle +} \end{matrix} \begin{matrix} \\ \\ \cdot(-1) \end{matrix} \rightsquigarrow \begin{pmatrix} \boxed{1} & 1 & 1 & 1 & 0 \\ 0 & 1 & 2 & 8 & 4 \\ 0 & 2 & 0 & 4 & 4 \end{pmatrix}.$$

Now the first step is complete.

We move to the second step, which is essentially a copy of the first, but now we work with the second row instead and completely ignore the first row. First, we need to find a non-zero entry in the second column. We find such an entry in the second and in the third row (also in the first, but this row we ignore). Let us select the third row:

$$A \rightsquigarrow \begin{pmatrix} 1 & 1 & 1 & 1 & 0 \\ 0 & 1 & 2 & 8 & 4 \\ 0 & \boxed{2} & 0 & 4 & 4 \end{pmatrix}.$$

(Once more, if we had not found a non-zero entry in the second column [ignoring any entries in the first row], then we would move to the third column and so on.)

We swap it with the second row:

$$A \rightsquigarrow \begin{pmatrix} 1 & 1 & 1 & 1 & 0 \\ 0 & 1 & 2 & 8 & 4 \\ 0 & \boxed{2} & 0 & 4 & 4 \end{pmatrix} \begin{array}{l} \leftarrow \\ \leftarrow \end{array} \rightsquigarrow \begin{pmatrix} 1 & 1 & 1 & 1 & 0 \\ 0 & \boxed{2} & 0 & 4 & 4 \\ 0 & 1 & 2 & 8 & 4 \end{pmatrix}.$$

Now we rescale the new second row as to turn the boxed entry into a 1:

$$A \rightsquigarrow \begin{pmatrix} 1 & 1 & 1 & 1 & 0 \\ 0 & \boxed{2} & 0 & 4 & 4 \\ 0 & 1 & 2 & 8 & 4 \end{pmatrix} \; | \cdot \tfrac{1}{2} \rightsquigarrow \begin{pmatrix} 1 & 1 & 1 & 1 & 0 \\ 0 & \boxed{1} & 0 & 2 & 2 \\ 0 & 1 & 2 & 8 & 4 \end{pmatrix}.$$

Below the box, there is one entry which is non-zero. We now turn this into a zero by adding -1 times the second row to the third:

$$A \rightsquigarrow \begin{pmatrix} 1 & 1 & 1 & 1 & 0 \\ 0 & \boxed{1} & 0 & 2 & 2 \\ 0 & 1 & 2 & 8 & 4 \end{pmatrix} \begin{array}{l} \cdot(-1) \\ + \end{array} \rightsquigarrow \begin{pmatrix} 1 & 1 & 1 & 1 & 0 \\ 0 & \boxed{1} & 0 & 2 & 2 \\ 0 & 0 & 2 & 6 & 2 \end{pmatrix}.$$

Remark 3.19. Optionally, we could now also subtract the second row from the first, to clear the entry *above* our boxed 1. This comment (with the obvious adaptations) also applies to subsequent step.

Now we move to the third step. We now look in the third column for a row with non-zero entry, ignoring the first and the second row. This only leaves one row anyway, namely the third:

$$A \rightsquigarrow \begin{pmatrix} 1 & 1 & 1 & 1 & 0 \\ 0 & 1 & 0 & 2 & 2 \\ 0 & 0 & \boxed{2} & 6 & 2 \end{pmatrix}.$$

Again, we normalise, getting

$$A \rightsquigarrow \begin{pmatrix} 1 & 1 & 1 & 1 & 0 \\ 0 & 1 & 0 & 2 & 2 \\ 0 & 0 & \boxed{2} & 6 & 2 \end{pmatrix} \; | \cdot \tfrac{1}{2} \rightsquigarrow \begin{pmatrix} 1 & 1 & 1 & 1 & 0 \\ 0 & 1 & 0 & 2 & 2 \\ 0 & 0 & \boxed{1} & 3 & 1 \end{pmatrix}. \tag{3.22}$$

Since there are no rows below the boxed one, there are no non-zero entries to get rid of. Therefore, the third step is done. In fact, the matrix we have obtained is in row echelon form.

Example. Let us solve the linear system

$$\begin{pmatrix} 0 & 1 & 2 & 8 \\ 2 & 2 & 2 & 2 \\ 1 & 3 & 1 & 5 \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \end{pmatrix} \overset{!}{=} \begin{pmatrix} 4 \\ 0 \\ 4 \end{pmatrix}. \tag{3.23}$$

The corresponding augmented coefficient matrix is

$$A = \left(\begin{array}{cccc|c} 0 & 1 & 2 & 8 & 4 \\ 2 & 2 & 2 & 2 & 0 \\ 1 & 3 & 1 & 5 & 4 \end{array}\right),$$

our matrix from our description of Gauß's algorithm. Above we have seen that A can be transformed into the following matrix in row echelon form (recall (3.22)):

$$A \rightsquigarrow \left(\begin{array}{cccc|c} 1 & 1 & 1 & 1 & 0 \\ 0 & 1 & 0 & 2 & 2 \\ 0 & 0 & 1 & 3 & 1 \end{array}\right).$$

Now we list all solutions (x_1, x_2, x_3, x_4) to our system (3.23) by the bottom-up procedure described in Example 3.18. The value of x_4 is allowed to be arbitrary. Then $x_3 = 1 - 3x_4$ by the last row of the above matrix. Furthermore, $x_2 = 2 - 2x_4$ by the second row, and $x_1 = -x_2 - x_3 - x_4$ by the first row.

To give an even more explicit description of the solutions to (3.23), we can use our expressions for the various variables to simplify even more. Indeed,

$$x_1 = -x_2 - x_3 - x_4 = -(2 - 2x_4) - (1 - 3x_4) - x_4 = -3 + 4x_4.$$

Thinking of x_4 as a parameter, it is customary to replace it by a different symbol, λ say. Then we can write the set of solutions (x_1, x_2, x_3, x_4) of our system (3.23) in parametric form:[13]

$$\{(-3 + 4\lambda, 2 - 2\lambda, 1 - 3\lambda, \lambda) : \lambda \in \mathbb{R}\}.$$

For instance, taking $\lambda = 0$, we obtain the particular solution $(-3, 2, 1, 0)$.

Remark. When transforming a matrix into row echelon form, one use the row operations described above in a different succession, as one sees fit. As we have argued above, none of these change the set of solutions to the underlying system of linear equations. Hence, no such deviation constitutes a mistake and if one comes up with a matrix in row echelon form faster, then so be it. For instance, when facing a matrix of the shape

$$\left(\begin{array}{cccccc} 1 & 1 & 1 & 1 & 1 & 1 \\ \hline \star & ? & ? & ? & ? & ? \\ \star & ? & ? & ? & ? & ? \\ \star & ? & ? & ? & ? & ? \\ \hline 0 & 1 & 1 & 1 & 1 & 1 \end{array}\right),$$

one may wish to start by subtracting the last row from the first, before subtracting the first row from the middle rows as to annihilate all entries marked with '$\star$'. Doing it this way seems easier, because then only zeros get added to the '?' entries, which makes the computation particularly easy.

[13] We have written '$\lambda \in \mathbb{R}$', but if one is interested in complex solutions, one need only write '$\lambda \in \mathbb{C}$' instead.

Let us now use Gauß's algorithm on the linear system from Example 3.17.

Example 3.20. The system (3.13) of linear equations from Example 3.17 can be written using the following augmented coefficient matrix

$$\left(\begin{array}{ccc|c} 7 & 3 & 8 & 1 \\ 2 & 1 & 0 & 1 \end{array}\right)$$

We transform this matrix into row echelon form as follows:

$$\begin{pmatrix} 7 & 3 & 8 & 1 \\ 2 & 1 & 0 & 1 \end{pmatrix} \overset{+}{\underset{\cdot(-3)}{\leftarrow\!\!\rightarrow}} \rightsquigarrow \begin{pmatrix} 1 & 0 & 8 & -2 \\ 2 & 1 & 0 & 1 \end{pmatrix} \overset{\cdot(-2)}{\underset{+}{\rightarrow\!\!\leftarrow}} \rightsquigarrow \left(\begin{array}{ccc|c} 1 & 0 & 8 & -2 \\ 0 & 1 & -16 & 5 \end{array}\right).$$

Now we can read off the solutions (x_1, x_2, x_3): x_3 is arbitrary, $x_2 = 5 + 16x_3$ and $x_1 = -2 - 8x_3$. In parametric form, the set of solutions can be given in the form

$$\{(-2-8\lambda, 5+16\lambda, \lambda) : \lambda \in \mathbb{R}\}.$$

3.6.5. Gauß–Jordan algorithm. A variant of Gauß's algorithm, the so-called ***Gauß–Jordan algorithm***, computes the inverse matrix A^{-1} of a given $n \times n$-matrix A if this inverse exists. The inverse matrix of A is determined by the equation

$$AA^{-1} \overset{!}{=} \mathbf{1}_n.$$

If $\vec{b}_1, \ldots, \vec{b}_n$ are the columns of A^{-1}, then the above can be rewritten as a system of systems of linear equations:

$$\begin{cases} A\vec{b}_1 \overset{!}{=} \vec{e}_1, & \text{(system 1)} \\ \quad\vdots \\ A\vec{b}_n \overset{!}{=} \vec{e}_n. & \text{(system } n\text{)} \end{cases} \tag{3.24}$$

To solve these systems, one could apply Gauß's algorithm n times. However, the Gauß–Jordan algorithm proposes, to solve all these systems simultaneously. To do this, look at the augmented coefficient matrix

$$(A \mid \mathbf{1}_n) \in \mathbb{R}^{n \times (n+n)},$$

which consists of the columns of A followed by the columns of the $n \times n$ identity matrix. Now apply Gauß's algorithm with the optional steps mentioned in Remark 3.19 to transform the above matrix to

$$(A \mid \mathbf{1}_n) \rightsquigarrow \text{(row operations)} \rightsquigarrow (\mathbf{1}_n \mid B),$$

where B is a suitable matrix. If this fails, i.e., if during the algorithm one fails to find a suitable pivot element to eliminate further entries in the matrix, then A is not invertible. However, if this procedure succeeds, then our discussion from above guarantees that the columns of B satisfy (3.24). Therefore, $B = A^{-1}$.

Example. We wish to invert $A = \begin{pmatrix} 7 & 5 \\ 4 & 3 \end{pmatrix}$. To this end, consider

$$
\begin{aligned}
(A \mid \mathbf{1}_2) &= \left(\begin{array}{cccc} 7 & 5 & 1 & 0 \\ 4 & 3 & 0 & 1 \end{array}\right) \begin{array}{l} \leftarrow\!\!\rfloor{\scriptstyle +} \\ \ \ \rfloor{\scriptstyle \cdot(-1)} \end{array} \rightsquigarrow \left(\begin{array}{cccc} 3 & 2 & 1 & -1 \\ 4 & 3 & 0 & 1 \end{array}\right) \begin{array}{l} \ \ \rfloor{\scriptstyle \cdot(-1)} \\ \leftarrow\!\!\rfloor{\scriptstyle +} \end{array} \\
&\rightsquigarrow \left(\begin{array}{cccc} 3 & 2 & 1 & -1 \\ 1 & 1 & -1 & 2 \end{array}\right) \begin{array}{l} \leftarrow\!\!\rfloor{\scriptstyle +} \\ \ \ \rfloor{\scriptstyle \cdot(-3)} \end{array} \rightsquigarrow \left(\begin{array}{cccc} 0 & -1 & 4 & -7 \\ 1 & 1 & -1 & 2 \end{array}\right) \begin{array}{l} \ \ \rfloor \\ \leftarrow\!\!\rfloor{\scriptstyle +} \end{array} \\
&\rightsquigarrow \left(\begin{array}{cccc} 0 & -1 & 4 & -7 \\ 1 & 0 & 3 & -5 \end{array}\right) \begin{array}{l} \leftarrow\!\!\rfloor \\ \ \ \ \end{array} \rightsquigarrow \left(\begin{array}{cccc} 1 & 0 & 3 & -5 \\ 0 & -1 & 4 & -7 \end{array}\right) \begin{array}{l} \\ \mid \cdot(-1) \end{array} \\
&\rightsquigarrow \left(\begin{array}{cc|cc} 1 & 0 & 3 & -5 \\ 0 & 1 & -4 & 7 \end{array}\right) = (\mathbf{1}_2 \mid B).
\end{aligned}
$$

Now, indeed, one can check that $AB = \mathbf{1}_2 = BA$. Therefore, $B = A^{-1}$.

CHAPTER 4

Fourier analysis

In solid state physics one models the atomic (or molecular) structure of solids of interest on lattices, i.e., structures in space that admit some translational symmetry. This assumption on symmetry makes the study of *wave* and *heat propagation* in these solids mathematically feasible. The mathematics of waves and heat diffusion is a topic that we shall discuss later in this course. At the present we only mention that higher dimensional analogues of the "differential operator" $\frac{\mathrm{d}^2}{\mathrm{d}x^2}$ will have a major role to play. We seek to understand the 1-dimensional setting first. This chapter deals with that theory in particular and closes with an outlook of higher-dimensional generalisations.

4.1. Motivation

We let the "differential operator" $\frac{\mathrm{d}^2}{\mathrm{d}x^2}$ act on the infinitely often (continuously) differentiable ("C^∞ functions") 1-periodic functions $\mathbb{R} \to \mathbb{R}$ (such a function f is required to satisfy $f(x+1) = f(x)$ for every $x \in \mathbb{R}$). It is easily seen that the derivative of a differentiable 1-periodic function is again 1-periodic. Therefore, $\frac{d^2}{\mathrm{d}x^2}$ constitutes a map from the set V of 1-periodic C^∞ functions $\mathbb{R} \to \mathbb{R}$. Moreover, any two such functions $f, g \in V$ can be added and multiplied by scalars $\lambda \in \mathbb{R}$:

$$(f+g)(x) \coloneqq f(x) + g(x), \quad (\lambda f)(x) \coloneqq \lambda \cdot f(x) \quad \text{(for all } x \in \mathbb{R}\text{)}.$$

One has again $f+g, \lambda f \in V$. Note that V is similar to $\mathbb{R}^n$ in this regard. Moreover, the map

$$\frac{\mathrm{d}^2}{\mathrm{d}x^2} \colon V \to V, \quad f \mapsto f'',$$

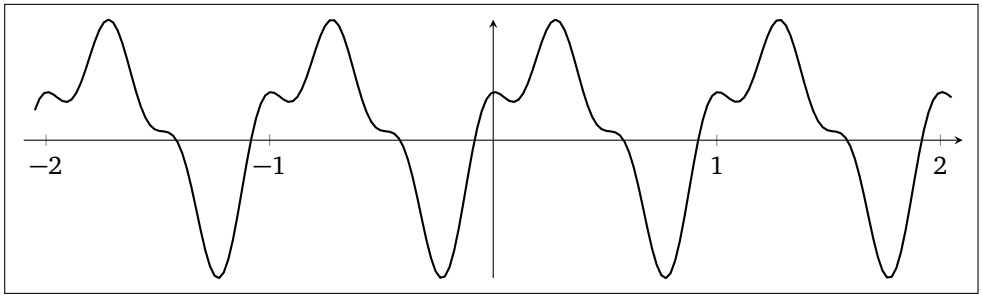


Figure 38. A 1-periodic function.

is *linear* in the sense that

$$\frac{\mathrm{d}^2}{\mathrm{d}x^2}(\lambda f + \mu g) = \lambda \frac{\mathrm{d}^2}{\mathrm{d}x^2} f + \mu \frac{\mathrm{d}^2}{\mathrm{d}x^2} g$$

for any $f, g \in V$ and any $\lambda, \mu \in \mathbb{R}$. In this context (working on a function space V rather than working on $\mathbb{R}^n$) one usually calls $\frac{\mathrm{d}^2}{\mathrm{d}x^2}$ a ***linear operator*** rather than a linear map. For obvious reasons one often also calls this a linear *differential* operator.

Note that unlike $\mathbb{R}^n$ which naturally comes equipped with the standard unit vectors $\vec{e}_1, \ldots, \vec{e}_n$, the space V does not naturally come with such useful luggage. We would have encountered such phenomena even in a finite-dimensional setting if we had spoken abstractly about vector spaces rather than restrict our attention to the base case $\mathbb{R}^n$ altogether. However, we do not feel inclined to remedy this. Instead, we recall that in our earlier discussion on eigenvalues and eigenvectors (whose sole purpose it was to better understand the action of a linear map $f\colon \mathbb{R}^n \to \mathbb{R}^n$ on $\mathbb{R}^n$) we had good reason to exchange our initial coordinate system on $\mathbb{R}^n$ furnished by the standard unit vectors $\vec{e}_1, \ldots, \vec{e}_n$ for a coordinate system consisting of eigenvectors of f (assuming, of course, that such a system existed, which need not always be the case). On the other hand, on V we do not have a "standard" coordinate system. (Actually, we start out with *no* coordinate system at all to begin with!) We now take this opportunity to choose one and nothing seems better suited than choosing such a system consisting of eigenvectors of $\frac{\mathrm{d}^2}{\mathrm{d}x^2}$ on V.

What are eigenvectors of $\frac{\mathrm{d}^2}{\mathrm{d}x^2}$ on V? Or perhaps even more basic: what are the eigenvalues of $\frac{\mathrm{d}^2}{\mathrm{d}x^2}$ on V? As our previous approach with characteristic polynomials seems inapplicable in this setting, we make the basic ansatz of letting $\lambda \in \mathbb{R}$ be as general as possible and trying to find a non-trivial element f ("vector") of V such that

$$\frac{\mathrm{d}^2}{\mathrm{d}x^2} f = \lambda f.$$

This amounts to solving the ordinary differential equation

$$f'' - \lambda f = 0.$$

We find the two independent solutions

$$x \mapsto \exp(x\sqrt{\lambda}) \quad \text{and} \quad x \mapsto \exp(-x\sqrt{\lambda})$$

and any solution f of the above equation is a linear combination of these. However, one can check that, in order for there to be non-trivial (i.e., not identically zero) solutions in V, the above solutions themselves must be periodic. By properties of the exponential function, this happens if and only if $\sqrt{\lambda}$ has vanishing real part. Moreover, the resulting function must have period 1 in order to be an element of V. This implies $\sqrt{\lambda} = 2\pi i k$ for some $k \in \mathbb{Z}$. We arrive at the following candidates(!) for

giving rise to a coordinate system on V tailor-made to study the action of $\frac{\mathrm{d}^2}{\mathrm{d}x^2}$:

$$x \mapsto \exp(2\pi i k x) = e^{2\pi i k x} \quad (k = 0, \pm 1, \pm 2, \ldots).$$

A more rigorous inspection of our above arguments would reveal many flaws and there is a whole branch of mathematics called "functional analysis" that grew out of the desire to correctly address flaws like these. The details are far beyond the scope of this course, but let us still provide a few pointers.

- First, the eigenvectors we have found are $b_k\colon x \mapsto \exp(2\pi i k x)$ and they are complex-valued. This is easily fixed by working with complex-valued functions $\mathbb{R} \to \mathbb{C}$ to begin with or by taking suitable linear combinations to get real-valued functions (cosines and sines!).
- Second, the hope that we may write any $f \in V$ as a linear combination

  $$f(x) = \sum\nolimits_k \hat{f}(k) b_k$$

  where the summation is taken over a *finite* range of k, is ill-advised. Indeed, it can be shown that the kind of infinite-dimensional spaces suitable to do analysis on do *not* admit a *countable* number of elements giving rise to a coordinate system for the full space. (Observe that the sequence of b_k is countable: $b_0, b_{-1}, b_1, b_{-2}, \ldots$) In this sense the operator $\frac{\mathrm{d}^2}{\mathrm{d}x^2}$ admits far too few eigenvalues to satisfy our problem. This can be remedied by allowing for the above sum to be over the set infinite set of all $k \in \mathbb{Z}$. However, then convergence of the resulting series becomes a problem. (This effect is not visible with finite sums, as these always make sense in the present setting.)
- Third, and very much related to our second point: the space C^∞ is quite restrictive to work with. In particular the question of convergence raised in the second point is a nasty one. One way out of this is by enlarging the space under consideration and extending the operator acting on it. In particular the last part is a non-trivial task.

4.2. Fourier series and pointwise convergence

Our previous discussion of diagonalising the operator $\frac{\mathrm{d}^2}{\mathrm{d}x^2}$ acting on 1-periodic C^∞ functions motivates the following question: which functions $f\colon \mathbb{R} \to \mathbb{C}$ can be written as

$$f(x) = \sum_{k=-\infty}^{\infty} \hat{f}(k) e^{2\pi i k x} := \lim_{K \to \infty} \sum_{k=-K}^{K} \hat{f}(k) e^{2\pi i k x}$$

for every (or most?) $x \in \mathbb{R}$, where $\hat{f}(k)$ are supposed to be complex numbers independent of x which we shall view as "the coordinates" of f with respect to the coordinate system $(x \mapsto e^{2\pi i k x})_{k \in \mathbb{Z}}$. It is worth-while to mention that one *should* require some further technical assumptions here, e.g., that the above limit is approached locally uniformly with respect to x (whatever that means). However, we do not do this here and shall be aptly happy with glossing over any technical problems

arising from the lack of such assumptions, well aware that the statements below are not technically correct for all f.

We start with the following observation for $k, \ell \in \mathbb{Z}$:

$$\int_0^1 e^{2\pi i k x} e^{-2\pi i \ell x}\,\mathrm{d}x = \int_0^1 e^{2\pi \mathrm{i}(k-\ell)x}\,\mathrm{d}x = \begin{cases} \displaystyle\int_0^1 1\,\mathrm{d}x = 1 & \text{if } k = \ell, \\ \displaystyle\frac{1}{2\pi \mathrm{i}(k-\ell)} e^{2\pi \mathrm{i}(k-\ell)x}\Big|_{x=0}^{1} = 0 & \text{if } k \neq \ell. \end{cases}$$

We write this succinctly as

$$\int_0^1 e^{2\pi i k x} e^{-2\pi i \ell x}\,\mathrm{d}x = \delta_{k\ell}, \tag{4.1}$$

where $\delta_{k\ell}$ is the ***Kronecker delta symbol*** which equals 1 if $k = \ell$ and zero otherwise. (The index $k\ell$ is to be understood as k, ℓ and not as the product of k and ℓ.) The above formula is known as ***orthogonality relations*** for the functions $x \mapsto e^{2\pi i k x}$ ($k \in \mathbb{Z}$) and can be interpreted as the geometrical statement that these functions are "perpendicular" to each other if the map

$$(f, g) \mapsto \int_0^1 f(x)\overline{g(x)}\,\mathrm{d}x.$$

is interpreted as a generalisation of the well-known dot product of two vectors. (The complex conjugation being present here may seem confusing but would be less so if we had worked with $\mathbb{C}^n$ rather than $\mathbb{R}^n$ earlier. The interested reader is encouraged to consult books on linear algebra for more insight.)

Now assume that f is of the above form and sufficiently nice so that the integration that follows makes sense. We wish to find a formula for the numbers $\hat{f}(k)$. To this end, consider

$$\int_0^1 f(x) e^{-2\pi i \ell x}\,\mathrm{d}x = \int_0^1 \sum_{k=-\infty}^{\infty} \hat{f}(k) e^{2\pi \mathrm{i}(k-\ell)x}\,\mathrm{d}x.$$

We assume that the integral and the infinite series can be readily interchanged.[1] This yields

$$\int_0^1 f(x) e^{-2\pi i \ell x}\,\mathrm{d}x = \sum_{k=-\infty}^{\infty} \int_0^1 \hat{f}(k) e^{2\pi \mathrm{i}(k-\ell)x}\,\mathrm{d}x = \sum_{k=-\infty}^{\infty} \hat{f}(k)\delta_{k\ell} = \hat{f}(\ell).$$

This is a formula for $\hat{f}(\ell)$ using only the values of f (in the integral on the left hand side)!

[1] Actually, interchanging two limiting processes is almost always a subtle endeavour and, indeed, we are scrubbing a serious technical problem under the rug here.

With the previous discussion in mind, let $f\colon \mathbb{R} \to \mathbb{C}$ be a 1-periodic, piecewise continuously differentiable function. Then we define the k-th ***Fourier coefficient*** of f as

$$\hat{f}(k) \coloneqq \int_0^1 f(x) e^{-2\pi i k x} \, \mathrm{d}x. \tag{4.2}$$

Then we have:

Theorem 4.1 (Dirichlet's theorem)**.** *Let $f\colon \mathbb{R} \to \mathbb{C}$ be a 1-periodic, piecewise continuously differentiable function. Then, for every $x \in \mathbb{R}$,*

$$\frac{f(x-) + f(x+)}{2} = \sum_{k=-\infty}^{\infty} \hat{f}(k) e^{2\pi i (k-\ell) x}, \tag{4.3}$$

where $\displaystyle\sum_{k=-\infty}^{\infty} \coloneqq \lim_{K\to\infty} \sum_{k=-K}^{K}$ *and*

$$f(x-) \coloneqq \lim_{\xi \nearrow x} f(\xi) \quad \text{and} \quad f(x+) \coloneqq \lim_{\xi \searrow x} f(\xi)$$

are the left- and right-sided limits of f at the point x. If f is continuous at x, then both of these limits are identical to $f(x)$ and the left hand side of (4.3) equals $f(x)$ as well.

The series in (4.3) is called the ***Fourier series*** of f (evaluated at x). We give a less precise version of Theorem 4.1 which captures the essence of what one might want to remember here:

Corollary 4.2. *Sufficiently "nice" 1-periodic functions can be represented by their Fourier series.*

Proof of Theorem 4.1 (sketch). Looking at the so-called ***Dirichlet kernel***

$$D_K(x) \coloneqq \sum_{k=-K}^{K} e^{2\pi i k x} = \frac{\sin(\pi(2K+1)x)}{\sin(\pi x)}, \tag{4.4}$$

where the last equality can be derived by means of the well-known formula for summing geometric progressions,[2] we compute

$$\begin{aligned}(f * D_K)(x) &\coloneqq \int_0^1 f(\xi) D_K(x-\xi) \, \mathrm{d}\xi = \int_0^1 f(\xi) \sum_{k=-K}^{K} e^{2\pi i k (x-\xi)} \, \mathrm{d}\xi \\ &= \sum_{k=-K}^{K} \int_0^1 f(\xi) e^{-2\pi i k \xi} \, \mathrm{d}\xi \, e^{2\pi i k x} = \sum_{k=-K}^{K} \hat{f}(k) e^{2\pi i k x}.\end{aligned}$$

[2] $x^0 + x^1 + x^2 + \ldots + x^r = \frac{x^{r+1}-1}{x-1}$ for every $x \neq 1$.

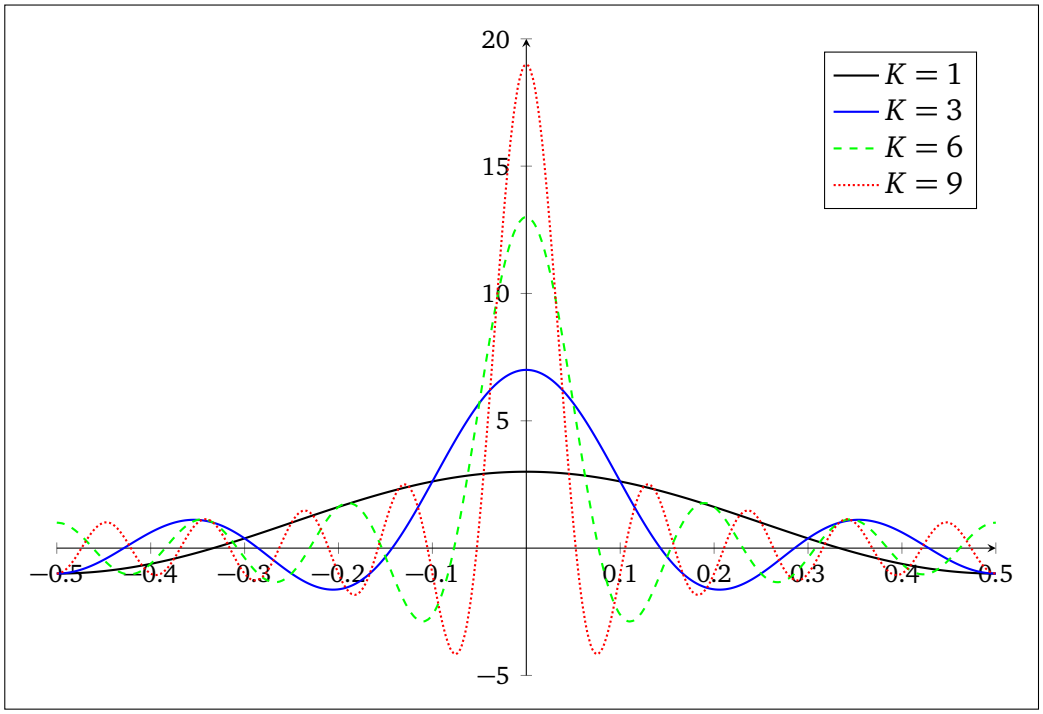


Figure 39. Plot of the Dirichlet kernel D_K from (4.4) for $K = 1, 3, 6, 9$. Observe the increase of the maximal amplitude as well as the increase in oscillations as K increases. (Compare also with Figure 14.)

Now Theorem 4.1 follows if we can show that

$$\lim_{K\to\infty} (f * D_K)(x) = \frac{f(x-) + f(x+)}{2}.$$

This can be accomplished by exploiting the very explicit formula for $D_K(x)$ given on the right hand side of (4.4). By periodicity, we may assume that $0 \leq x < 1$. One can show that the wild oscillations of $D_K(x-\xi)$ about any $\xi \neq x$ (which become ever more wilder as $K \to \infty$) cannot be compensated by the integration against $f(\xi)$. (Technically, this can be shown using partial integration.) This only leaves the contribution to the integral arising from the region where $\xi \approx x$ and this is where one extracts the sought-after term $(f(x-) + f(x+))/2$. The details (which we do not intend to give) are remarkably similar those for Theorem 2.2. □

4.3. Examples

Lemma 4.3 (Linearity of taking Fourier coefficients)**.** *The operation of taking Fourier coefficients is linear: if $f, g\colon \mathbb{R} \to \mathbb{C}$ are 1-periodic and $\lambda, \mu \in \mathbb{C}$ are complex numbers, then*

$$\widehat{(\lambda f + \mu g)} = \lambda \hat{f} + \mu \hat{g}.$$

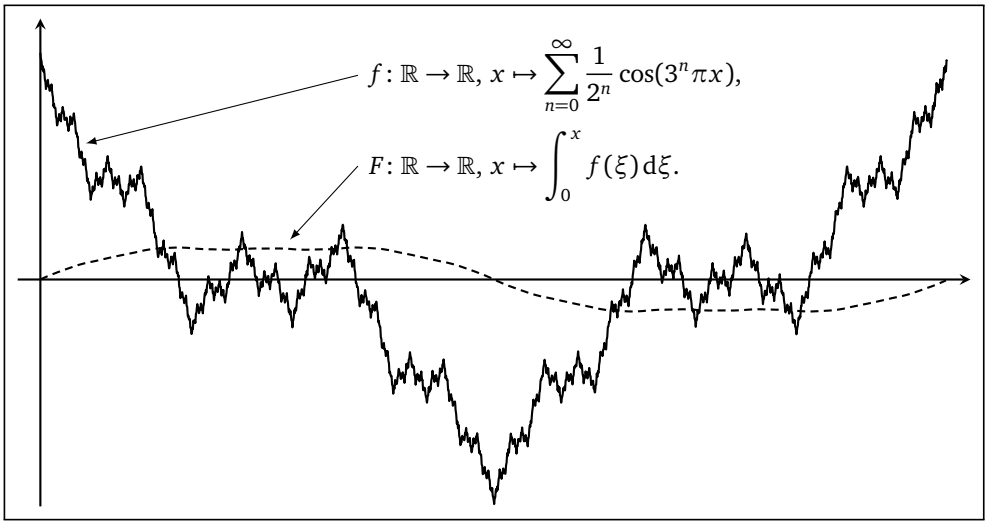


Figure 40. Innocently looking Fourier series can converge to weird things. The above example shows a function f discovered by Weierstraß which is continuous, yet nowhere differentiable. (Note that this also provides an example of a function F which is continuously differentiable, yet whose derivative is itself nowhere differentiable.) This shocked mathematicians at the time.

Proof. This follows immediately from the definition of $\hat{\cdot}$ via integration. □

Example 4.4. The Fourier coefficients of the function $f\colon \mathbb{R} \to \mathbb{C},\ x \mapsto \exp(2\pi \mathrm{i} \ell)$, ($\ell \in \mathbb{Z}$) are given by the Kronecker delta symbol:

$$\hat{f}(k) = \delta_{k\ell} \quad (k \in \mathbb{Z})$$

(recall (4.2) and (4.1)). In particular, taking $\ell = 0$, we find that the Fourier coefficients of the constant-one function $c\colon x \mapsto 1$ are

$$\hat{c}(k) = \delta_{k0} = \begin{cases} 1 & \text{if } k = 0, \\ 0 & \text{if } k \neq 0. \end{cases}$$

(This statement corresponds to the obvious statement that in $\mathbb{R}^n$ the first standard unit vector $\vec{e}_1$ has a 1 in its first component and zeros everywhere else. The previous statement about $\hat{f}$ is the analogue of making the obvious generalisation to the other standard unit vectors.)

$$\vec{e}_1 = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \begin{matrix} \leftarrow \text{one here,} \\ \leftarrow \text{zero elsewhere.} \\ \vdots \end{matrix}$$

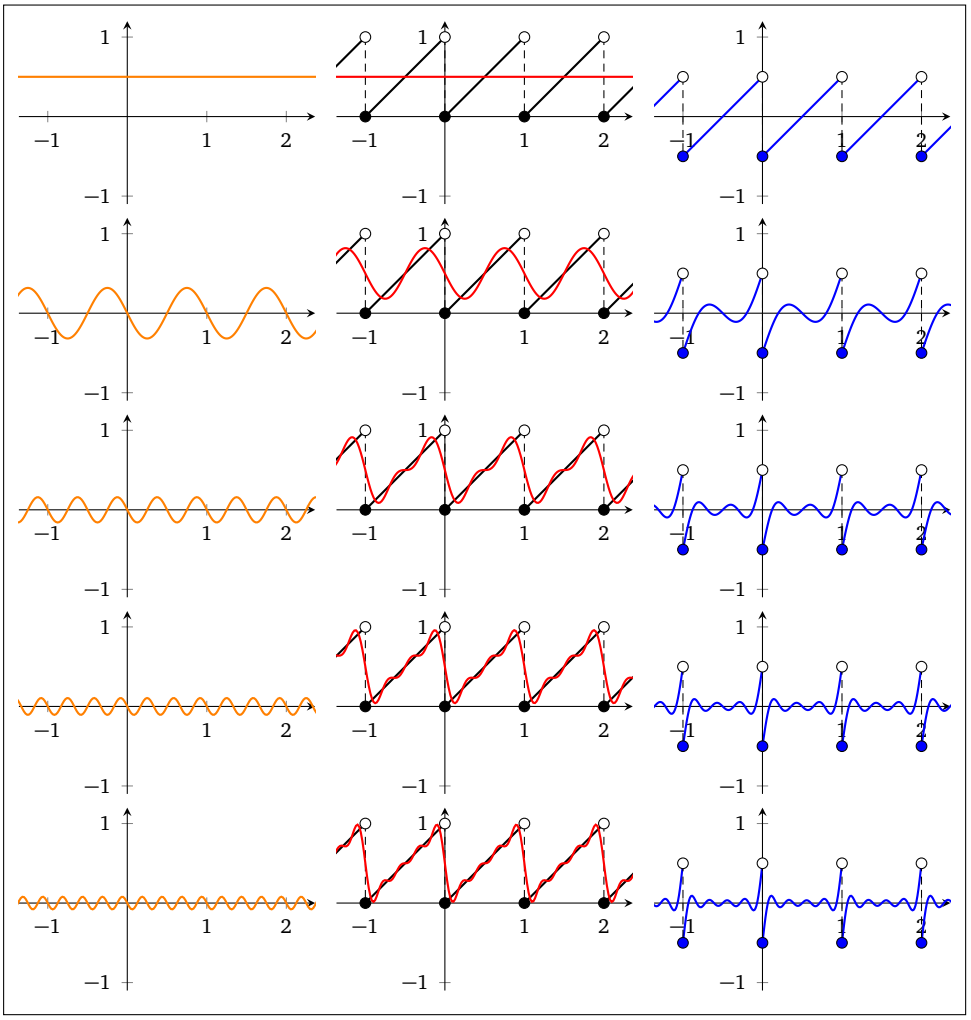


Figure 41. The convergence of the Fourier series of $f = s$ from Example 4.5. The function f is shown in the middle column (black graph). The left column shows

$$\hat{f}(-k)e^{-2\pi ikx} + \hat{f}(k)e^{2\pi ikx},$$

and the middle column shows (red graph)

$$\sum_{|k|\leq K} \hat{f}(k)e^{2\pi ikx} \quad \text{for } K = 0, 1, 2, 3, 4.$$

The left column shows the approximation error, that is, the difference of f and its cut-off Fourier series. One can see clearly that the Fourier series converges to the average of the left- and right-sided limits of s at its jump discontinuities.

Example 4.5 (Saw-tooth function)**.** We shall try to compute the Fourier coefficients of the 1-periodic function $s\colon \mathbb{R} \to \mathbb{R}$ given by $s(x) = x$ for $0 \le x < 1$. We have

$$\hat{s}(k) = \int_0^1 s(x) e^{-2\pi i k x}\,\mathrm{d}x = \int_0^1 x e^{-2\pi i k x}\,\mathrm{d}x.$$

For $k \neq 0$, computing the last integral is an easy exercise in integration by parts:

$$\hat{s}(k) = x \frac{1}{-2\pi i k} e^{-2\pi i k x}\Big|_{x=0}^{1} - \int_0^1 \frac{1}{-2\pi i k} e^{-2\pi i k x}\,\mathrm{d}x.$$

Noting that the last integrand admits the 1-periodic anti-derivative $\frac{1}{(-2\pi i k)^2} e^{-2\pi i k x}$, we see that the corresponding integral vanishes. As so does the first term for $x = 0$, we arrive at

$$\hat{s}(k) = \frac{1}{-2\pi i k} = \frac{i}{2\pi k} \quad (k \neq 0).$$

On the other hand, for $k = 0$ we immediately obtain

$$\hat{s}(0) = \int_0^1 x\,\mathrm{d}x = \frac{x^2}{2}\Big|_{x=0}^{1} = \frac{1}{2}.$$

Proposition 4.6 (Shift and convolution)**.** *Let $f, g\colon \mathbb{R} \to \mathbb{C}$ be two piecewise continuous 1-periodic functions. For $c \in \mathbb{R}$, define the right shift $g_{\to c}\colon \mathbb{R} \to \mathbb{C}$ of g by $g_{\to c}(x) = g(x-c)$ and the convolution $f * g\colon \mathbb{R} \to \mathbb{C}$ of f and g by*[3]

$$(f * g)(x) = \int_0^1 f(\xi) g(x-\xi)\,\mathrm{d}\xi$$

for all $x \in \mathbb{R}$. Then $g_{\to c}$ and $f * g$ are both also 1-periodic and piecewise continuous and we have the formulas

$$\widehat{g_{\to c}}(k) = e^{-2\pi i k c} \hat{g}(k), \quad \widehat{(f * g)}(k) = \hat{f}(k)\hat{g}(k)$$

for all $k \in \mathbb{Z}$.

Proof. We skip proving the periodicity and continuity statements and move straight to the claimed formulas for the Fourier coefficients. We have

$$\widehat{g_{\to c}}(k) = \int_0^1 g_{\to c}(x) e^{-2\pi i k x}\,\mathrm{d}x = \int_0^1 g(x-c) e^{-2\pi i k ((x-c)+c)}\,\mathrm{d}x.$$

Substituting y for $x - c$ we deduce that

$$\widehat{g_{\to c}}(k) = \int_{0-c}^{1-c} g(y) e^{-2\pi i k (y+c)}\,\mathrm{d}y = e^{-2\pi i k c} \int_{-c}^{1-c} g(y) e^{-2\pi i k y}\,\mathrm{d}y.$$

[3] Note that both the definition of the right shift and the definition of the convolution $f * g$ differ from the ones given in Proposition 2.4.

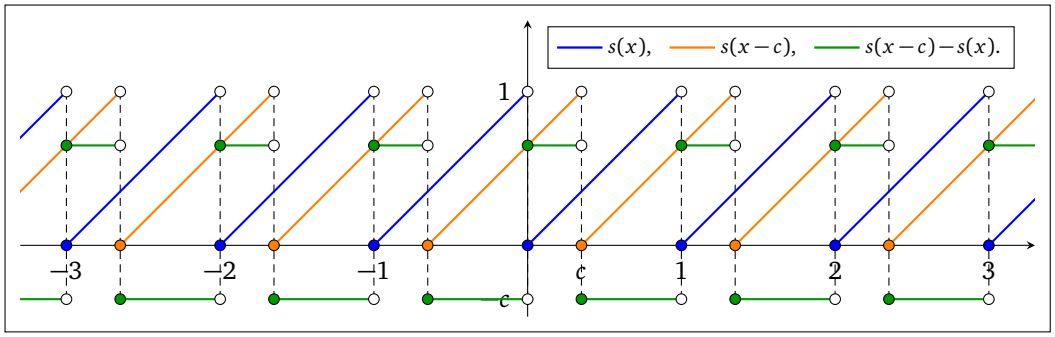


Figure 42. Illustration of the claim made in Example 4.7 that $\chi\colon x \mapsto s(x-c)-s(x)+c$ takes only the values 0 and 1.

We can split the last integral as $\displaystyle\int_{-c}^{1-c} = \int_{-c}^{0} + \int_{0}^{1-c}$ and use 1-periodicity of the integrand on the first integral to see that it equals $\displaystyle\int_{1-c}^{1}$. Putting both integrals back together, we arrive at an integral from 0 to 1 and see that

$$\widehat{g_{\to c}}(k) = e^{-2\pi ikc}\int_0^1 g(y)e^{-2\pi iky}\,\mathrm{d}y = e^{-2\pi ikc}\hat{g}(k).$$

Now we turn to the Fourier coefficients of $f*g$. Using the definition and interchanging the order of integration, we find that

$$\widehat{(f*g)}(k) = \int_0^1\int_0^1 f(\xi)g(x-\xi)\,\mathrm{d}\xi\, e^{-2\pi ikx}\,\mathrm{d}x = \int_0^1 f(\xi)\int_0^1 g(x-\xi)e^{-2\pi ikx}\,\underbrace{\mathrm{d}x\,\mathrm{d}\xi}.$$

The inner integral is seen to equal $\widehat{g_{\to\xi}}(k)$. Therefore, using the formula we have just proved, we have

$$\widehat{(f*g)}(k) = \int_0^1 f(\xi)e^{-2\pi ik\xi}\,\mathrm{d}\xi\,\hat{g}(k) = \hat{f}(k)\hat{g}(k). \qquad \square$$

Example 4.7 (Fourier coefficients of a characteristic function)**.** We shall use Proposition 4.6 to calculate the Fourier coefficients of the characteristic function of an interval (shifted periodically). Indeed, fix a number $c \in (0,1)$ and consider the sawtooth function $s\colon \mathbb{R}\to\mathbb{C}$ from Example 4.5 given by $s(x) = x$ for $0 \le x < 1$ and repeated with period 1. Then the function $\chi\colon \mathbb{R}\to\mathbb{C}$, $x \mapsto s(x-c)-s(x)+c$, is 1-periodic and piecewise continuous. For $0 \le x < 1$ we have

$$\chi(x) = \begin{cases} 1 & \text{if } 0 \le x < c, \\ 0 & \text{if } c \le x < 1. \end{cases}$$

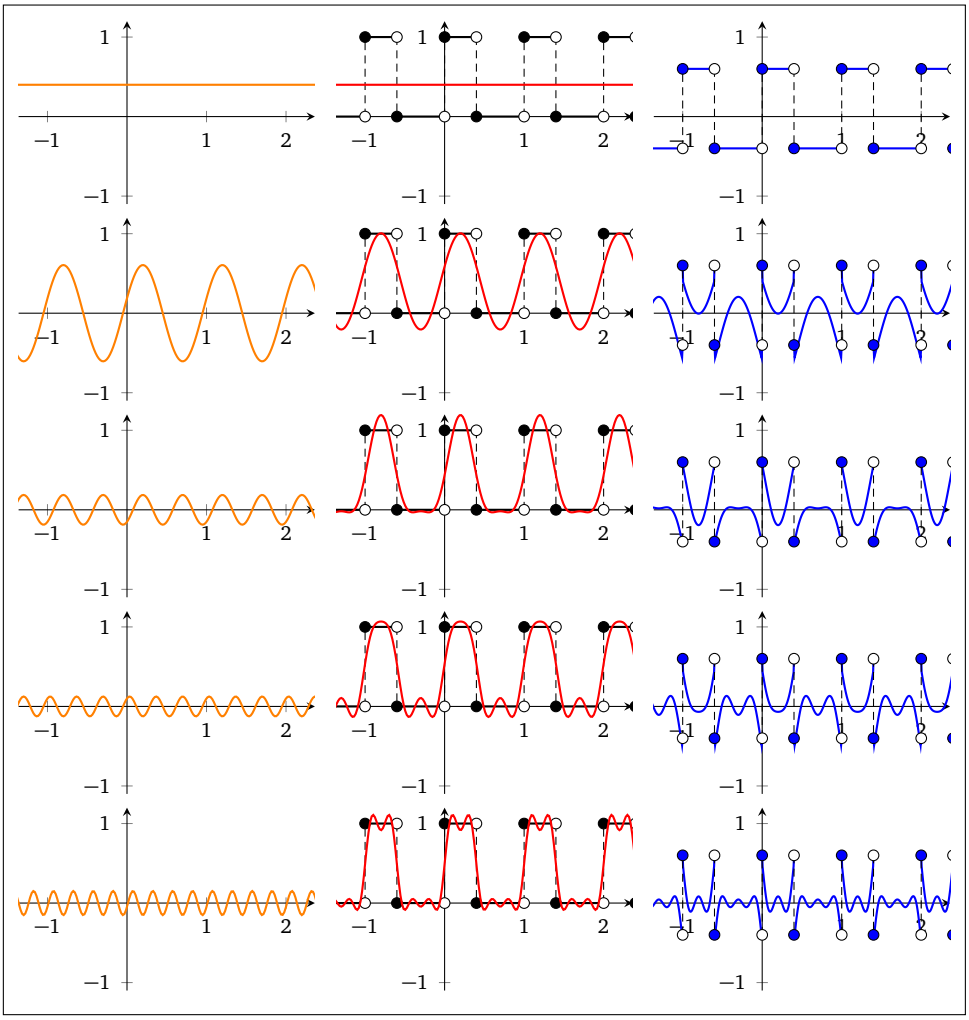


Figure 43. The convergence of the Fourier series of χ with $c = 2/5$ from Example 4.7. (The explanation of the plots is as in Figure 41.)

(Compare Figure 42.) We want to compute the Fourier coefficients of g. By linearity (Lemma 4.3) we have

$$\hat{\chi}(k) = \hat{s}_{\to c}(k) - \hat{s}(k) + \widehat{(x \mapsto c)}(k).$$

Now using Proposition 4.6 and Example 4.4 we find that

$$\hat{\chi}(k) = (e^{-2\pi i k c} - 1)\hat{s}(k) + c\delta_{k0}.$$

From our knowledge of $\hat{s}$ from Example 4.5 we find that for $k \neq 0$

$$\hat{\chi}(k) = \frac{i(e^{-2\pi i k c} - 1)}{2\pi k} \quad \text{and} \quad \hat{\chi}(0) = (e^{-2\pi i 0 c} - 1)\frac{1}{2} + c = c.$$

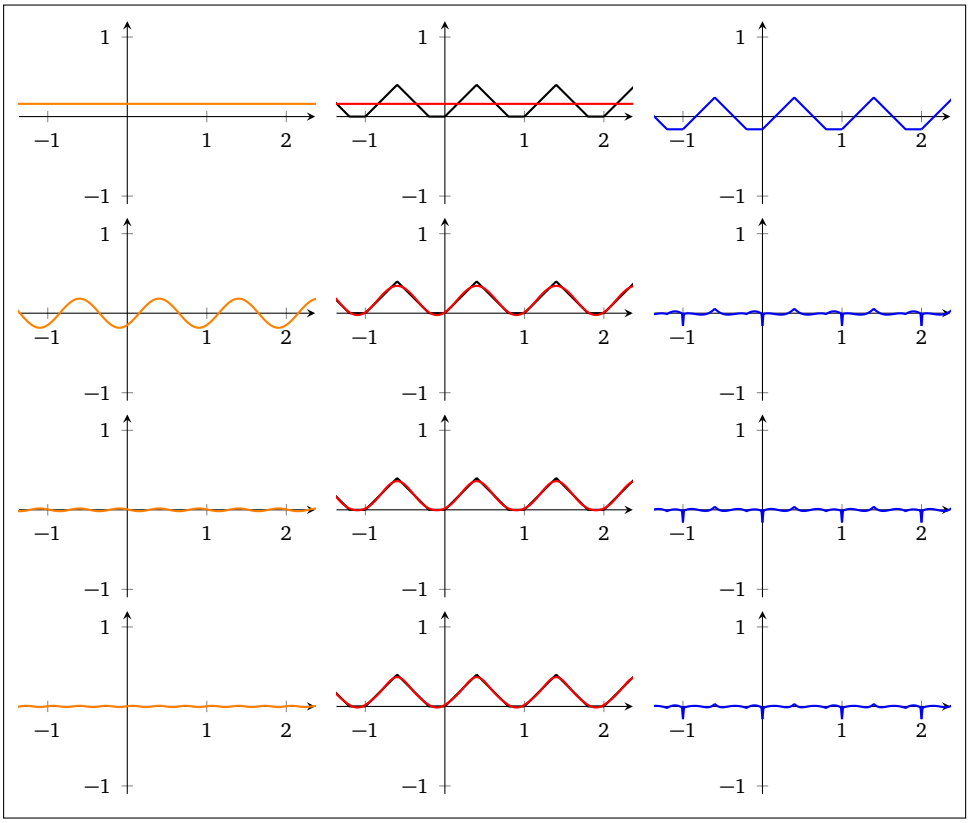


Figure 44. The convergence of the Fourier series of $\chi * \chi$ from Example 4.8. (The explanation of the plots is as in Figure 41.)

Example 4.8 (Fourier coefficients of a triangular wave)**.** Let $\chi\colon \mathbb{R} \to \mathbb{C}$ denote the function from the previous example. We know from Proposition 4.6 that for $k \neq 0$

$$\widehat{(\chi * \chi)}(k) = \hat{\chi}(k)^2 = -\frac{(e^{-2\pi i k c} - 1)^2}{(2\pi k)^2} \quad \text{and} \quad \widehat{(\chi * \chi)}(0) = \hat{\chi}(0)^2 = c^2.$$

We now wish to compute $\chi * \chi$. We have

$$(\chi * \chi)(x) = \int_0^1 \chi(\xi)\chi(x-\xi)\,\mathrm{d}\xi = \int_0^c \chi(x-\xi)\,\mathrm{d}\xi.$$

One sees that this integral gives the "measure" of the set

$$[0,c] \cap \bigcup_{u \in \mathbb{Z}} (x + u + (-c, 0]).$$

A bit of thinking then shows that we have

$$(\chi * \chi)(x) = \max\{0, 2c-1, c - |x-c|\}.$$

4.4. Fourier series in higher dimensions

In this section, we give a quick glimpse at multi-dimensional Fourier series. Moreover, we shall use linear algebra to make our results (which we like to formulate for 1-periodic functions) applicable to functions which are periodic with respect to more general period vectors.

4.4.1. Periodic functions on $[0,1]^n$. Suppose that $f\colon \mathbb{R}^n \to \mathbb{C}$ is continuously differentiable[4] and periodic with respect to the standard unit vectors, i.e.,

$$f(\vec{x} + \lambda_1\vec{e}_1 + \ldots + \lambda_n\vec{e}_n) = f(\vec{x})$$

for every $\vec{x} \in \mathbb{R}^n$ and all *integers(!)* $\lambda_1, \ldots, \lambda_n$. Then we define the ***Fourier coefficients*** of f to be

$$\hat{f}(\vec{k}) := \int_{[0,1]^n} f(\vec{x}) e^{-2\pi i \vec{k}\cdot\vec{x}}\, \mathrm{d}^n\vec{x},$$

where $\vec{k} \in \mathbb{Z}^n$. (The precise definition of such integrals will be given later in § 7.1 of Chapter 7.) Then we have the following analogue of Corollary 4.2:

$$f(\vec{x}) = \sum_{\vec{k}}{}^{\natural} \hat{f}(\vec{k}) e^{2\pi i \vec{k}\cdot\vec{x}}, \quad \text{where} \quad \sum_{\vec{k}}{}^{\natural} := \lim_{K\to\infty} \sum_{\substack{\vec{k}\in\mathbb{Z}^n \\ |k_1|+\ldots+|k_n|_1 \le K}} .$$

Observe that, much akin to our discussion in § 4.1, for each non-zero $\vec{k} \in \mathbb{Z}^n$, the function

$$b_{\vec{k}}\colon \mathbb{R}^n \to \mathbb{R}^n, \quad \vec{x} \mapsto e^{2\pi i \vec{k}\cdot\vec{x}},$$

is a 1-periodic eigenfunction of *each* of the following differential operators:

$$\frac{\partial^2}{\partial x_i^2} \quad (i = 1, \ldots, n).$$

Indeed,

$$\frac{\partial^2}{\partial x_i^2} b_{\vec{k}}(\vec{x}) = (2\pi i k_i)^2 e^{2\pi i \vec{k}\cdot\vec{x}} = -4\pi^2 k_i^2 \cdot b_{\vec{k}}(\vec{x}).$$

More information on partial derivatives such as the ones considered above is given in Chapter 5. Interestingly, the fact that one should *hope* to find such functions which are eigenfunctions to all of the above operators *simultaneously* is related to a commutativity property of partial derivatives known as *Schwarz's theorem* (stated as Theorem 6.1 below).

[4] We shall review the fundamentals of higher-dimensional differentiation later in Chapter 5.

4.4.2. Dealing with more general periods. In practice (for instance, in solid state physics when analysing functions defined on crystals) one also has to deal with functions $f\colon \mathbb{R}^n \to \mathbb{C}$ whose "period vectors" are not the standard unit vectors. Let us assume that we have vectors $\vec{a}_1, \ldots, \vec{a}_n \in \mathbb{R}^n$ and f satisfies

$$f(\vec{x} + \lambda_1\vec{a}_1 + \ldots + \lambda_n\vec{a}_n) = f(\vec{x})$$

for every $\vec{x} \in \mathbb{R}^n$ and all integers $\lambda_1, \ldots, \lambda_n$. Assume further that[5]

$$\det A \neq 0, \quad \text{where} \quad A \coloneqq \begin{pmatrix} | & \ldots & | \\ \vec{a}_1 & \ldots & \vec{a}_n \\ | & \ldots & | \end{pmatrix} \in \mathbb{R}^{n\times n}.$$

(Recall that this means that the vectors $\vec{a}_1, \ldots, \vec{a}_n$ span a ***parallelotope***[6] with non-zero volume, or—equivalently—that the vectors $\vec{a}_1, \ldots, \vec{a}_n$ do not all lie in some hyperplane and genuinely span the whole space $\mathbb{R}^n$.)

We now define a function $f_\square\colon \mathbb{R}^n \to \mathbb{R}^n$ via

$$f_\square(\vec{x}) \coloneqq f(A\vec{x}).$$

This new function is now periodic with respect to the standard unit vectors and, intuitively speaking, behaves on the box

$$\{\, \lambda_1\vec{e}_1 + \ldots + \lambda_n\vec{e}_n : 0 \leq \lambda_1, \ldots, \lambda_n < 1 \,\}$$

like f behaves on the parallelotope

$$\{\, \lambda_1\vec{a}_1 + \ldots + \lambda_n\vec{a}_n : 0 \leq \lambda_1, \ldots, \lambda_n < 1 \,\}.$$

Indeed, we have

$$\begin{aligned} f_\square(\lambda_1\vec{e}_1 + \ldots + \lambda_n\vec{e}_n) &= f(A(\lambda_1\vec{e}_1 + \ldots + \lambda_n\vec{e}_n)) \\ &= f(\lambda_1 A\vec{e}_1 + \ldots + \lambda_n A\vec{e}_n) \\ &= f(\lambda_1\vec{a}_1 + \ldots + \lambda_n\vec{a}_n). \end{aligned}$$

If f is continuously differentiable, then so is $f_\square$. This is not difficult to show, but we shall not attempt it, as it would distract us. (Also we review differentiability only later in Chapter 5. The reader should just get the following message here: "if f is a nice function, then so is $f_\square$".) Now we have

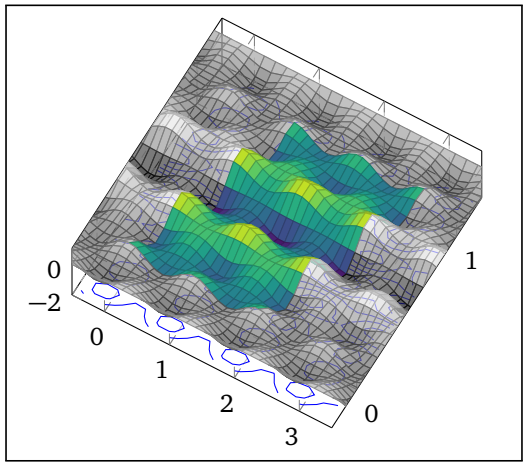
$$f_\square(\vec{y}) = \sum_{\vec{k}}{}^{\!\prime} \hat{f}_\square(\vec{k}) e^{2\pi i \vec{k}\cdot\vec{y}}$$

for every $\vec{y} \in \mathbb{R}^n$. Using this with $\vec{y} = A^{-1}\vec{x}$ for some $\vec{x} \in \mathbb{R}^n$ we find that

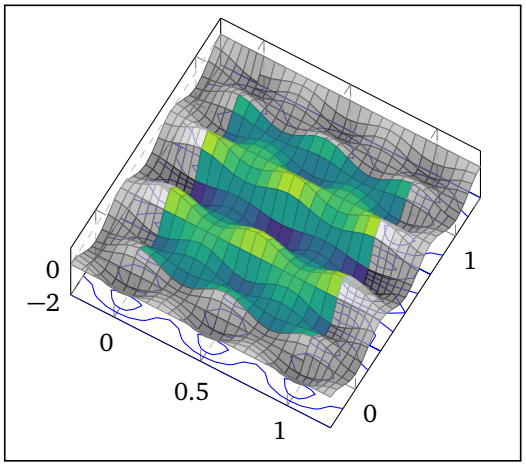
$$f(\vec{x}) = f(AA^{-1}\vec{x}) = f_\square(A^{-1}\vec{x}) = \sum_{\vec{k}}{}^{\!\prime} \hat{f}_\square(\vec{k}) e^{2\pi i \vec{k}\cdot(A^{-1}\vec{x})}. \tag{4.5}$$

[5]Technically, we have only discussed determinants in dimensions $n = 2, 3$, so the readers are welcome to specialise to these cases should they so desire.

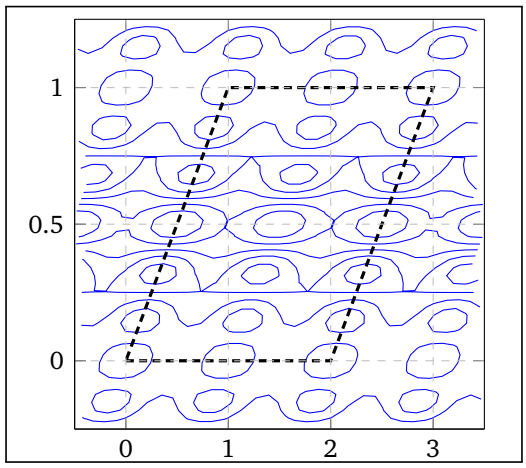
[6]For $n = 2$ this is a parallelogram, for $n = 3$ a parallelepiped. For arbitrary n one speaks of parallelotopes.



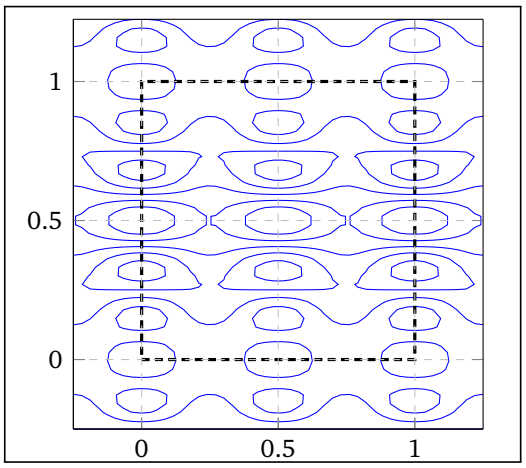
(a) Plot of f.



(b) Plot of $f_\square$.



(c) Contour plot of f.



(d) Contour plot of $f_\square$.

Figure 45. A periodic function $f\colon \mathbb{R}^2 \to \mathbb{R}$ (left) with period vectors $\vec{a}_1 = (2,0)$ and $\vec{a}_2 = (1,1)$ and the corresponding function $f_\square$ (right). The bottom row shows a contour plot the functions (at least approximately).

We have

$$\hat{f}_\square(\vec{k}) = \int_{[0,1]^n} f_\square(\vec{y}) e^{-2\pi \mathrm{i}\vec{k}\cdot\vec{y}} \,\mathrm{d}^n\vec{y} = \int_{[0,1]^n} f(A\vec{y}) e^{-2\pi\mathrm{i}\vec{k}\cdot(A^{-1}A\vec{y})} \,\mathrm{d}^n\vec{y}.$$

Using the transformation formula,[7] we arrive at

$$\begin{aligned}\hat{f}_\square(\vec{k}) &= \int_{A\cdot[0,1]^n} f(\vec{x}) e^{-2\pi \mathrm{i}\,\vec{k}\cdot(A^{-1}\vec{x})} (\det A^{-1})\,\mathrm{d}^n\vec{x} \\ &= \frac{1}{\det A}\int_{A\cdot[0,1]^n} f(\vec{x}) e^{-2\pi \mathrm{i}\,((A^{-1})^\intercal \vec{k})\cdot\vec{x}}\,\mathrm{d}^n\vec{x}.\end{aligned}$$

Plugging this into (4.5), this show that

$$f(\vec{x}) = \sum_{\vec{k}}{}^{\!\natural} \left(\frac{1}{\det A}\int_{A\cdot[0,1]^n} f(\vec{\xi}) e^{-2\pi \mathrm{i}\,((A^{-1})^\intercal \vec{k})\cdot\vec{\xi}}\,\mathrm{d}^n\vec{\xi}\right) e^{2\pi \mathrm{i}\,((A^{-1})^\intercal \vec{k})\cdot\vec{x}}.$$

(Here we have replaced the integration variable $\vec{x}$ by $\vec{\xi}$, because the variable $\vec{x}$ is already taken.) This leads us to define

$$\hat{f}(\vec{r}) := \frac{1}{\det A}\int_{A\cdot[0,1]^n} f(\vec{x}) e^{-2\pi \mathrm{i}\,\vec{r}\cdot\vec{x}}\,\mathrm{d}^n\vec{x}$$

for $\vec{r} \in (A^{-1})^\intercal \mathbb{Z}^n$. Then

$$f(\vec{x}) = \sum_{\vec{k}:\ \vec{r}=(A^{-1})^\intercal\vec{k}}{}^{\!\natural} \hat{f}(\vec{r}) e^{2\pi \mathrm{i}\,\vec{r}\cdot\vec{x}}.$$

A physicist may write the above as

$$f(\vec{x}) = \sum_{\vec{r}} \hat{f}(\vec{r}) e^{2\pi \mathrm{i}\,\vec{r}\cdot\vec{x}}$$

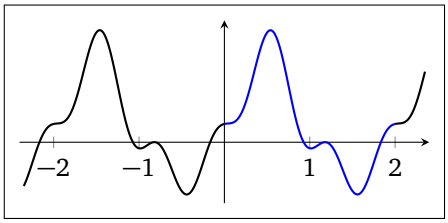
together with some accompanying words amounting to saying that the sum is to be taken over $\vec{r} = (A^{-1})^\intercal\vec{k}$ where $\vec{k}$ ranges over $\mathbb{Z}^n$. The readers should be familiar with such expressions from solid state physics. In this context, the formula in Theorem 3.13 is often used (after transposing!) to compute $(A^{-1})^\intercal$.

4.4.3. Examples. We shall discuss two examples of the above reduction technique. First, we generalise our treatment of 1-periodic functions to T-periodic functions in Example 4.9. Then, in Example 4.10, we illustrate how to deal with periodic functions on $\mathbb{R}^2$ whose period vectors are different from the standard unit vectors.

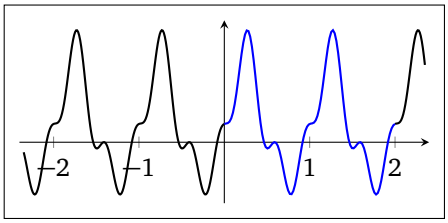
Example 4.9 (T-periodic functions)**.** We shall treat the example of a T-periodic, continuously differentiable function $f\colon \mathbb{R} \to \mathbb{C}$. Here T may denote any positive real number. The matrix A from above is then a 1×1-matrix (i.e., a number), namely $A = (T)$. We have

$$f_\square(x) = f(Tx).$$

[7] This will be discussed later; see Theorem 7.2.



(a) A 2-periodic function f.



(b) The 1-periodic function $f_\square$.

Figure 46. Illustration of the process of passing from f to $f_\square$ as in Example 4.9.

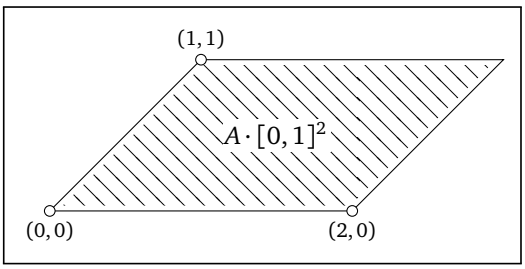


Figure 47. The parallelogram $A \cdot [0,1]^2$ from Example 4.10.

Moreover, $\det A = T$ and we have

$$\hat{f}(r) = \frac{1}{T}\int_0^T f(x)e^{-2\pi i r x}\,\mathrm{d}x,$$

as well as

$$f(x) = \sum_r{}^{\!\prime} \hat{f}(r)e^{2\pi i r x},$$

where r runs over the numbers $r = (A^{-1})^\intercal k = k/T$, $k \in \mathbb{Z}$. Hence,

$$f(x) = \lim_{K\to\infty} \sum_{k=-K}^{K} \left(\frac{1}{T}\int_0^T f(\xi)e^{-2\pi \mathrm{i}(k/T)\xi}\,\mathrm{d}\xi \right) e^{2\pi \mathrm{i}(k/T)x}.$$

Example 4.10. We shall treat the example where a continuously differentiable function $f\colon \mathbb{R}^2 \to \mathbb{C}$ is periodic with respect to the period vectors $\vec{a}_1 = \binom{2}{0}$ and $\vec{a}_2 = \binom{1}{1}$. Then

$$A = \begin{pmatrix} 2 & 1 \\ 0 & 1 \end{pmatrix}, \quad A^\intercal = \begin{pmatrix} 2 & 0 \\ 1 & 1 \end{pmatrix}, \quad (A^\intercal)^{-1} = \frac{1}{\det A^\intercal}\begin{pmatrix} 1 & 0 \\ -1 & 2 \end{pmatrix} = \begin{pmatrix} 1/2 & 0 \\ -1/2 & 1 \end{pmatrix}.$$

(For the computation of $(A^\intercal)^{-1}$ recall Proposition 3.2.) Then, for $\vec{r} = (A^{-1})^\intercal \vec{k}$ we have

$$\vec{r}\cdot\vec{x} = \begin{pmatrix} k_1/2 \\ -k_1/2 + k_2 \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \tfrac{1}{2}k_1x_1 + (k_2 - \tfrac{1}{2}k_1)x_2.$$

Hence,

$$f(\vec{x}) = \sum_{\vec{k}}{}' \left(\frac{1}{2} \int_{A\cdot[0,1]^2} f(\vec{\xi}) e^{-2\pi \mathrm{i}(\frac{1}{2}k_1\xi_1 + (k_2 - \frac{1}{2}k_1)\xi_2)} \, \mathrm{d}^2\vec{\xi} \right) e^{2\pi \mathrm{i}(\frac{1}{2}k_1 x_1 + (k_2 - \frac{1}{2}k_1)x_2)}.$$

In this case

$$A\cdot[0,1]^2 = \left\{ \begin{pmatrix} 2\lambda_1 + \lambda_2 \\ \lambda_2 \end{pmatrix} : 0 \le \lambda_1, \lambda_2 \le 1 \right\}$$

(a parallelogram; see Figure 47).

CHAPTER 5

Differentiability

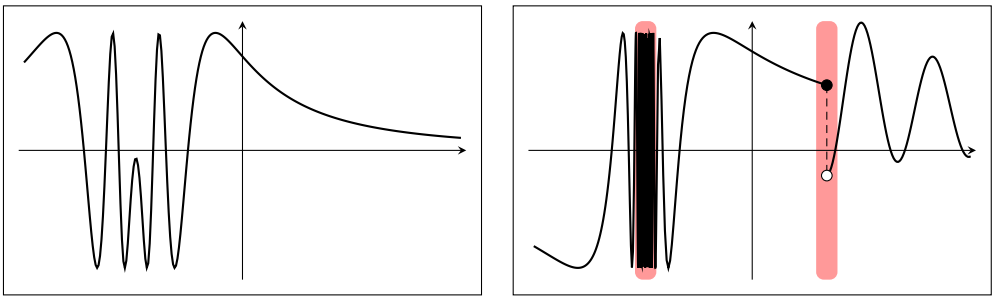
A fundamental concept in the analysis of functions is *differentiation*. Differentiation amounts to *linearisation* and this transports one into the realm of linear algebra; the latter being more easy to navigate in. In this chapter we review the basic concepts of this linearisation process. The importance of differentiation for physical phenomena should be evident from the ubiquity of differential equations in physics. It turns out, however, that a thorough understanding of differentiation is also crucial for integration. This will show up in the next chapter and the process of integrating a function ultimately relies on a quite similar process: approximating it with something like a linear function whose integral is easy to define (and grasp geometrically!) and let the general integral be the limit attained by this procedure.

5.1. Total differential

5.1.1. Differentiability at a point. We now review the basics of higher-dimensional differentiability. Let $f\colon U \to \mathbb{R}^m$ be an arbitrary map defined on some set $U \subseteq \mathbb{R}^n$. (Later we might write $\vec{f}$ instead of f if $m \geq 2$, but we shall not do this here.) We say that f is ***continuous at a point*** $\vec{x}_0 \in U$ if

$$\lim_{\vec{x}\to\vec{x}_0} f(\vec{x}) = f(\vec{x}_0).$$

We say plainly that f is ***continuous*** if it is continuous at every point of its domain of definition U.



(a) A continuous function. **(b)** A discontinuous function.

Figure 48. Illustration of the concept of continuity.

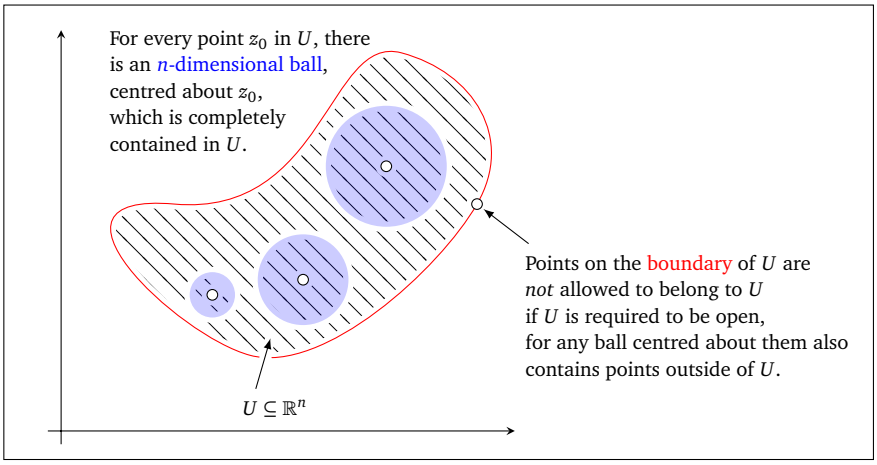


Figure 49. Illustration of the concept of an open set.

Example. Consider $f\colon \mathbb{R}^3 \to \mathbb{R}^2$, $(x_1, x_2, x_3) \mapsto (x_2, x_1^2 + x_3)$. Then

$$\lim_{\vec{x}\to(1,2,3)} f(\vec{x}) = \lim_{\vec{x}\to(1,2,3)} \begin{pmatrix} x_2 \\ x_1^2 + x_3 \end{pmatrix} = \begin{pmatrix} \lim\limits_{\vec{x}\to(1,2,3)} x_2 \\ \lim\limits_{\vec{x}\to(1,2,3)} (x_1^2 + x_3) \end{pmatrix} = \begin{pmatrix} 2 \\ 1^2 + 3 \end{pmatrix} = \begin{pmatrix} 2 \\ 4 \end{pmatrix}$$

and this value equals $f(1,2,3)$. Hence, f is continuous at $(1,2,3)$. In fact, f is continuous at every point of $\mathbb{R}^3$. (Observe here that limits of vectors are computed coordinate-wise and how limits are computed should be known [2]. The reader should conclude that in practice, checking for continuity is *usually* quite easy.)

If $\vec{x}_0 \in U$ is such that U contains some n-dimensional ball $B(\vec{x}_0, \epsilon) = \{\, \vec{y} \in \mathbb{R}^n : \|\vec{y} - \vec{x}_0\| < \epsilon \,\}$ for some $\epsilon > 0$, then we can ask differentiability of f at $\vec{x}_0$. Such a point $\vec{x}_0 \in U$ is called an ***inner point*** of U and sets U which consist only of inner points are called ***open***.

Example. The so-called open ball $\{\, \vec{y} \in \mathbb{R}^3 : \|\vec{y}\| < 1 \,\}$ is open in the sense just defined. On the contrary, the "closed ball" $\{\, \vec{y} \in \mathbb{R}^3 : \|\vec{y}\| \leq 1 \,\}$ is not open, because the point $\vec{y} = (1,0,0)$ is contained in it, yet any ϵ-ball about $\vec{y}$ contains the point $(1 + \epsilon/2, 0, 0)$, which is not contained in our closed ball.

If $\vec{x}_0$ is some inner point of U, then f is called ***differentiable at $\vec{x}_0$*** if there is some linear map $\mathrm{d}f_{\vec{x}_0}\colon \mathbb{R}^n \to \mathbb{R}^m$ such that

$$f(\vec{x}_0 + \underbrace{\vec{v}}_{\text{distortion}}) \approx \overbrace{f(\vec{x}_0)}^{\text{point evaluation}} + \underbrace{\mathrm{d}f_{\vec{x}_0}(\vec{v})}_{\substack{\text{linearisation evaluated} \\ \text{at distortion } \vec{v}}} \tag{5.1}$$

for any $\vec{v}$ sufficiently close to $\vec{0}$ (and such that $\vec{x}_0 + \vec{v} \in U$). By general policy of this lecture, we tend to ignore most issues that arise with approximation errors.

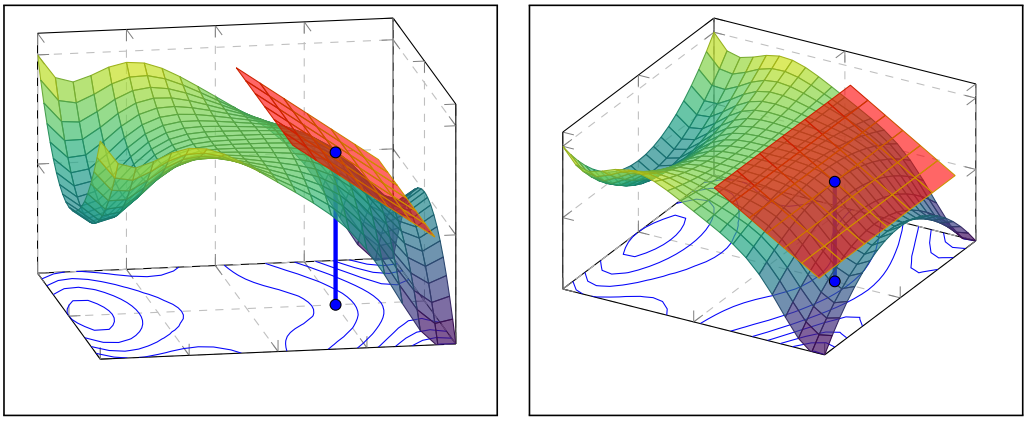


Figure 50. A real-valued function f defined on the unit disk in $\mathbb{R}^2$ and its linearisation $\vec{x} \mapsto f(\vec{x}_0) + \mathrm{d}f_{\vec{x}_0}(\vec{x} - \vec{x}_0)$ (red) at a point $\vec{x}_0$ (black). Observe that here we do include the term $f(\vec{x}_0)$ as to have the graph of the linearisation "attached" to the graph of f at $f(\vec{x}_0)$ above the point $\vec{x}_0$. Usually, though, when speaking of the linearisation of f at $\vec{x}_0$, we mean $\mathrm{d}f_{\vec{x}_0}$.

Nevertheless, here we intend to at least sketch how '$\approx$' in (5.1) is made precise. To this end, consider the difference of both sides of (5.1), namely

$$r(\vec{v}) := f(\vec{x}_0 + \vec{v}) - f(\vec{x}_0) - \mathrm{d}f_{\vec{x}_0}(\vec{v}). \tag{5.2}$$

Now '$\approx$' in (5.1) should mean that $r(\vec{v})$ goes to zero more quickly than linearly, as $\vec{v}$ approaches $\vec{0}$. Precisely, we postulate that

$$\lim_{\substack{\vec{v}\to\vec{0} \\ \vec{v}\neq\vec{0} \\ \vec{x}_0+\vec{v}\in U}} \frac{\|r(\vec{v})\|}{\|\vec{v}\|} \stackrel{!}{=} 0. \tag{5.3}$$

(One requires $\vec{x}_0$ to be an inner point of U, because otherwise the approximation (5.1) would in general be too weak a condition to ensure the validity of various theorems.)

The map $f : U \to \mathbb{R}^m$ is called ***differentiable*** if U is open and f is differentiable at every point of U. One views $\mathrm{d}f_{\vec{x}_0}$ as the "linearisation of f at $\vec{x}_0$." It can be shown that this linearisation is unique (if it exists at all). Since $\mathrm{d}f_{\vec{x}_0} : \mathbb{R}^n \to \mathbb{R}^m$ (called the ***(total) differential of f at $\vec{x}_0$***) is a linear map, it can be represented by a matrix $J_f(\vec{x}_0) \in \mathbb{R}^{m\times n}$ called the ***Jacobian matrix of f at $\vec{x}_0$***.

Example. In the one-dimensional setting (set $n = m = 1$ in the above), we have the well-known derivative:

$$f'(x_0) = \lim_{\substack{h\to 0 \\ h\neq 0}} \frac{f(x_0 + h) - f(x_0)}{h}.$$

In particular,

$$\lim_{\substack{h\to 0\\ h\neq 0}} \frac{f(x_0+h)-f(x_0)-f'(x_0)h}{h} = 0,$$

so that

$$f(x_0+h) \approx f(x_0)+f'(x_0)h$$

with '$\approx$' having the aforementioned meaning. A comparison with (5.1) (and appealing to the aforementioned uniqueness of the total differential of a function) shows that $h \mapsto f'(x_0)h$ coincides with $\mathrm{d}f_{x_0}$. The matrix representing this linear map from $\mathbb{R}^1 \to \mathbb{R}^1$ is the 1×1-matrix $J_f(x_0) = f'(x_0)$.

5.1.2. Directional and partial derivatives. Before computing the Jacobian matrix, we remind the reader of another way of thinking of differentiability. Indeed, taking any vector $\vec{v} \in \mathbb{R}^n$ with $\|\vec{v}\| = 1$, called a ***direction***, we construct a function[1] $f_{\vec{x}_0,\vec{v}}\colon (-\epsilon,\epsilon) \to \mathbb{R}^m$ of a *single* variable from f via

$$f_{\vec{x}_0,\vec{v}}(t) \coloneqq f(\vec{x}_0 + t\vec{v}).$$

We define[2]

$$\partial_{\vec{v}} f(\vec{x}_0) \coloneqq \frac{\partial f}{\partial \vec{v}}(\vec{x}_0) \coloneqq f'_{\vec{x}_0,\vec{v}}(0) \coloneqq \lim_{\substack{t\to 0\\ t\neq 0}} \frac{f_{\vec{x}_0,\vec{v}}(t)-f_{\vec{x}_0,\vec{v}}(0)}{t} \tag{5.4}$$

if the limit on the right hand side exists and call this the ***directional derivative*** of f at $\vec{x}_0$ with respect to the direction $\vec{v}$. Moreover, we reserve the following special notation for directional derivatives with respect to coordinate functions:

$$\partial_1 f \coloneqq \frac{\partial f}{\partial \vec{e}_1}, \quad \partial_2 f \coloneqq \frac{\partial f}{\partial \vec{e}_2}, \quad \ldots \quad \partial_n f \coloneqq \frac{\partial f}{\partial \vec{e}_n}.$$

These are called ***partial derivatives*** of f (with respect to the first, second, … n-th variable). If one agrees that the arguments of f are denoted by $x_1, x_2, \ldots$, then one also writes

$$\frac{\partial f}{\partial x_1} \coloneqq \partial_1 f, \quad \frac{\partial f}{\partial x_2} \coloneqq \partial_2 f, \quad \ldots,$$

and one defines similar notation when the variables of f are denoted by $x, y, \ldots$, getting $\frac{\partial f}{\partial x}$, $\frac{\partial f}{\partial y}$ etc. In this situation one often also writes f_x, f_y and f_z for these derivatives.

[1] Here $\epsilon > 0$ ought to be chosen small enough so that $\vec{x}_0 + t\vec{v}$ belongs to U for every t in the interval $(-\epsilon,\epsilon)$. This ensures that f may be evaluated at each point $\vec{x}_0 + t\vec{v}$.

[2] Note that the quantities involved (after evaluating f) are *vectors* in $\mathbb{R}^m$. The limit is supposed to be taken component-wise, e.g., $\lim_{t\to 0}(t, t+1) = (\lim_{t\to 0} t, \lim_{t\to 0}(t+1)) = (0,1)$.

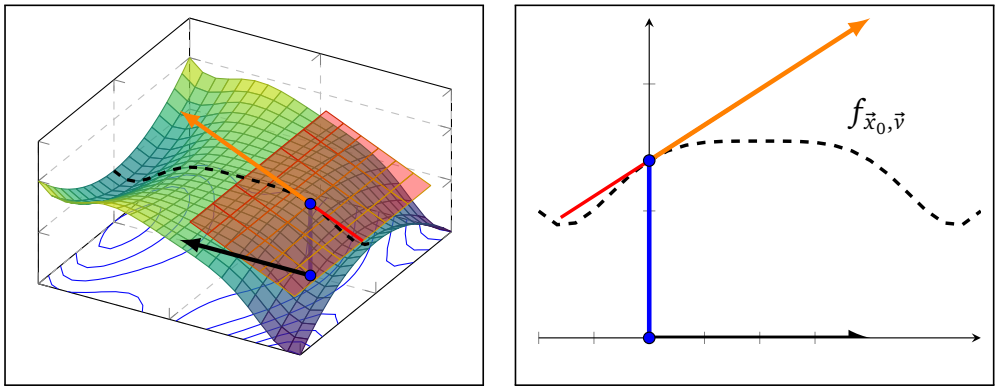


Figure 51. Illustration of the directional derivative $\partial_{\vec{v}} f(\vec{x}_0)$ of the function f from Figure 50 in the direction $\vec{v} = (-1, 0)$. The dashed curve on the left if f restricted to the line $\vec{x}_0 + t\vec{v}$, $t \in (-\epsilon, \epsilon)$; compare the definition of $f_{\vec{x}_0,\vec{v}}$.

Example. Consider the function $f : \mathbb{R}^2 \to \mathbb{R}$, $(x_1, x_2) \mapsto x_1^3 x_2^2$. We wish to calculate its directional derivatives with respect to $\vec{v} = \vec{e}_1 = (1, 0)$ at the point $\vec{x}_0 = (x_1, x_2)$. We have

$$f_{\vec{x}_0,\vec{v}}(t) := f((x_1, x_2) + t(1, 0)) = f(x_1 + t, x_2) = (x_1 + t)^3 x_2^2.$$

Using the chain rule, we find that

$$\frac{\partial f}{\partial x_1}(x_1, x_2) = \partial_1 f(x_1, x_2) = \frac{\partial f}{\partial \vec{e}_1}(x_1, x_2) = f'_{\vec{x}_0,\vec{v}}(0) = 3x_1^2 x_2^2.$$

Observe that this is exactly the result we get from considering the expression $x_1^3 x_2^2$ as a function of x_1 (with x_2 being considered a constant) and taking its derivative with respect to x_1.
Similarly, we now wish to find the directional derivative of f at $\vec{x}_0$ with respect to $\vec{w} = \frac{1}{\sqrt{2}}(1, 1)$. Then

$$f_{\vec{x}_0,\vec{w}}(t) := f((x_1, x_2) + t\tfrac{1}{\sqrt{2}}(1, 1)) = f(x_1 + t\tfrac{1}{\sqrt{2}}, x_2 + t\tfrac{1}{\sqrt{2}}) = (x_1 + t\tfrac{1}{\sqrt{2}})^3 (x_2 + t\tfrac{1}{\sqrt{2}})^2.$$

Therefore, by the product and chain rule,

$$\frac{\partial f}{\partial \vec{w}}(x_1, x_2) = f'_{\vec{x}_0,\vec{w}}(0) = 3x_1^2 \tfrac{1}{\sqrt{2}} x_2^2 + 2x_1^3 x_2 \tfrac{1}{\sqrt{2}}.$$

We leave it as an exercise for the reader to also compute $\frac{\partial f}{\partial \vec{e}_2}(x_1, x_2)$ and verify that

$$\frac{\partial f}{\partial \vec{w}} = \frac{1}{\sqrt{2}} \frac{\partial f}{\partial \vec{e}_1} + \frac{1}{\sqrt{2}} \frac{\partial f}{\partial \vec{e}_2}.$$

The last equation may be seen as a consequence of Lemma 5.1 below and the identity $\vec{w} = \frac{1}{\sqrt{2}}\vec{e}_1 + \frac{1}{\sqrt{2}}\vec{e}_2$ (see also Proposition 5.2).

If f is differentiable at $\vec{x}_0$, then one can substitute (5.1) into (5.4), getting the following result:

Lemma 5.1. *Under the above hypotheses,* $\partial_{\vec{v}} f(\vec{x}_0) = \mathrm{d}f_{\vec{x}_0}(\vec{v})$.

Proof. Let $f\colon U \to \mathbb{R}^m$ be differentiable at $\vec{x}_0 \in U \subseteq \mathbb{R}^n$. By (5.4) and the definition of $f_{\vec{x}_0,\vec{v}}$, we have

$$\partial_{\vec{v}} f(\vec{x}_0) = \lim_{\substack{t\to 0\\ t\neq 0}} \frac{f(\vec{x}_0 + t\vec{v}) - f(\vec{x}_0)}{t}.$$

In view of (5.1) and using the notation $r(\vec{v})$ from (5.2), we may rewrite this as

$$\partial_{\vec{v}} f(\vec{x}_0) = \lim_{\substack{t\to 0\\ t\neq 0}} \frac{f(\vec{x}_0) + \mathrm{d}f_{\vec{x}_0}(t\vec{v}) + r(t\vec{v}) - f(\vec{x}_0)}{t} = \lim_{\substack{t\to 0\\ t\neq 0}} \frac{\mathrm{d}f_{\vec{x}_0}(t\vec{v})}{t} + \lim_{\substack{t\to 0\\ t\neq 0}} \frac{r(t\vec{v})}{t}.$$

By linearity, $\mathrm{d}f_{\vec{x}_0}(t\vec{v}) = t\,\mathrm{d}f_{\vec{x}_0}(\vec{v})$, so the first limit on the right hand side of the above equation turns out to just be $\mathrm{d}f_{\vec{x}_0}(\vec{v})$. Moreover, the second limit turns out to be the zero vector (in $\mathbb{R}^m$): indeed, by looking at the limit of the lengths of $r(t\vec{v})/t$ as $t \to \infty$, we see that this limit of lengths must be zero by (5.3):

$$\lim_{\substack{t\to 0\\ t\neq 0}} \left\| \frac{r(t\vec{v})}{t} \right\| = \lim_{\substack{t\to 0\\ t\neq 0}} \frac{\|r(t\vec{v})\|}{|t|} = \lim_{\substack{t\to 0\\ t\neq 0}} \frac{\|r(t\vec{v})\|}{\|t\vec{v}\|} = \lim_{\substack{\vec{w}\to\vec{0}\\ \vec{w}\neq\vec{0}}} \frac{\|r(\vec{w})\|}{\|\vec{w}\|} = 0.$$

Consequently, we have $\partial_{\vec{v}} f(\vec{x}_0) = \mathrm{d}f_{\vec{x}_0}(\vec{v})$, as claimed in the lemma. □

Now clearly the columns of the Jacobian matrix $J_f(\vec{x}_0)$ are given by

$$\mathrm{d}f_{\vec{x}_0}(\vec{e}_1),\ \ldots,\ \mathrm{d}f_{\vec{x}_0}(\vec{e}_n).$$

From this and Lemma 5.1 it follows that

$$J_f(\vec{x}_0) = \begin{pmatrix} | & \cdots & | \\ \partial_1 f(\vec{x}_0) & \cdots & \partial_n f(\vec{x}_0) \\ | & \cdots & | \end{pmatrix}.$$

Observe that this allows us to calculate J_f (and, thus, the differential $\mathrm{d}f$), because in all practical situations, where f is given by some formula, the partial derivatives $\partial_i f$ can be calculated in the same fashion in which one has learned to calculate the derivative of a function from $\mathbb{R}$ to $\mathbb{R}$.

5.1.3. Criterion for differentiability. To show that a function $f\colon U \to \mathbb{R}^m$ is differentiable at $\vec{x}_0$ one often uses the following criterion:

Proposition 5.2. *Let* $\vec{x}_0$ *be an inner point of some set* $U \subseteq \mathbb{R}^n$ *and* $f\colon U \to \mathbb{R}^m$ *be some function. Suppose that all the partial derivatives of* f *exist in an open ball* $B(\vec{x}_0, \epsilon) = \{\, \vec{x} \in \mathbb{R}^n : \|\vec{x} - \vec{x}_0\| < \epsilon \,\} \subseteq U$ *with centre* $\vec{x}_0$ *and radius* $\epsilon > 0$ *and are continuous there (when viewed as functions on the ball* $B(\vec{x}_0, \epsilon)$*). Then* f *is differentiable at* $\vec{x}_0$.

A function having continuous partial derivatives at every point of its domain of definition in the sense of Proposition 5.2 is called ***continuously differentiable*** or "***C^1-function***" for short. If the partial derivatives are not just continuous but themselves continuously differentiable at every point of the domain of definition, then we speak of the function being ***twice continuously differentiable*** or being a "***C^2-function***". Similarly, one speaks of C^k-functions ($k = 1, 2, 3, \ldots$) and a function which is C^k for every $k \geq 1$ is called ***C^∞-function*** or ***smooth***.

Example. Consider the function

$$f\colon \mathbb{R}^2 \to \mathbb{R}^4, \quad \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \mapsto \begin{pmatrix} x_2^2 \\ x_1 x_2^4 \\ 42 \\ x_1 + x_2 \end{pmatrix}.$$

We have

$$\partial_1 f(\vec{x}) = \begin{pmatrix} 0 \\ x_2^4 \\ 0 \\ 1 \end{pmatrix} \quad \text{and} \quad \partial_1 f(\vec{x}) = \begin{pmatrix} 2x_2 \\ 4x_1 x_2^3 \\ 0 \\ 1 \end{pmatrix}.$$

It is obvious that the partial derivatives of f are continuous at every point $\vec{x}$ (the expressions in the coordinates above are polynomials in the coordinates of $\vec{x}$ after all). Therefore, f is differentiable at every point $\vec{x} \in \mathbb{R}^2$ (use Proposition 5.2) and we have

$$J_f(\vec{x}) = \begin{pmatrix} 0 & 2x_2 \\ x_2^4 & 4x_1 x_2^3 \\ 0 & 0 \\ 1 & 1 \end{pmatrix}.$$

Plugging in $\vec{x} = (0, 1)$ for illustration, we find that

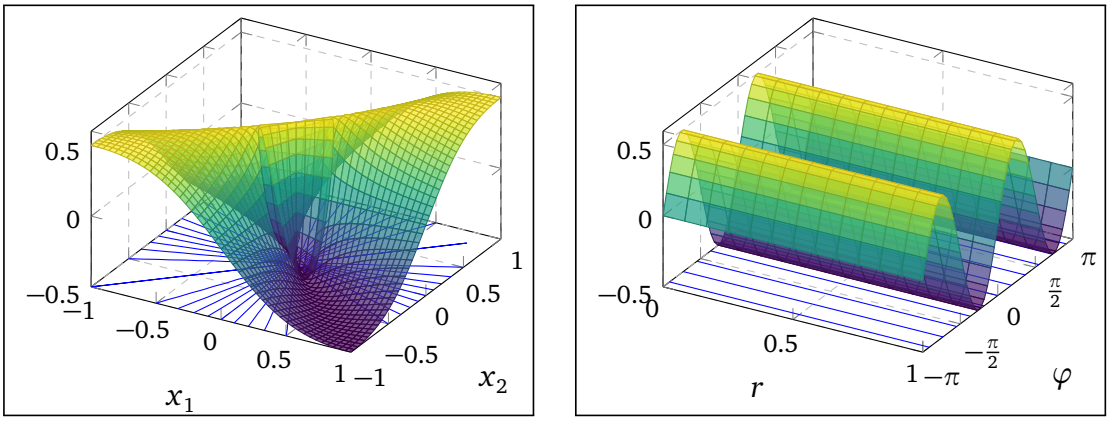
$$J_f(0,1) = \begin{pmatrix} 0 & 2 \\ 1 & 0 \\ 0 & 0 \\ 1 & 1 \end{pmatrix}$$

and $\mathrm{d}f_{(0,1)}\colon \mathbb{R}^2 \to \mathbb{R}^4$ is the linear map given by

$$\mathrm{d}f_{(0,1)}(\vec{v}) = \begin{pmatrix} 0 & 2 \\ 1 & 0 \\ 0 & 0 \\ 1 & 1 \end{pmatrix} \vec{v} = \begin{pmatrix} 2v_2 \\ v_1 \\ 0 \\ v_1 + v_2 \end{pmatrix}.$$

Example 5.3. The function

$$f\colon \mathbb{R}^2 \to \mathbb{R}, \quad (x_1, x_2) \mapsto \begin{cases} \dfrac{x_1 x_2}{x_1^2 + x_2^2} & \text{if } (x_1, x_2) \neq (0,0), \\ 0 & \text{if } x_1 = x_2 = 0 \end{cases}$$



(a) Plot of f.

(b) Plot of $f \circ \vec{P}$.

Figure 52. Plot of the function from Example 5.3 (left). On the right one can see a plot of the function f in polar coordinates. See § 6.4 and, in particular, Example 6.6.

has partial derivatives with respect to x_1 and x_2:

$$\frac{\partial f}{\partial x_1}(x_1, y_2) = \begin{cases} x_2 \dfrac{x_2^2 - x_1^2}{(x_1^2 + x_2^2)^2} & \text{if } (x_1, x_2) \neq (0,0), \\ 0 & \text{if } x_1 = x_2 = 0 \end{cases}$$

and similarly for $\frac{\partial f}{\partial x_2}$. However, neither partial derivative is continuous at $(0,0)$ and f is not differentiable at $(0,0)$. In fact, f is not even continuous there:

$$\lim_{t \to 0} f(t,t) = \lim_{t \to 0} \frac{t^2}{t^2 + t^2} = \frac{1}{2} \neq 0 = f(0,0).$$

A plot of the function f is shown in Figure 52. For a continuation of this example see Example 6.6.

To see that f in the previous example is not differentiable at $(0,0)$, note the following result:

Proposition 5.4. *If a function $f\colon U \to \mathbb{R}^m$ is differentiable on some inner point $\vec{x}_0$ of the set $U \subseteq \mathbb{R}^n$, then f is continuous at $\vec{x}_0$.*

Proof. Write

$$f(\vec{x}) = f(\vec{x}_0) + \mathrm{d}f_{\vec{x}_0}(\vec{x} - \vec{x}_0) + r(\vec{x} - \vec{x}_0)$$

with the error term r defined in (5.2). One can then check that the last two terms on the right hand side converge to zero as $\vec{x} \to \vec{x}_0$. (For r this limit is approached even more quickly than linearly—by definition of differentiability! As for $\mathrm{d}f_{\vec{x}_0}$ this is a consequence of linear maps $\mathbb{R}^n \to \mathbb{R}^m$ always being continuous—a fact which one can verify quite easily by looking at the corresponding matrix representation and

calculating in coordinates.) This implies that $f(\vec{x})$ converges to $f(\vec{x}_0)$ as $\vec{x} \to \vec{x}_0$ and this says precisely that f is continuous at $\vec{x}_0$. □

5.2. Gradient

Now suppose that $m = 1$, i.e., $f\colon U \to \mathbb{R}$ is a *real-valued* function on some set $U \subseteq \mathbb{R}^n$. Then

$$J_f(\vec{x}_0) = \begin{pmatrix} \partial_1 f(\vec{x}_0) & \dots & \partial_n f(\vec{x}_0) \end{pmatrix} \in \mathbb{R}^{1\times n}$$

is a "row-vector". Transposing this yields a (column-)vector called the ***gradient*** of f at $\vec{x}_0$,[3]

$$\operatorname{grad} f(\vec{x}_0) \coloneqq J_f(\vec{x}_0)^\intercal = \begin{pmatrix} \partial_1 f(\vec{x}_0) \\ \vdots \\ \partial_n f(\vec{x}_0) \end{pmatrix}.$$

For any $\vec{v} \in \mathbb{R}^n$ we have

$$(\operatorname{grad} f(\vec{x}_0)) \cdot \vec{v} = J_f(\vec{x}_0)\vec{v} = \mathrm{d}f_{\vec{x}_0}(\vec{v}) = \partial_{\vec{v}} f(\vec{x}_0) = f'_{\vec{x}_0,\vec{v}}(0).$$

Moreover,

$$f'_{\vec{x}_0,\vec{v}}(0) = \lim_{\substack{h\to 0 \\ h\neq 0}} \frac{f_{\vec{x}_0,\vec{v}}(h) - f_{\vec{x}_0,\vec{v}}(0)}{h} = \lim_{\substack{h\to 0 \\ h\neq 0}} \frac{f(\vec{x}_0 + h\vec{v}) - f(\vec{x}_0)}{h}.$$

Hence,

$$(\operatorname{grad} f(\vec{x}_0)) \cdot \vec{v} = \lim_{\substack{h\to 0 \\ h\neq 0}} \frac{f(\vec{x}_0 + h\vec{v}) - f(\vec{x}_0)}{h}.$$

We now ask when the right hand side becomes maximal. This corresponds to taking $\vec{v}$ to be the direction in which f increases most in the vicinity of the point $\vec{x}_0$. Using the definition of the dot product and $\|\vec{v}\| = 1$, we find that the left hand side of the above equals

$$\|\operatorname{grad} f(\vec{x}_0)\| \cos \sphericalangle(\operatorname{grad} f(\vec{x}_0), \vec{v}).$$

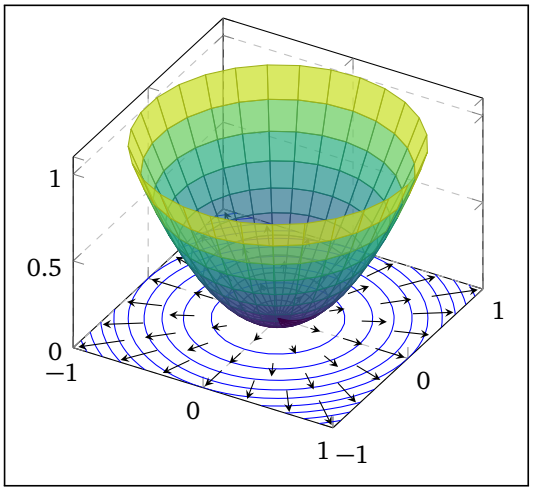
The only influence of $\vec{v}$ here is on the cosine of the angle between $\operatorname{grad} f(\vec{x}_0)$ and $\vec{v}$. The maximum value of the cosine is 1 and this is attained for an angle of 0°, i.e., when $\operatorname{grad} f(\vec{x}_0)$ and $\vec{v}$ point in the same direction.

Proposition 5.5. *The gradient* $\operatorname{grad} f(\vec{x}_0)$ *of a real-valued, differentiable function* $f\colon U \to \mathbb{R}$ *(*$U \subseteq \mathbb{R}^n$*) at* $\vec{x}_0 \in U$ *points in the direction of the steepest ascent of* f *in the vicinity of* $\vec{x}_0$.

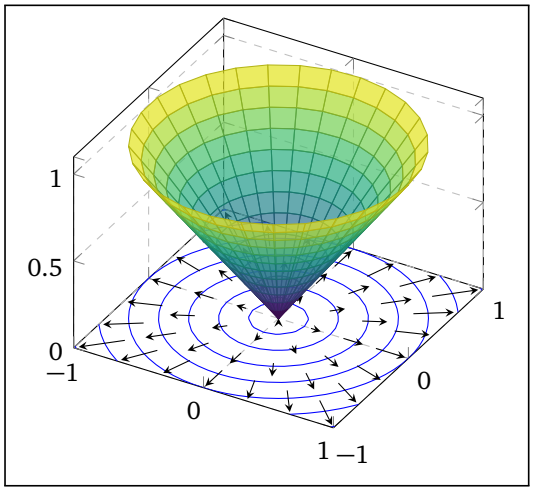
Example.

- Consider $f\colon \mathbb{R}^n \to \mathbb{R}$, $\vec{x} \mapsto \|\vec{x}\|^2 = x_1^2 + \ldots + x_n^2$. Then $\operatorname{grad} f(\vec{x}) = 2\vec{x}$ (compare Figure 53).

[3] The notation is rather clumsy here. Actually one should read $\operatorname{grad} f(\vec{x}_0)$ as $\operatorname{grad} f$ evaluated at $\vec{x}_0$, i.e., $(\operatorname{grad} f)(\vec{x}_0)$, but adding parentheses here seems to be unusual in the literature. Some authors write $\operatorname{grad}_{\vec{x}_0} f$ instead.



(a) $f\colon \mathbb{R}^2 \to \mathbb{R}$, $\vec{x} \mapsto \|\vec{x}\|^2$.



(b) $g\colon \mathbb{R}^2 \to \mathbb{R}$, $\vec{x} \mapsto \|\vec{x}\|$.

Figure 53. Illustration of Proposition 5.5. The graph of f, is an elliptic paraboloid (a parabola rotated about an axis passing perpendicularly through its extremal point); the sets on which the function stays constant are circles about the origin and the steepest increase is always straight away from the origin and indeed $\operatorname{grad}\|\vec{x}\|^2 = 2\vec{x}$ points precisely in this direction. The arrows show this gradient field.

- Consider $g\colon \mathbb{R}^n \to \mathbb{R}$, $\vec{x} \mapsto \|\vec{x}\| = \sqrt{x_1^2 + \ldots + x_n^2}$. Then $\operatorname{grad} g(\vec{x}) = \vec{x}/\|\vec{x}\|$ for every $\vec{x} \neq \vec{0}$. Moreover, g is not differentiable at $\vec{0}$.

Recall that we consider differentiable functions $f\colon \mathbb{R}^n \to \mathbb{R}$ (mapping to $\mathbb{R}^1$). If one defines "coordinate functions" $x_i\colon U \to \mathbb{R}$ for $i = 1, \ldots, n$ via

$$x_i(\vec{v}) = i\text{-th coordinate of } \vec{v},$$

then $\mathrm{d}x_{i,\vec{x}_0}$ is the linear map $\mathbb{R}^n \to \mathbb{R}$ mapping any vector $\vec{v}$ to its i-th coordinate:

$$\mathrm{d}x_{i,\vec{x}_0}\left(\begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix}\right) = v_i.$$

(Hence, the linear map $\mathrm{d}x_{i,\vec{x}_0}$ coincides with the map x_i apart from having extended the domain of definition from U to $\mathbb{R}^n$. This should not be surprising: the map x_i *is* linear[4] and *must*, therefore, coincide with its linearisation $\mathrm{d}x_i$.) One derives the cute formula

$$\mathrm{d}f_{\vec{x}_0} = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(\vec{x}_0)\,\mathrm{d}x_{i,\vec{x}_0}, \tag{5.5}$$

[4] In the sense of $x_i(\lambda\vec{v} + \mu\vec{w}) = \lambda x_i(\vec{v}) + \mu x_i(\vec{w})$ whenever $\lambda\vec{v} + \mu\vec{w}, \vec{v}, \vec{w} \in U$.

or

$$\mathrm{d}f = \frac{\partial f}{\partial x_1}\,\mathrm{d}x_1 + \ldots + \frac{\partial f}{\partial x_n}\,\mathrm{d}x_n = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}\,\mathrm{d}x_i$$

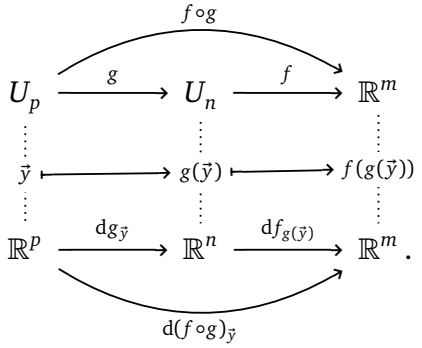
for short. One can also view (5.5) in terms of matrices. Indeed, the Jacobian matrix $J_f(\vec{x}_0)$ is an $\mathbb{R}^{1\times n}$-matrix (a row!) with entries $\frac{\partial f}{\partial x_i}(\vec{x}_0)$. The matrix representing $\mathrm{d}x_{i,\vec{x}_0}$ is also a row of the same length. Its entries are the images of the standard unit vectors under $\mathrm{d}x_{i,\vec{x}_0}$. However, $\mathrm{d}x_{i,\vec{x}_0}(\vec{e}_j) = 0$ for $j \neq i$ and $= 1$ for $j = i$. Consequently, the formula (5.5) says nothing but

$$\begin{pmatrix} \frac{\partial f}{\partial x_1}(\vec{x}_0) \\ \frac{\partial f}{\partial x_2}(\vec{x}_0) \\ \vdots \\ \frac{\partial f}{\partial x_n}(\vec{x}_0) \end{pmatrix} = \frac{\partial f}{\partial x_1}(\vec{x}_0)\begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} + \frac{\partial f}{\partial x_2}(\vec{x}_0)\begin{pmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{pmatrix} + \ldots + \frac{\partial f}{\partial x_n}(\vec{x}_0)\begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix},$$

where we have written $\mathbb{R}^{n\times 1}$-matrices (columns) instead of $\mathbb{R}^{1\times n}$-matrices (rows) for typographical reasons.

5.3. Chain rule

Let $U_p \subseteq \mathbb{R}^p$, $U_n \subseteq \mathbb{R}^n$ be two sets consisting only of inner points and consider two differentiable functions $g\colon U_p \to U_n$ and $f\colon U_n \to \mathbb{R}^m$. We now take on the mantra *"differentiation means linearisation"*. Hence, we may linearise g at $\vec{y} \in U_p$, getting the linear map $\mathrm{d}g_{\vec{y}}\colon \mathbb{R}^p \to \mathbb{R}^n$. Similarly, we may linearise f at the point $g(\vec{y}) \in U_n$, getting the linear map $\mathrm{d}f_{g(\vec{y})}\colon \mathbb{R}^n \to \mathbb{R}^m$. We can also compose f and g, getting $f \circ g$ and try to linearise this map at $\vec{y}$. What should this linearisation be?



The beautiful ***chain rule*** asserts that this linearisation is *precisely* what it ought be: the composition of the respective linearisations of f and g!

Theorem 5.6 (Chain rule)**.** *With the notation above, $f \circ g\colon U_p \to \mathbb{R}^n$ is differentiable and we have*

$$\mathrm{d}(f \circ g)_{\vec{y}} = (\mathrm{d}f_{g(\vec{y})}) \circ \mathrm{d}g_{\vec{y}},$$

or—equivalently—in terms for Jacobian matrices:

$$J_{f\circ g}(\vec{y}) = J_f(g(\vec{y}))J_g(\vec{y}).$$

Remark. For $p = n = m = 1$, each Jacobian matrix in the above is a 1×1-matrix whose single entry equals the derivative of the function of which the Jacobian matrix was being computed. Hence, in this case, the chain rule above simply states that

$$(f \circ g)'(y) = f'(g(y))\,g'(y).$$

This is just the well-known 1-dimensional chain rule!

Proving Theorem 5.6 would not be particularly hard: one needs only check that $(\mathrm{d}f_{g(\vec{y})}) \circ \mathrm{d}g_{\vec{y}}$ linearises $f \circ g$ at $\vec{y}$ in the sense of the definition of differentiability given at the start of the chapter. However, we refrain from doing this and give an example instead.

Example. Consider

$$g\colon \mathbb{R} \to \mathbb{R}^3, \quad y \mapsto \begin{pmatrix} y \\ y^2 \\ 3y-2 \end{pmatrix} \quad \text{and} \quad f\colon \mathbb{R}^3 \to \mathbb{R}, \quad \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \mapsto x_1^2 + x_2 + x_3.$$

Then $(f \circ g)(y) = f(y, y^2, 3y-2) = 2y^2 + 3y - 2$ and

$$J_{f\circ g}(y) = (f \circ g)'(y) = 4y + 3.$$

On the other hand, we have

$$J_f(\vec{x}) = \begin{pmatrix} 2x_1 & 1 & 1 \end{pmatrix} \in \mathbb{R}^{1\times3}, \quad J_g(y) = \begin{pmatrix} 1 \\ 2y \\ 3 \end{pmatrix} \in \mathbb{R}^{3\times1},$$

and

$$J_f(g(y))J_g(y) = \begin{pmatrix} 2y & 1 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ 2y \\ 3 \end{pmatrix} = (2y + 2y + 3) \in \mathbb{R}^{1\times1} = \mathbb{R}.$$

We see that the chain rule does indeed hold here.

CHAPTER 6

Topics in differentiability

This chapter offers a selection of applications and further concepts in differentiability. Amongst other things, we discuss various differential operators which bear significance for integration theory (§ 6.1), study approximation of functions by polynomials and power series (§ 6.2), present a method for solving non-linear equations numerically (§ 6.3), and consider various ways of describing points in space (§ 6.4).

6.1. Nabla, rotation and divergence

One writes formally(!)[1]

$$\nabla := \begin{pmatrix} \frac{\partial}{\partial x_1} \\ \vdots \\ \frac{\partial}{\partial x_n} \end{pmatrix} := \begin{pmatrix} \partial_1 \\ \vdots \\ \partial_n \end{pmatrix}.$$

This is called the ***nabla operator*** or ***del operator***. Then

$$\nabla f = \begin{pmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{pmatrix} = \operatorname{grad} f.$$

(In the literature one also often finds the notation $\vec{\nabla}$ which is meant to remind the reader that $\vec{\nabla} f$ is a vector-valued function. We refrain from adapting this due to laziness, although it would otherwise integrate well with our general choice of notation.) Moreover, for a vector field $\vec{F}\colon U \to \mathbb{R}^3$, $\vec{v} \mapsto (F_1(\vec{v}), F_2(\vec{v}), F_3(\vec{v}))$, on some set $U \subseteq \mathbb{R}^3$ one defines

$$\operatorname{rot} \vec{F} := \nabla \times \vec{F} := \det \begin{pmatrix} \vec{e}_1 & \frac{\partial}{\partial x_1} & F_1 \\ \vec{e}_2 & \frac{\partial}{\partial x_2} & F_2 \\ \vec{e}_3 & \frac{\partial}{\partial x_3} & F_3 \end{pmatrix} := \begin{pmatrix} \frac{\partial F_3}{\partial x_2} - \frac{\partial F_2}{\partial x_3} \\ \frac{\partial F_3}{\partial x_1} - \frac{\partial F_1}{\partial x_3} \\ \frac{\partial F_2}{\partial x_1} - \frac{\partial F_1}{\partial x_2} \end{pmatrix}.$$

[1] We assign no independent meaning to the symbols $\frac{\partial}{\partial x_i}$ or ∂_i; they only become meaningful after being "attached" to some function f.

One defines further

$$\operatorname{div}\vec{F} := \nabla\cdot\vec{F} := \begin{pmatrix} \frac{\partial}{\partial x_1} \\ \vdots \\ \frac{\partial}{\partial x_n} \end{pmatrix} \cdot \begin{pmatrix} F_1 \\ \vdots \\ F_n \end{pmatrix} := \frac{\partial F_1}{\partial x_1} + \ldots + \frac{\partial F_n}{\partial x_n}.$$

The significance of these "vector differential operators" will become apparent later when we discuss the classical integral theorems of Kelvin–Stokes and Gauß in § 7.4: these relate integration of $\operatorname{rot}\vec{F}$ (or $\operatorname{div}\vec{F}$) along some surface (or some solid) to integration over its boundary. In Chapter 8 we shall see that this together with conservation laws from physics (e.g., particles in some solid only leave said solid only through the boundary) this naturally gives rise to certain partial differential equations. These also involve the Laplace operator (defined below).

Example.

$$\begin{aligned}\operatorname{div}\begin{pmatrix} x_1^2x_2 - 4\sin(x_1) \\ 2\cos(x_1x_3) \\ 45 \end{pmatrix} &= \frac{\partial(x_1^2x_2 - 4\sin(x_1))}{\partial x_1} + \frac{\partial(2\cos(x_1x_3))}{\partial x_2} + \frac{\partial(45)}{\partial x_3} \\ &= \frac{\partial(x_1^2x_2 - 4\sin(x_1))}{\partial x_1} + 0 + 0 = 2x_1x_2 - 4\cos(x_1).\end{aligned}$$

Let $U \subseteq \mathbb{R}^3$ be some open set. We define

$$\begin{aligned} C^\infty(U) &:= \{\text{smooth functions } f\colon U \to \mathbb{R}\}, \\ \mathscr{V}(U) &:= \{\text{smooth vector fields } \vec{F}\colon U \to \mathbb{R}^3\}. \end{aligned}$$

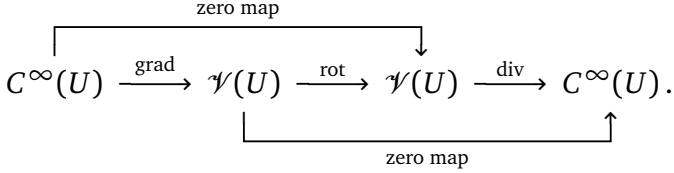
(A vector field $\vec{F}\colon U \to \mathbb{R}^3$, $\vec{x} \mapsto (F_1(\vec{x}), F_2(\vec{x}), F_3(\vec{x}))$, is called ***smooth*** if all of its component functions $F_1, F_2, F_3\colon U \to \mathbb{R}$ are smooth.) Then we obtain a sequence of maps

$$C^\infty(U) \xrightarrow{\text{grad}} \mathscr{V}(U) \xrightarrow{\text{rot}} \mathscr{V}(U) \xrightarrow{\text{div}} C^\infty(U).$$

By using Schwarz's theorem (Theorem 6.1 below) it can be checked that the composition of any of two consecutive maps is zero:

$$\begin{aligned} \operatorname{rot}\circ\operatorname{grad} &= (\text{zero map: } C^\infty(U) \longrightarrow \mathscr{V}(U)), \\ \operatorname{div}\circ\operatorname{rot} &= (\text{zero map: } \mathscr{V}(U) \longrightarrow C^\infty(U)). \end{aligned}$$

One illustrates this with the following diagram:



Theorem 6.1 (Schwarz). *If $f\colon U \to \mathbb{R}$ is a C^2-function on some open set $U \subseteq \mathbb{R}^n$, then, for any $1 \leq i, j \leq n$,*

$$\frac{\partial}{\partial x_i}\left(\frac{\partial}{\partial x_j} f\right) = \frac{\partial}{\partial x_j}\left(\frac{\partial}{\partial x_i} f\right).$$

(Similarly, one may rearrange up to k partial derivatives if f is a C^k function.)

Example. We consider the function $f\colon \mathbb{R}^3 \to \mathbb{R}$, $(x_1, x_2, x_3) \mapsto x_1^2 x_2 - \cos(x_3)$. Then

$$\operatorname{grad} f(\vec{x}) = \begin{pmatrix} 2x_1x_2 \\ x_1^2 \\ \sin(x_3) \end{pmatrix}.$$

Consequently, we find that

$$\operatorname{rot}\operatorname{grad} f(\vec{x}) = \begin{pmatrix} 0-0 \\ 0-0 \\ 2x_1 - 2x_1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} = \vec{0}.$$

Furthermore, we have

$$\operatorname{div}\operatorname{grad} f = \operatorname{div}\begin{pmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{pmatrix} = \frac{\partial}{\partial x_1}\frac{\partial f}{\partial x_1} + \ldots + \frac{\partial}{\partial x_n}\frac{\partial f}{\partial x_n} = \Delta f,$$

where

$$\Delta := \frac{\partial^2}{\partial x_1^2} + \ldots + \frac{\partial^2}{\partial x_n^2}$$

is the so-called ***Laplace operator*** or ***Laplacian***. One may also remember the above calculation using the nabla operator, although the following is rather sketchy:

$$\operatorname{div}\operatorname{grad} f = \nabla \cdot (\nabla f) \stackrel{?}{=} (\nabla \cdot \nabla) f = \Delta f.$$

Remark. In § 7.3 we shall see how vector fields give rise to 'things that can be integrated along curves or surfaces'. A glimpse at a more intrinsic viewpoint is given in § 7.6, when we discuss differential forms. To give a hint at where such things are applied, consider the subject of thermodynamics. The basic objects of interest therein, dubbed '(thermodynamic) systems', are described by their state (a collection of variables for pressure, temperature, volume, etc.). One then considers functions on the space of all possible states (called 'state functions'). One such function is called the *internal energy* of the system, denotes by the letter U. The ***first law of thermodynamics*** then states that for a closed[2] system any change in internal energy U equals the amount of heat supplied (δQ) to the system plus the work done on[3] the

[2]Meaning no mass being allowed to leave the system.

[3]Depending on whether one speaks of work done 'by' or 'on' the system one gets a sign $\pm$ in front of the δW term. Here we have just picked one, keeping in mind that the author of the present notes is not qualified to have a professional opinion on which choice of sign is more natural here.

system (δW). One phrases this as

$$\mathrm{d}U = \delta Q + \delta W,$$

and maintains that a change from one state to another along some path C in the state space is given by the integral

$$\int_C \mathrm{d}U = \int_C (\delta Q + \delta W).$$

In the language of § 7.6, $\mathrm{d}U$, δQ and δW are all 'differential 1-forms', although the 1-form $\mathrm{d}U$ arises as the total differential of a genuine (state) function U, whereas δQ and δW generally do not arise in this fashion. (Some authors still write $\mathrm{d}Q$ and $\mathrm{d}W$ and then stress that Q and W are not state functions.) For the connection to vector fields, see (7.9).

6.2. Taylor expansion and consequences

In this section, we mostly restrict to the one-dimensional setting.

6.2.1. Drawing conclusions from approximations. Recall that we use differentiation to approximate a given function a simpler function (see (5.1)). The goal is then to draw conclusions about the behaviour of the original function from the behaviour of the simpler function (the latter being easier to analyse, hopefully). We illustrate this principle via an example which should be familiar to the reader from school.

Example 6.2. Let $f\colon \mathbb{R} \to \mathbb{R}$ be a linear map, i.e., $f(x) = ax$ for some $a \in \mathbb{R}$ and all x. Under what conditions (on a) does f admit a maximum on $\mathbb{R}$? This is easy. If $a \neq 0$, then f takes on arbitrarily large values. Indeed, $\lim_{x\to\pm\infty} f(x) = \infty$, where the sign $\pm$ ought to be chosen according to the sign of a. Hence, f does not admit a maximum on $\mathbb{R}$. On the other hand, if $a = 0$, then f is constant. In particular, it *does* attain a maximum on $\mathbb{R}$, namely 0 and this value is attained everywhere. The same argument applies when f is affine-linear, i.e., $f(x) = ax + b$ for some $a, b \in \mathbb{R}$ and all x. Here the existence (and position) of a maximum of f depend only on f, but not on b. (On the other hand, if f does admit a maximum, then $a = 0$ and f is the constant function $x \mapsto b$ which attains its maximal value b everywhere.)

Next, suppose that $f\colon \mathbb{R} \to \mathbb{R}$ is an arbitrary differentiable map, not necessarily linear. If x_0 is a local maximum of f, i.e., if $f(x) \leq f(x_0)$ for all x in some neighbourhood of x_0, then

$$\frac{f(x)-f(x_0)}{x-x_0} \begin{cases} \geq 0 & \text{if } x < x_0, \\ \leq 0 & \text{if } x > x_0. \end{cases}$$

From this it follows that

$$f'(x_0) = \lim_{\substack{x \to x_0 \\ x \neq x_0}} \frac{f(x) - f(x_0)}{x - x_0} \begin{cases} \geq 0, \\ \leq 0, \end{cases}$$

whence $f'(x_0) = 0$. One can interpret this as saying that if f admits a local maximum, then so does its linearisation $x \mapsto f(x_0) + f'(x_0)(x - x_0)$ at x_0. We record this result as follows:

Proposition 6.3. *Let $f : I \to \mathbb{R}$ be a differentiable function on some open interval $I \subseteq \mathbb{R}$. If f attains a local maximum or minimum at $x_0 \in I$, then $f'(x_0) = 0$.*

Remark. The converse of Proposition 6.3 may fail to hold. For instance, f given by $f(x) = x^3$ satisfies $f'(x) = 3x^2$, which vanishes at 0, but f does not admit a local maximum at 0.

The above discussion should show that sometimes linear approximation is insufficient for drawing the desired conclusions. One way to remedy this deficiency is to consider higher-degree approximations.

6.2.2. Higher-degree approximations. Note that for differentiable $f : I \to \mathbb{R}$ defined on some open interval $I \subseteq \mathbb{R}$, the map $f' : I \to \mathbb{R}$ which maps x to $f'(x)$ is again a perfectly good function. We may even try to differentiate f'. This may not be possible (see, e.g., § 4.2), but let us suppose that our function f is 'nice enough' so that it is. In fact, let us suppose that f'' is not only defined throughout I, but is even continuous. By the fundamental theorem of calculus, we may write

$$f(x) = f(x_0) + \int_{x_0}^{x} f'(\xi)\,\mathrm{d}\xi$$

for any $x_0, x \in I$. Observe that dropping the integral here furnishes the (potentially not so good) approximation $f(x) \approx f(x_0)$. We now write the integrand $f'(\xi)$ as $f'(\xi) \cdot (x - \xi)^0$ and use integration by parts, differentiating f' and integrating $(x - \xi)^0$ (with respect to the variable ξ). This yields

$$f(x) = f(x_0) \underbrace{- f'(\xi)(x - \xi)\Big|_{\xi = x_0}^{x}}_{= f'(x_0)(x - x_0)} + \int_{x_0}^{x} f''(\xi)(x - \xi)\,\mathrm{d}\xi.$$

If we drop the integral here, we obtain the arguably better approximation

$$f(x) \approx f(x_0) + f'(x_0)(x - x_0). \tag{6.1}$$

This suggests that we may obtain even better approximations by iterating this process. Indeed, silently assuming that f'' is continuously differentiable, we may use integration by parts once more, differentiating f'' and integrating $x - \xi$. This yields

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + f''(x_0)\frac{1}{2}(x - x_0)^2 + \int_{x_0}^{x} f'''(\xi)\frac{1}{2}(x - \xi)^2\,\mathrm{d}\xi.$$

Again, discarding the integral, we arrive at the approximation

$$f(x) \approx f(x_0) + f'(x_0)(x - x_0) + f''(x_0)\frac{1}{2}(x - x_0)^2. \tag{6.2}$$

We pause for a moment, to see that the approximation (6.2) carries more information than (6.1). Suppose that we are interested in maximising $f\colon I \to \mathbb{R}$. By Proposition 6.3 we certainly have $f'(x_0) = 0$ at any point x_0 where f attains a local maximum. If f is given explicitly, then we may hope to solve $f'(x_0) = 0$ for the 'unknown' x_0 and thus get some *candidates* for extreme points. Suppose now that x_0 was obtained in this manner. How do we tell that f does admit a local maximum at x_0? By the remark following Proposition 6.3 we know that the fact that $f'(x_0) = 0$ does not suffice to ensure that f has a local extremum at x_0. Note that (6.2) takes the form

$$f(x) \approx f(x_0) + f''(x_0)\frac{1}{2}(x - x_0)^2.$$

Suppose that $f''(x_0) \neq 0$. Then, as x varies, the graph of the right hand side of the above is a parabola with extreme point at x_0. If $f''(x_0) > 0$, then the extreme point of the parabola is a local (in fact, global) minimum. Similarly, if $f''(x_0) < 0$, then the extreme point of the parabola is a local (in fact, global) maximum. We shall not attempt to prove this here, but the main insight (already known from school!) is, that the situation with extreme points of the quadratic approximation to f as given by (6.2) carries over to f:

Proposition 6.4. *Let $f\colon I \to \mathbb{R}$ be a differentiable function on some open interval $I \subseteq \mathbb{R}$. Suppose that $x_0 \in I$ is such that $f''(x_0) = 0$ and $f''(x_0) < 0$ (or > 0). Then f admits a local maximum (respectively, minimum) at x_0.*

Note that Proposition 6.4 fails to draw a conclusion about points x_0 for which $f'(x_0) = f''(x_0) = 0$. Higher-order approximations are needed for this.

To get hold of such higher-order approximations, we repeat the integration by parts procedure that brought us to (6.2) further. In this way, one finds that, for any $n = 0, 1, 2, \ldots$,

$$f(x) = \sum_{k=0}^{n} \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k + \int_{x_0}^{x} \frac{f^{(n+1)}(\xi)}{n!}(x - \xi)^n \, \mathrm{d}\xi, \tag{6.3}$$

where $f^{(0)} = f$, $f^{(1)} = f'$, $f^{(2)} = f''$, $\ldots$ are the higher-order derivatives of f. (Once more, the above is under the assumption that f is continuously differentiable sufficiently often.)

The sum on the right hand side of (6.3) is called the ***n-th Taylor polynomial for f at x_0***.[4] Upon letting $n \to \infty$, the some on the right hand side of (6.3) turns into an

[4]Observe that our numbering starts with $n = 0$, so the n-Taylor polynomial is actually the n+1-th in some sense.

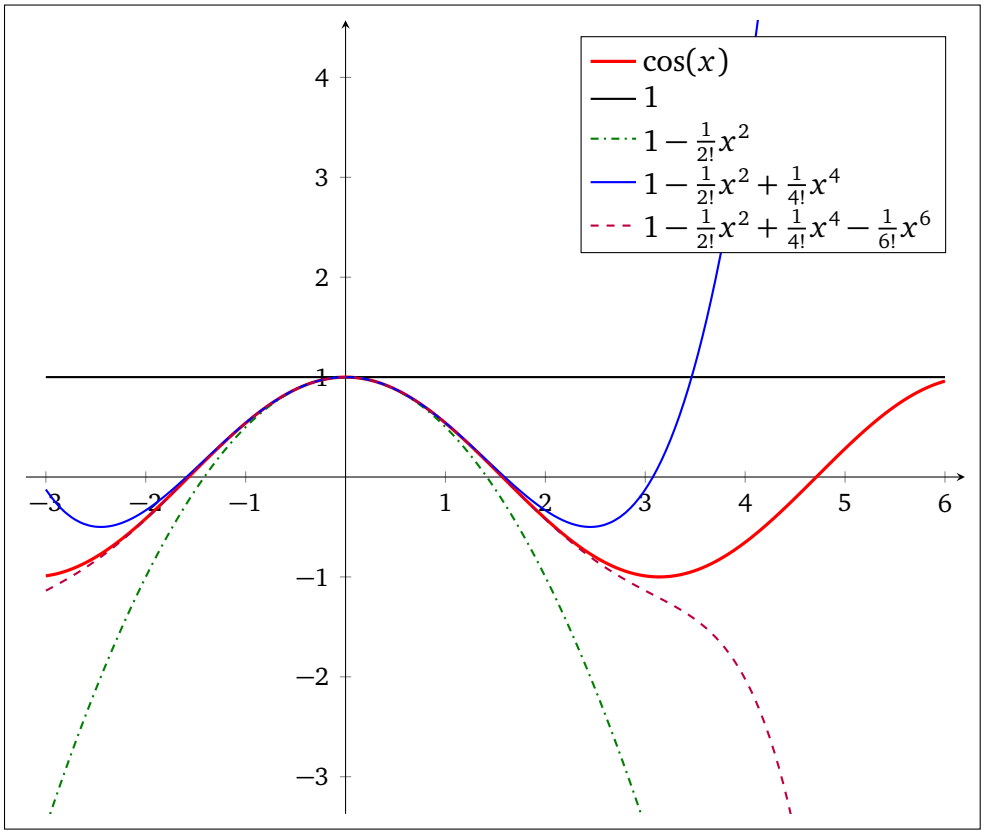


Figure 54. Plot of the cosine function and its k-th Taylor polynomials about 0 with $k = 0, 2, 4, 6$. Note that the k+1-th Taylor polynomial (about 0) of the cosine function coincides with the k-th Taylor polynomial for even k, because $\cos^{(k+1)}(0)$ vanishes.

infinite series and, assuming that the derivatives of f do not increase to rapidly, the denominator $n!$ should (hopefully) kill off the contribution from the integral. Thus,

$$f(x) \approx \sum_{k=0}^{\infty} \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k.$$

The question as to if '$\approx$' here is, in fact, '$=$', is a rather subtle one. We shall not pursue this line of thought further and mention only that the above is known as the ***Taylor series expansion of f at x_0***.

Example. Take $f(x) = x^2 + 3x + 7$ and $x_0 = 0$. Then

$$f(x_0) = x_0^2 + 3x_0 + 7 = 7, \quad f'(x_0) = 2x_0 + 3 = 3, \quad f''(x_0) = 2,$$

and $f^{(k)}(x_0) = 0$ for $k \geq 3$. In particular, the first three Taylor polynomials for f at $x_0 = 0$ are

$$7, \quad 7 + \frac{3}{1!}x, \quad 7 + \frac{3}{1!}x + \frac{2}{2!}x^2 = f(x).$$

Higher-order Taylor polynomials for f coincide with the third Taylor polynomials, which itself coincides with f.

For illustration's sake, let us also compute the relevant Taylor polynomials at $x_0 = 1$. Here we find

$$f(x_0) = x_0^2 + 3x_0 + 7 = 11, \quad f'(x_0) = 2x_0 + 3 = 5, \quad f''(x_0) = 2,$$

and, again, $f^{(k)}(x_0) = 0$ for $k \geq 3$. Therefore, the first three Taylor polynomials for f at $x_0 = 1$ are

$$11, \quad 11 + \frac{5}{1!}(x-1), \quad 11 + \frac{5}{1!}(x-1) + \frac{2}{2!}(x-1)^2.$$

The last Taylor polynomials here is

$$= 11 + 5(x-1) + (x-1)^2 = 11 + 5x - 5 + x^2 - 2x + 1 = 7 + 3x + x^2 = f(x).$$

The above example shows a general phenomenon: the n-th Taylor polynomial of a polynomial with degree not exceeding n coincides with that polynomial. In particular, the Taylor *series* of a polynomial coincides with that polynomial as well.

Example. Take $f = \cos$ and $x_0 = 0$. We have

$$f' = -\sin, \quad f'' = -\cos = -f, \quad f''' = \sin, \quad f'''' = \cos = f.$$

From then on, this pattern repeats with period 4. This shows that

$$f^{(k)} = \begin{cases} \cos & \text{if } k = 0, 4, \ 8, 12, \ldots, \\ -\sin & \text{if } k = 1, 5, \ 9, 13, \ldots, \\ -\cos & \text{if } k = 2, 6, 10, 14, \ldots, \\ \sin & \text{if } k = 3, 7, 11, 15, \ldots. \end{cases}$$

Upon plugging in x_0, we get the following table of values:

k	0	1	2	3	4	5	6	7	8	9	...
$f^{(k)}(0)$	1	0	-1	0	1	0	-1	0	1	0	...

Taking $k = 5$ in (6.3), for instance, we find that

$$\cos(x) = 1 + \frac{-1}{2!}x^k + \frac{1}{4!}x^4 + \int_0^x \frac{f^{(6)}(\xi)}{5!}\xi^5 \, \mathrm{d}\xi.$$

Observe that the fifth Taylor polynomial of cos at 0 has degree 4 and not 5. This is due to the fifth derivative $\cos^{(5)}$ vanishing at 0. See also Figure 54. Noting that

$f^{(6)} = -\cos$ takes only values in $[-1,1]$, we can estimate the integral above by

$$\left|\int_0^x \frac{f^{(6)}(\xi)}{5!}\xi^5\,\mathrm{d}\xi\right| \le \int_0^x \left|\frac{f^{(6)}(\xi)}{5!}\xi^5\right|\mathrm{d}\xi \le \frac{1}{5!}\int_0^x |\xi|^5\,\mathrm{d}\xi = \frac{1}{5!}\frac{1}{6}x^6 = \frac{1}{6!}x^6. \tag{6.4}$$

For illustration's sake, we plug in $x = 1$. Then

$$\cos(1) = 1 + \frac{-1}{2!} + \frac{1}{4!} + \int_0^1 \ldots\,\mathrm{d}\xi = \frac{13}{24} + \int_0^1 \ldots\,\mathrm{d}\xi,$$

where, by (6.4), the integral does not exceed $1/6! = 720$. Hence, we know that

$$\cos(1) \in \left[\frac{13}{24} - \delta, \frac{13}{24} + \delta\right], \quad \text{where} \quad \delta = \frac{1}{720}.$$

Using a calculator, one may find that

$$\cos(1) = 0.5403023\ldots, \quad \frac{13}{24} = 0.541\overline{6}, \quad \frac{1}{720} = 0.00138\overline{8},$$

and, indeed, the difference $\cos(1) - \frac{13}{24} = -0.0013643\ldots$ does not exceed δ.

The previous example is characteristic. In applications, we hope to deal with functions which do not change too quickly (i.e., have a small derivative), or whose derivative does not change too quickly (i.e., have a small second derivative), or whose derivative of its derivative does not change too quickly etc. In such cases, we expect the integral

$$\int_{x_0}^x \frac{f^{(n+1)}(\xi)}{n!}(x-\xi)^n\,\mathrm{d}\xi$$

in (6.3) to be small if n is chosen suitably and x is not too far away from x_0. How 'small' constitutes 'small enough' and what 'not too far away' means, does obviously depend on the application one has in mind.

6.2.3. A closer look at the error terms. We shall take another look at (6.3). Suppose that $[x_0, x_1]$ is contained in the interval I on which f is defined. If $f^{(n+1)}$ is continuous on $[x_0, x_1]$ (which we have assumed earlier anyway), then $|f^{(n+1)}|$ assumes a maximum M there. (This is a theorem due to Karl Weierstraß and we will not prove this.) In particular, for $x \in [x_0, x_1]$, the integral in (6.3) can be bounded by

$$\int_{x_0}^x \left|\frac{f^{(n+1)}(\xi)}{n!}(x-\xi)^n\right|\mathrm{d}\xi \le \frac{M}{n!}\int_{x_0}^x |(x-\xi)^n|\,\mathrm{d}\xi = \frac{M}{(n+1)!}|x-x_0|^{n+1}. \tag{6.5}$$

Note that as x approaches x_0, the right hand side of the above goes to zero as quickly as $|x-x_0|^{n+1}$.

We shall write $O_{x\to x_0}(g(x))$ as a placeholder for a suitable function f for which f/g remains bounded in a neighbourhood of x_0. For instance, $x^2 = O_{x\to 0}(x)$ or $x^2 + 18 = 18 + O_{x\to 0}(x)$. (Note that we do not mean to say that $x^2 + 18 = 18 + g(x)$ holds for *every* function g such that $g(x) = O_{x\to 0}(x)$, but merely that there is *some*

function g which makes this equation true; clearly that function is $x \mapsto x^2$.) Moreover, by (6.5), we have the following version of (6.2):

$$f(x) = f(x_0) + f'(x_0)(x-x_0) + f''(x_0)\frac{1}{2}(x-x_0)^2 + O_{x\to x_0}(|x-x_0|^3).$$

Writing $x = x_0 + h$ this can also be written as

$$f(x_0+h) = f(x_0) + f'(x_0)h + f''(x_0)\frac{1}{2}h^2 + O_{h\to 0}(|h|^3). \tag{6.6}$$

6.2.4. Numerical differentiation. Our above discussion can be used for numerical problems. We consider the following problem:

> For a function $f\colon \mathbb{R} \to \mathbb{R}$ which we may assume to be sufficiently often differentiable, how well can we determine $f'(x_0)$ if we are given only the values of f at the h-spaced points
>
> $$\ldots,\ f(x_0-h),\ f(x_0),\ f(x_0+h),\ f(x_0+2h),\ \ldots\ ?$$

In the above, $h > 0$ is assumed to be small, but thinking of implementing the above situation in a computer, taking h to be outrageously small may have an unacceptable impact on performance. Hence, we shall try to get the most out of the information available. In view of the definition of $f'(x_0)$ as

$$\lim_{\substack{h\to 0\\ h\neq 0}} \frac{f(x_0+h)-f(x_0)}{h},$$

we may guess that

$$\frac{f(x_0+h)-f(x_0)}{h}$$

should be a good approximation to $f'(x_0)$. Let us try to see what accuracy we may hope for with this approach. From (6.6) we obtain

$$f(x_0 \pm h) = f(x_0) \pm f'(x_0)h + f''(x_0)\frac{1}{2}h^2 + O_{h\to 0}(|h|^3), \tag{6.7}$$

The difference quotient

$$\begin{aligned} \frac{f(x_0+h)-f(x_0)}{h} &= f'(x_0) + f''(x_0)\frac{1}{2}h + O_{h\to 0}(h^2) \\ &= f'(x_0) + O_{h\to 0}(|h|), \end{aligned} \tag{6.8}$$

whose limit as $h \to 0$ is used to *define* the first derivative of f at x_0 obviously approaches $f'(x_0)$ with only 'linear speed' in h if $f''(x_0)$ happens to be non-zero.

Can we do better than linear speed? Yes! By taking (6.7) with both choices of sign and subtracting, we see that the terms with $f''(x_0)$ cancel with each other. Indeed,

$$f(x_0+h) - f(x_0-h) = 2f'(x_0)h + O_{h\to 0}(|h|^3),$$

In particular, the ***central difference quotient***

$$\frac{f(x_0+h)-f(x_0-h)}{2h} = f'(x_0)+O_{h\to 0}(h^2)$$

is expected to yield an even better approximation to $f'(x_0)$ than (6.8). Similarly, one can check that

$$\frac{f(x_0-2h)-8f(x_0-h)+8f(x_0+h)-f(x_0+2h)}{12h} = f'(x_0)+O_{h\to 0}(h^4).$$

For an illustration of this, see Figure 55.

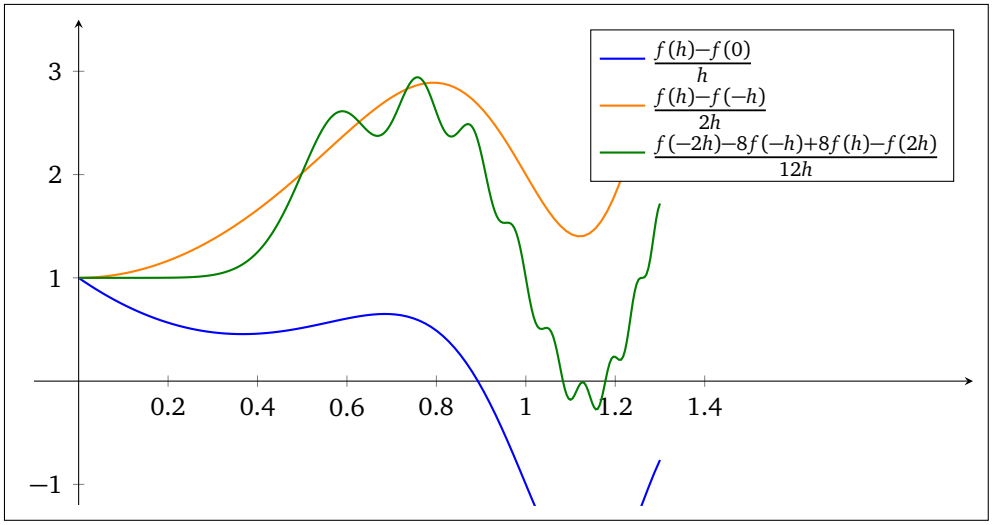


Figure 55. Illustration of various types of difference quotients used to approximate the derivative of $f(x) = x^3 - 3x^2 + x + \sin(\pi x^3)$ at zero.

6.3. Newton's method

6.3.1. Derivation of the method. In this section, we shall discuss a method used to solve the equation

$$\vec{f}(\vec{x}) \stackrel{!}{=} \vec{b} \tag{6.9}$$

numerically for $\vec{x}$, where $f\colon \mathbb{R}^n \to \mathbb{R}^n$ is assumed to be differentiable and $\vec{b} \in \mathbb{R}^n$ is a fixed vector. We start by choosing some arbitrary value for $\vec{x}$. Let us call this value $\vec{x}_0$, say. Certainly, there is no guarantee that $\vec{x}_0$ may not solve the equation (6.9). ***Newton's method*** makes an attempt at constructing a better guess, $\vec{x}_1$ say, from $\vec{x}_0$. This process can then be iterated, producing further guesses $\vec{x}_2$, $\vec{x}_3$, ..., and by stopping at some point, one may hope to have found a sufficiently good approximation

to an actual solution of (6.9). To describe Newton's method, let $\vec{x}$ be a solution to (6.9). Recall that, by (5.1),

$$\begin{aligned}\vec{b} = \vec{f}(\vec{x}) = \vec{f}(\vec{x}_0 + (\vec{x} - \vec{x}_0)) &\approx \vec{f}(\vec{x}_0) + \mathrm{d}\vec{f}_{\vec{x}_0}(\vec{x} - \vec{x}_0) \\ &= \vec{f}(\vec{x}_0) + J_{\vec{f}}(\vec{x}_0)\vec{x} - J_{\vec{f}}(\vec{x}_0)\vec{x}_0.\end{aligned} \tag{6.10}$$

Thus, we have the approximate equation

$$\vec{b} \approx \vec{f}(\vec{x}_0) + J_{\vec{f}}(\vec{x}_0)\vec{x} - J_{\vec{f}}(\vec{x}_0)\vec{x}_0.$$

We now treat this approximation as an equation and seek to solve it for $\vec{x}$. This should be easier, than solving the original equation (6.9), because our new equation is *linear*. Assuming that $J_{\vec{f}}(\vec{x}_0)$ is invertible, we infer

$$\vec{x} \approx \vec{x}_0 - J_{\vec{f}}(\vec{x}_0)^{-1}(\vec{f}(\vec{x}_0) - \vec{b}). \tag{6.11}$$

In lack of knowing the actual value of $\vec{x}$, we now do the next best thing that comes to mind and plainly *define*

$$\vec{x}_1 := \vec{x}_0 - J_{\vec{f}}(\vec{x}_0)^{-1}(\vec{f}(\vec{x}_0) - \vec{b}).$$

Clearly, we cannot expect the approximation (6.11) to be exact, so $\vec{x}_1$ should not be expected to solve (6.9). In particular, if $\vec{x}_0$ lies very far away from a solution $\vec{x}$ to (6.9), then the approximation (6.10) may be utterly insufficient, rendering the value $\vec{x}_1$ quite useless.

6.3.2. Examples for the one-dimensional case. For the remainder of this section, we shall restrict to the one-dimensional case, and drop the arrows from our notation. Then the matrix $J_{\vec{f}}(\vec{x}_0)$ is simply $f'(x_0)$ and the matrix inversion turns into a plain division. Newton's method for solving

$$f(x) \stackrel{!}{=} b$$

then takes the form

$$x_0 = \text{arbitrary real number},$$

$$x_1 = x_0 - \frac{f(x_0) - b}{f'(x_0)}, \quad x_2 = x_1 - \frac{f(x_1) - b}{f'(x_1)}, \quad x_3 = x_2 - \frac{f(x_2) - b}{f'(x_2)}, \quad \dots$$

Example. We shall attempt to solve the equation

$$x^2 - 2 \stackrel{!}{=} 0$$

numerically using Newton's method. As the left hand side factors as

$$(x - \sqrt{2})(x + \sqrt{2}),$$

we already know that there are precisely two solutions, namely $x = \pm\sqrt{2}$. In particular, we have a good reference point for seeing if Newton's method does yield approximative solutions. In fact, Newton's method applied to find square roots is also

known as ***Heron's method*** and has already been used ca. 1750 BC in Mesopotamia. Note that we have

$$f(x) = x^2 - 2, \quad f'(x) = 2x.$$

Suppose we took $x_0 = 0$. Then $f'(x_0) = 0$ and we cannot compute x_1, as we cannot divide by zero. Therefore, let us take $x_0 = 1$ instead. Then

$$x_1 = x_0 - \frac{x_0^2 - 2}{2x_0} = 1 - \frac{1^2 - 2}{2 \cdot 1} = 1 - \frac{-1}{2} = \frac{3}{2}.$$

(See Figure 56 (a).) Next

$$x_2 = x_1 - \frac{x_1^2 - 2}{2x_1} = \frac{3}{2} - \frac{(3/2)^2 - 2}{2(3/2)} = \ldots = \frac{17}{12}.$$

Similarly, one computes

$$x_3 = \frac{577}{408}, \quad x_4 = \frac{665857}{470832}, \quad x_5 = \frac{886731088897}{627013566048},$$

and the even more impressive looking

$$x_6 = \frac{1572584048032918633353217}{1111984844349868137938112}.$$

For comparison's sake, we also print the decimal expansions of the numbers involved and underline the part of the initial segment that already matches that of $\sqrt{2}$:

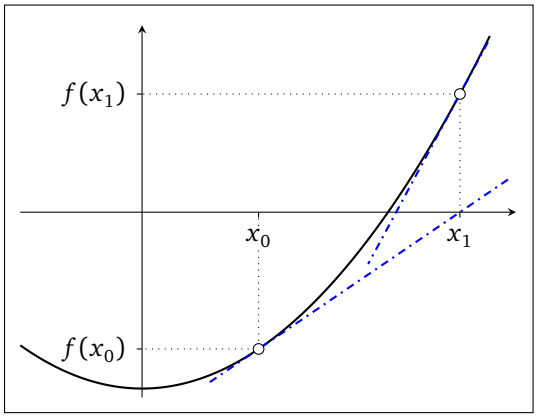
$$\begin{aligned}
x_0 &= \underline{1}.000000000000000000000000000000000000000000000000\ldots,\\
x_1 &= \underline{1}.500000000000000000000000000000000000000000000000\ldots,\\
x_2 &= \underline{1.41}666666666666666666666666666666666666666666666666\ldots,\\
x_3 &= \underline{1.41421}568627450980392156862745098039215686274509 80\ldots,\\
x_4 &= \underline{1.4142135623}74689910626295578890134910116559622115 7\ldots,\\
x_5 &= \underline{1.41421356237309504880168}962350253024361498192577 62\ldots,\\
x_6 &= \underline{1.41421356237309504880168872420969807856967187537}72\ldots,\\
&\vdots\\
\sqrt{2} &= 1.41421356237309504880168872420969807856967187537 69\ldots.
\end{aligned}$$

Incidentally, if we had chosen $x_0 = -1$ instead, then we would have obtained precisely the same numbers, but with opposite sign. Newton's method would have yielded approximations to $-\sqrt{2}$.

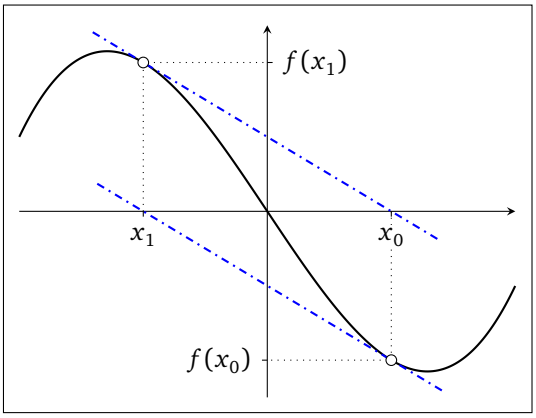
The following example shows that Newton's method may indeed fail in some cases.

Example. We shall apply Newton's method to the equation

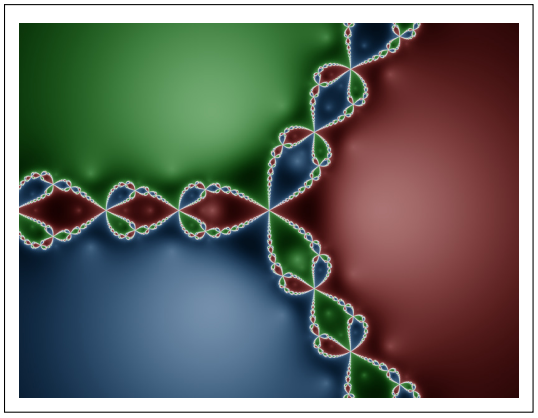
$$x^3 - 5x \stackrel{!}{=} 0$$



(a) $f(x) = x^2 - 2$ and initial value $x_0 = 1$. (The vertical axis is scaled down, however.)



(b) A situation where Newton's method exhibits periodic behaviour and does, therefore, not yield a sequence which converges to a solution of the equation in question.



(c) $f(x) = x^3 - 1$. The roots of f are the complex numbers $\zeta = \exp(2\pi i/3)$, ζ^2 and $\zeta^3 = 1$. A point $x_0 \in \mathbb{C}$ is coloured green, blue or red according to whether Newton's method, starting with that point, converges to ζ, ζ^2 or ζ^3, respectively. At the white points, Newton's method fails to converge.
(Image source: https://commons.wikimedia.org/wiki/File:Julia-set_N_z3-1.png)

Figure 56. Newton's method applied to solve $f(x) \stackrel{!}{=} 0$ with initial value x_0.

with initial guess $x_0 = \pm 1$. (Choose a sign!) We find that

$$x_1 = x_0 - \frac{x_0^3 - 5x_0}{3x_0^2 - 5} = \pm 1 - \frac{(\pm 1)^3 - 5(\pm 1)}{3 \cdot (\pm 1)^2 - 5} = \pm 1 - \frac{\pm 1 \mp 5}{-2} = \pm 1 \mp 2 = \mp 1.$$

It follows that

$$x_0 = \pm 1, \quad x_1 = \mp 1, \quad x_2 = \pm 1, \quad x_3 = \mp 1, \quad x_4 = \pm 1, \quad \ldots \text{ (periodically)}.$$

However, neither $+1$ nor -1 solves the equation from above. In particular, given the above choice of x_0, Newton's method does not yield a sequence converging to a solution (see also Figure 56 (b)).

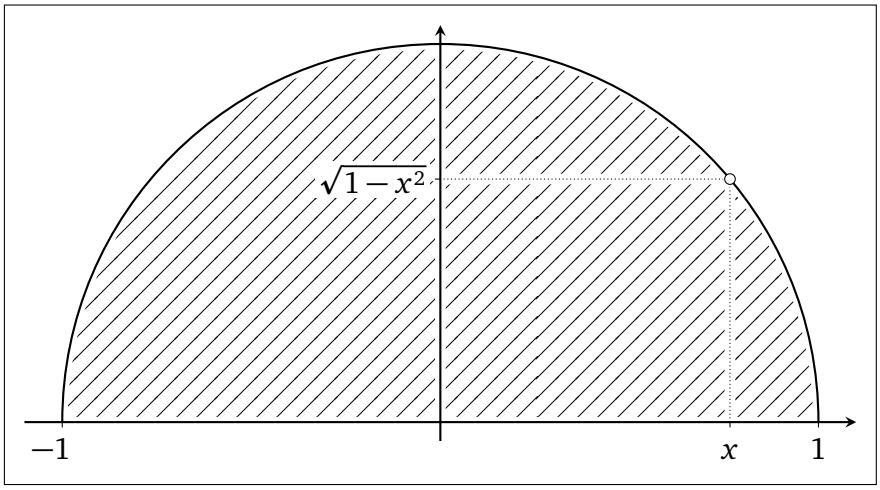


Figure 57. The graph of $[-1,1] \mapsto \mathbb{R}$, $x \mapsto \sqrt{1-x^2}$, is the top half of the circle with centre $(0,0)$ and radius 1. The area underneath is half the area of the circle. Thus, $\int_{-1}^{1} \sqrt{1-x^2}\,\mathrm{d}x = \frac{\pi}{2}$.

6.3.3. Fractals. We close this section by mentioning that the question as to whether Newton's method converges for a given initial value is a complicated, yet fascinating one, even when studied for simple functions like polynomials. Removing the restriction to starting with real values immediately takes one into a domain of mathematics known as 'complex dynamics'. Plotting the convergence behaviour exhibited by Newton's method produces mesmerising pictures (see Figure 56 (c)).

6.4. Polar, spherical and cylindrical coordinates

Models of physical problems frequently admit various sorts of symmetries; think, for instance, of electrons and their induced electric potential which admits spherical symmetry, or the magnetic field induced by current flowing through a thin, long wire which may be modelled as infinitely thin and infinitely long (the induced field will admit cylindrical symmetry). In such cases, the calculations that arise often become much easier if one chooses "coordinates" which are tailored to exploit the symmetries inherent to the problem at hand.

Example. We wish to compute $I = \displaystyle\int_{-1}^{1} \sqrt{1-x^2}\,\mathrm{d}x$. Note that the vector

$$\vec{v}(x) = (x, \sqrt{1-x^2})$$

has length $\|\vec{v}(x)\| = \sqrt{x^2 + (1-x^2)} = 1$. We see that the graph of the integrand in the integral defining I looks like part of the unit circle (centre $\vec{0}$ and radius 1); see Figure 57. Indeed, an easy monotonicity inspection shows that the graph for $-1 \leq x \leq 1$ is *exactly* the upper half of the unit circle. Thus I is the area under this

half circle. As the full unit circle has area π, we conclude that $I = \frac{\pi}{2}$. Can we show this "analytically"? Well, with a bit of tenacity one can actually compute the above integral using integration by parts, but our previous considerations suggest that we should somehow try to exploit the "circular structure" in our problem. This leads us to use the well-known substitution rule

$$\int_{g(a)}^{g(b)} f(x)\,\mathrm{d}x = \int_a^b f(g(\varphi))g'(\varphi)\,\mathrm{d}\varphi$$

with $g = \sin$ to arrive at

$$I = \int_{-\pi/2}^{\pi/2} \sqrt{1-(\sin\theta)^2}\sin'(\theta)\,\mathrm{d}\theta = \int_{-\pi/2}^{\pi/2} |\cos\theta|\cos(\theta)\,\mathrm{d}\theta = \int_{-\pi/2}^{\pi/2} \cos(\theta)^2\,\mathrm{d}\theta.$$

Using the trigonometric identity $2\cos(\theta)^2 = 1+\cos(2\theta)$ we conclude that

$$I = \int_{-\pi/2}^{\pi/2} \frac{1+\cos(2\theta)}{2}\,\mathrm{d}\theta = \frac{\pi}{2} + \frac{1}{4}\int_{-\pi/2}^{\pi/2} 2\cos(2\theta)\,\mathrm{d}\theta = \frac{\pi}{2} + \frac{1}{4}\sin(2\theta)\Big|_{\phi=-\pi/2}^{\pi/2} = \frac{\pi}{2}.$$

We now define the three most common "curved" coordinate functions.

- $\vec{P}\colon \mathbb{R}^2 \to \mathbb{R}^2$, $(r,\varphi) \mapsto \begin{pmatrix} r\cos\varphi \\ r\sin\varphi \end{pmatrix}$. (***Polar coordinates***.)
- $\vec{Z}\colon \mathbb{R}^3 \to \mathbb{R}^3$, $(r,\varphi,z) \mapsto \begin{pmatrix} r\cos\varphi \\ r\sin\varphi \\ z \end{pmatrix}$. (***Cylindrical coordinates***.)
- $\vec{K}\colon \mathbb{R}^3 \to \mathbb{R}^3$, $(r,\theta,\varphi) \mapsto \begin{pmatrix} r\cos(\varphi)\sin\theta \\ r\sin(\varphi)\sin\theta \\ r\cos\theta \end{pmatrix}$. (***Spherical coordinates***.)

Observe the counter-intuitive ordering of the arguments of $\vec{K}$: (r,θ,φ) and not (r,φ,θ) as one might expect when looking at $\vec{Z}$!

Proposition 6.5. *We have*

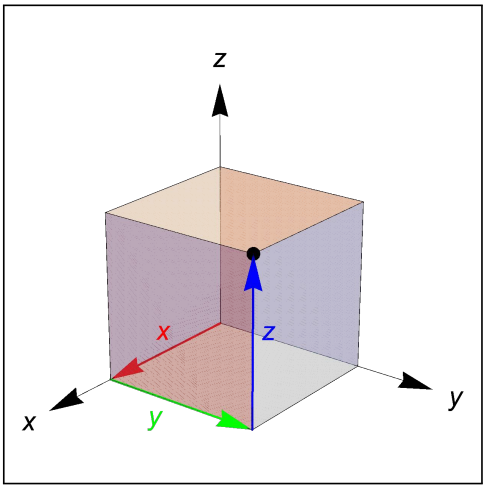
(1) $\det J_{\vec{P}}(r,\varphi) = r$,
(2) $\det J_{\vec{Z}}(r,\varphi,z) = r$,
(3) $\det J_{\vec{K}}(r,\theta,\varphi) = r^2\sin(\theta)$.

Proof. This is just a lengthy computation. We have

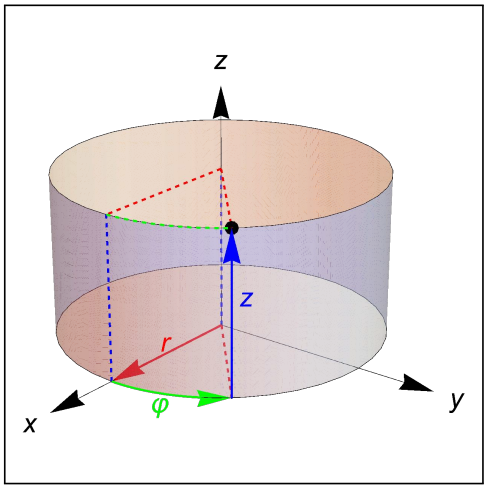
$$J_{\vec{P}}(r,\varphi) = \begin{pmatrix} \cos\varphi & -r\sin\varphi \\ \sin\varphi & r\cos\varphi \end{pmatrix}, \quad J_{\vec{Z}}(r,\varphi,z) = \begin{pmatrix} \cos\varphi & -r\sin\varphi & 0 \\ \sin\varphi & r\cos\varphi & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

and

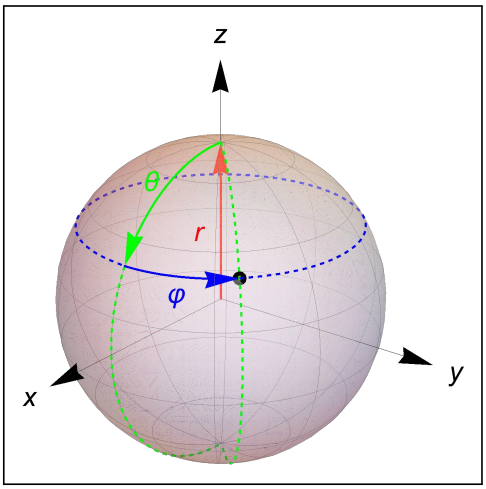
$$J_{\vec{K}}(r,\theta,\varphi) = \begin{pmatrix} \cos(\varphi)\sin\theta & r\cos(\varphi)\cos\theta & -r\sin(\varphi)\sin\theta \\ \sin(\varphi)\sin\theta & r\sin(\varphi)\cos\theta & r\cos(\varphi)\sin\theta \\ \cos\theta & -r\sin\theta & 0 \end{pmatrix}.$$



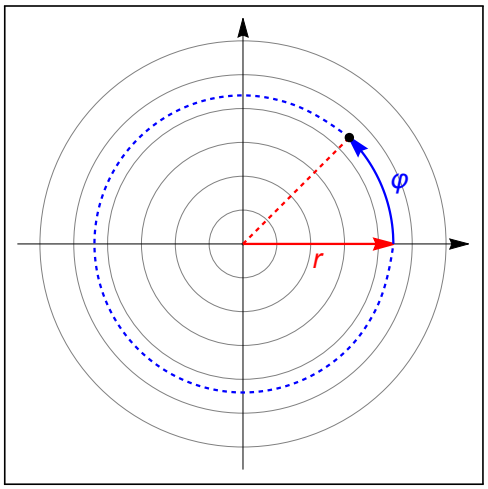
(a) Cartesian coordinates in $\mathbb{R}^3$.



(b) Cylindrical coordinates $\vec{Z}$.



(c) Spherical coordinates $\vec{K}$.



(d) Polar coordinates $\vec{P}$.

Figure 58. Some coordinate functions.

The first two determinants are not difficult to compute:

$$\det J_{\vec{P}}(r,\varphi) = r\cos(\varphi)^2 + r\sin(\varphi)^2 = r,$$
$$\det J_{\vec{Z}}(r,\varphi,z) = \det\begin{pmatrix}\cos\varphi & -r\sin\varphi \\ \sin\varphi & r\cos\varphi\end{pmatrix} = r\cos(\varphi)^2 + r\sin(\varphi)^2 = r.$$

As for $X := \det J_{\vec{K}}(r,\theta,\varphi)$, we use Leibniz's expansion with respect to the last column,[5] getting

$$X = -r\sin(\varphi)\sin(\theta)\det\begin{pmatrix} \boxtimes & \boxtimes & \boxtimes \\ \sin(\varphi)\sin\theta & r\sin(\varphi)\cos\theta & \boxtimes \\ \cos\theta & -r\sin\theta & \boxtimes \end{pmatrix} +$$
$$-r\cos(\varphi)\sin(\theta)\det\begin{pmatrix} \cos(\varphi)\sin\theta & r\cos(\varphi)\cos\theta & \boxtimes \\ \boxtimes & \boxtimes & \boxtimes \\ \cos\theta & -r\sin\theta & \boxtimes \end{pmatrix} + 0\det(\ldots)$$

After computing the determinants on the right hand side and using the formula $\cos(\alpha)^2+\sin(\alpha)^2 = 1$ multiple times one deduces that $\det J_{\vec{K}}(r,\theta,\varphi) = r^2\sin(\theta)$. □

One would like to restrict the domain of definition of $\vec{P}$, $\vec{Z}$ and $\vec{K}$ for the purpose of making them map bijectively ("one to one") onto $\mathbb{R}^2$ or $\mathbb{R}^3$. There are, however, some slight complications. Usual choices are

$$\vec{P}|_U\colon U \xrightarrow{1:1} \mathbb{R}^2, \qquad \vec{Z}|_V\colon V \xrightarrow{1:1} \mathbb{R}^3,$$
$$\vec{K}|_{\mathbb{R}_+\times(0,\pi)\times[0,2\pi)}\colon \mathbb{R}_+\times(0,\pi)\times[0,2\pi) \xrightarrow{1:1} \mathbb{R}^3\setminus\{(0,0,z) : z\in\mathbb{R}\},$$

where

$$U = \mathbb{R}_+\times[0,2\pi)\cup\{(0,0)\} \quad \text{or} \quad U = \mathbb{R}_+\times(-\pi,\pi]\cup\{(0,0)\},$$
$$V = \mathbb{R}_+\times[0,2\pi)\times\mathbb{R}\cup\{(0,0,0)\} \quad \text{or} \quad V = \mathbb{R}_+\times(-\pi,\pi]\times\mathbb{R}\cup\{(0,0,0)\}.$$

Sometimes it is useful to compute derivatives with respect to new coordinates. We shall illustrate the process with respect to polar coordinates. Suppose one is given a function $f\colon\mathbb{R}^2\to\mathbb{R}$. Then one would like to compute

$$\partial_1(f\circ\vec{P}) \quad \text{and} \quad \partial_2(f\circ\vec{P}).$$

One often writes this as $\dfrac{\partial(f\circ\vec{P})}{\partial r}$ or (careful: physicist's notation!) $\dfrac{\partial f}{\partial r}$. The chain rule tells us

$$\begin{aligned} J_{f\circ\vec{P}}(r,\varphi) &= J_f(\vec{P}(r,\varphi))J_{\vec{P}}(r,\varphi) \\ &= \begin{pmatrix}(\partial_1 f)(\vec{P}(r,\varphi)) & (\partial_2 f)(\vec{P}(r,\varphi))\end{pmatrix}\begin{pmatrix}\cos\varphi & -r\sin\varphi \\ \sin\varphi & r\cos\varphi\end{pmatrix} \\ &=: \begin{pmatrix}\ldots_{(1)} & \ldots_{(2)}\end{pmatrix}, \end{aligned}$$

where

$$\ldots_{(1)} = (\partial_1 f)(\vec{P}(r,\varphi))\cos\varphi + (\partial_2 f)(\vec{P}(r,\varphi))\sin\varphi,$$
$$\ldots_{(2)} = -(\partial_1 f)(\vec{P}(r,\varphi))r\sin\varphi + (\partial_2 f)(\vec{P}(r,\varphi))r\cos\varphi.$$

[5]Technically, we have only discussed this for expanding with respect to the first row, but the principle is the same.

Therefore,

$$\partial_1(f\circ\vec{P})(r,\varphi)=\ldots_{(1)}=(\partial_1 f)(\vec{P}(r,\varphi))\cos\varphi+(\partial_2 f)(\vec{P}(r,\varphi))\sin\varphi.$$

A physicist would have spoken of "a function $f(x,y)$", written

$$\begin{pmatrix}x\\y\end{pmatrix}=\begin{pmatrix}r\cos\varphi\\r\sin\varphi\end{pmatrix}$$

and concluded that

$$\frac{\partial f}{\partial r}=\frac{\partial f}{\partial x}\frac{\partial x}{\partial r}+\frac{\partial f}{\partial y}\frac{\partial y}{\partial r}.$$

We shall test this new device on the weird function from Example 5.3.

Example 6.6 (Continuation of Example 5.3)**.** Consider the function

$$f\colon\mathbb{R}^2\to\mathbb{R},\quad (x,y)\mapsto\begin{cases}\dfrac{xy}{x^2+y^2} & \text{if } (x,y)\neq(0,0),\\ 0 & \text{if } x=y=0,\end{cases}$$

from Example 5.3 (see also Figure 52). We know that it is not continuous at $(0,0)$, but has both partial derivatives there. We shall try to see what we can learn about this function when looking at it in polar coordinates. We have

$$(f\circ\vec{P})(r,\varphi)=\frac{(r\cos\varphi)(r\sin\varphi)}{(r\cos\varphi)^2+(r\sin\varphi)^2}=(\cos\varphi)\sin\varphi=\tfrac{1}{2}\sin(2\varphi).$$

Switching to physicist's notation,

$$\frac{\partial f}{\partial r}(r,\varphi)=0,\quad \frac{\partial f}{\partial\varphi}(r,\varphi)=\cos(2\varphi).$$

The first equation shows us that f is constant on lines through the origin (excluding the origin). We shall also like to test the physicist's version of the chain rule. We have (tacitly omitting some function arguments in the process)

$$\begin{aligned}\frac{\partial f}{\partial r}(r,\varphi)&=\frac{\partial f}{\partial x}\frac{\partial x}{\partial r}+\frac{\partial f}{\partial y}\frac{\partial y}{\partial r}=\frac{\partial f}{\partial x}\cos\varphi+\frac{\partial f}{\partial y}\sin\varphi\\&=y\frac{y^2-x^2}{(x^2+y^2)^2}\cos\varphi+x\frac{x^2-y^2}{(x^2+y^2)^2}\sin\varphi\\&=(x\sin\varphi-y\cos\varphi)\frac{x^2-y^2}{(x^2+y^2)^2}.\end{aligned}$$

Here a mathematician (e.g., the author of these notes) is usually very confused, because we have three variables now. A physicist, however, recalls that $x=r\cos\varphi$ and $y=r\sin\varphi$. Therefore,

$$\frac{\partial f}{\partial r}(r,\varphi)=(r\cos(\varphi)\sin\varphi-r\sin(\varphi)\cos\varphi)\frac{r^2(\cos\varphi)^2-r^2(\sin\varphi)^2}{r^4}=0.$$

Observe that this is also what we see in Figure 52: the function f is constant[6] along lines through the origin.

[6]Strictly speaking, we should exclude the origin here.

CHAPTER 7

Integration

Integration is a process for collecting together lots of locally given data. Given our previous discussion of differentiation in Chapter 5, we already have a decent means for understanding *local* phenomena by way of reducing to problems of a linear-algebraic nature: such problems often turn out to be more amenable to study, because linear maps are quite "rigid", that is, they obey strong laws. (For instance, a linear map is already fully determined by its action on the standard unit vectors—a property which is hopelessly false for arbitrary maps; even for continuous or C^∞ ones.) In turn, integration allows one to pass from local information given everywhere to *global* information. Applications range from computing lengths, areas and volumes of objects (which is easy "locally" by means of differentiation and determinants) to the invocation of conservation laws to physical phenomena; for the latter, see our derivation of the diffusion equation in Chapter 8.

7.1. The higher-dimensional Darboux integral

We start by considering an n-dimensional cuboid (rectangular box)

$$Q = [a_1, b_1] \times \ldots \times [a_n, b_n] \subset \mathbb{R}^n$$

along with a bounded function $f : Q \to \mathbb{R}$ on it. We aim to define the integral

$$\int_Q f(\vec{x})\, \mathrm{d}^n \vec{x}. \tag{7.1}$$

Intuitively, at least for positive f, this should be the "n+1-dimensional volume under the graph of f". To approximate this, we define a partition $\mathscr{P}$ of Q to be a collection $(x_{ij})_{i=1,j=0}^{n,N_i}$ of real numbers x_{ij} satisfying

$$a_i = x_{i0} < \ldots < x_{iN_i} = b_i.$$

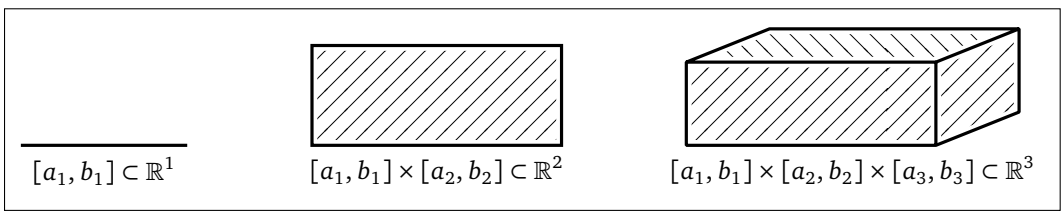


Figure 59. Cuboids in various dimensions.

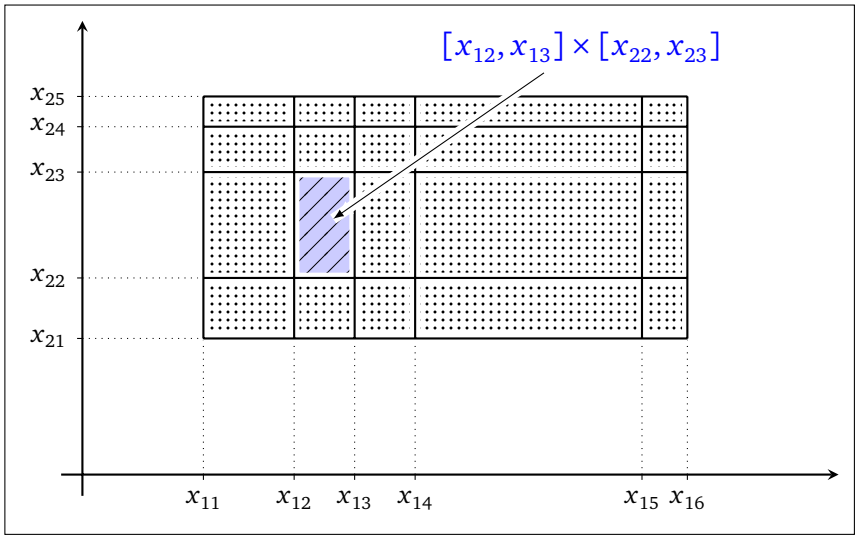


Figure 60. Partition of a two-dimensional cuboid.

(The numbers $N_1, \ldots, N_n \in \mathbb{N} \setminus \{1\}$ may also vary.) This decomposes Q into smaller cuboids given by

$$[\vec{x}, \vec{x}'] := [x_{1(j_1-1)}, x_{1j_1}] \times \ldots \times [x_{n(j_n-1)}, x_{1j_n}],$$

where $j_i \in \{1, \ldots, N_i\}$. We write $[\vec{x}, \vec{x}'] \in \mathscr{P}$ if $[\vec{x}, \vec{x}']$ is such a cuboid. Note, however, that the cuboids are not quite disjoint in general: they may have points on their boundaries in common. We write $\mathrm{vol}([\vec{x}, \vec{x}'])$ for the volume of the cuboid $[\vec{x}, \vec{x}']$. This is given by the formula

$$\mathrm{vol}([\vec{x}, \vec{x}']) := \prod_{i=1}^{n} \big(x_{ij_i} - x_{i(j_i-1)}\big).$$

Moreover, we define

$$\mathscr{U}(f, \mathscr{P}) := \sum_{[\vec{x}, \vec{x}'] \in \mathscr{P}} \inf(f|_{[\vec{x}, \vec{x}']}) \, \mathrm{vol}([\vec{x}, \vec{x}']),$$

$$\mathscr{O}(f, \mathscr{P}) := \sum_{[\vec{x}, \vec{x}'] \in \mathscr{P}} \sup(f|_{[\vec{x}, \vec{x}']}) \, \mathrm{vol}([\vec{x}, \vec{x}']).$$

(If the reader is not familiar with inf and sup, then these should be read as "min" and "max", although we have not assumed enough of f for these quantities to be guaranteed to exist.) We may view $\mathscr{U}(f, \mathscr{P})$ as a lower and $\mathscr{O}(f, \mathscr{P})$ as an upper approximation to the integral we aim to define.

One can now introduce what it means for one partition $\mathscr{P}'$ to be *finer* than another partition $\mathscr{P}$ and one can check that $\mathscr{U}(f, \mathscr{P}') \geq \mathscr{U}(f, \mathscr{P})$ and $\mathscr{O}(f, \mathscr{P}') \leq \mathscr{O}(f, \mathscr{P})$. One says that f is ***(Darboux-)integrable*** if

$$\sup\{\, \mathscr{U}(f, \mathscr{P}) : \text{partitions } \mathscr{P} \text{ of } Q \,\} = \inf\{\, \mathscr{O}(f, \mathscr{P}) : \text{partitions } \mathscr{P} \text{ of } Q \,\}.$$

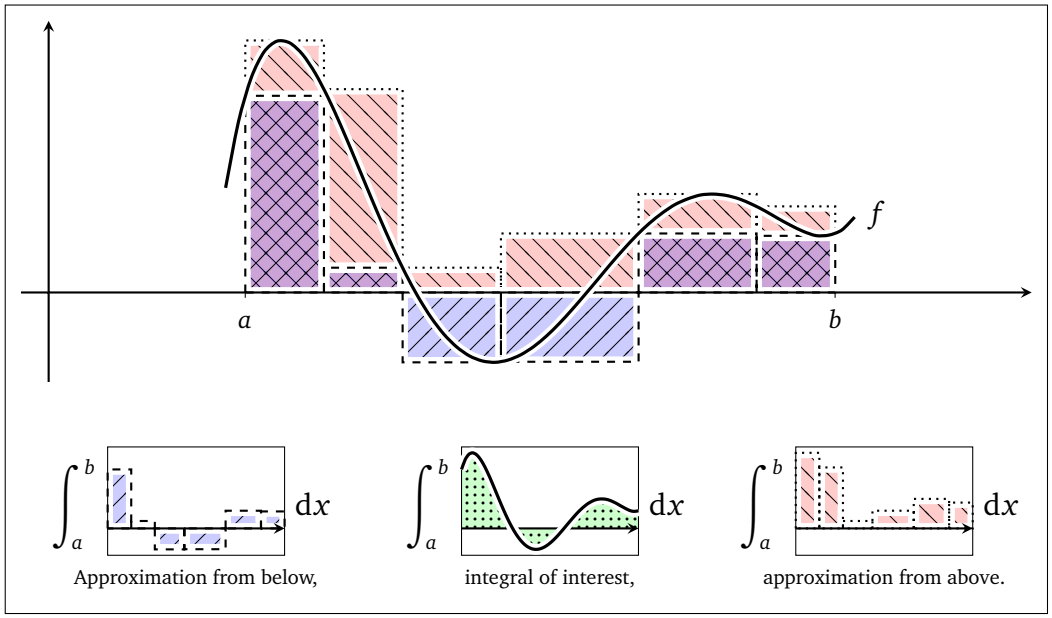


Figure 61. The Darboux integral.

In that case one defines the integral (7.1) to be equal to the above quantities.

This definition can be extended to complex-valued functions $f : Q \to \mathbb{C}$ by letting

$$\int_Q f(\vec{x})\,\mathrm{d}^n\vec{x} := \int_Q \operatorname{Re} f(\vec{x})\,\mathrm{d}^n\vec{x} + \mathrm{i}\int_Q \operatorname{Im} f(\vec{x})\,\mathrm{d}^n\vec{x}$$

provided that both $\operatorname{Re} f : Q \to \mathbb{R}$ and $\operatorname{Im} f : Q \to \mathbb{R}$ are integrable.

Now how would one define the integral of a function $g\colon U \to \mathbb{C}$ defined on some arbitrary *bounded* set $U \subseteq \mathbb{R}^n$? Easy: suppose that Q is some cuboid containing U. (Such a cuboid always exists due to our assumption that U be bounded.) Then we extend g as follows:

$$g_{\text{ext}}\colon Q \to \mathbb{C}, \quad \vec{x} \mapsto \begin{cases} g(\vec{x}) & \text{if } \vec{x} \in U, \\ 0 & \text{if } \vec{x} \in Q \setminus U. \end{cases}$$

We then simply define

$$\int_U g(\vec{x})\,\mathrm{d}^n\vec{x} := \int_Q g_{\text{ext}}(\vec{x})\,\mathrm{d}^n\vec{x}$$

provided that g_{ext} is integrable.

We may on occasion use the notation

$$\int_U g(x, y, \ldots)\,\mathrm{d}^n(x, y, \ldots) \quad \text{for} \quad \int_U g(\vec{x})\,\mathrm{d}^n\vec{x},$$

thus denoting the components of $\vec{x} = (x_1, x_2, \ldots, x_n)$ by other symbols.

There are further difficulties with finding defining the integral of a function defined on an *unbounded* domain $U \subseteq \mathbb{R}^n$ and also including unbounded functions is problematic:

Example. To illustrate some difficulties we look at the one-dimensional case. Observe that

$$\lim_{\epsilon \searrow 0} \int_\epsilon^1 \frac{1}{x}\,\mathrm{d}x = \lim_{\epsilon \searrow 0} \Big(\log x \Big|_{x=\epsilon}^1 \Big) = \lim_{\epsilon \searrow 0} (-\log \epsilon) = \infty.$$

However, for any $\delta > 0$,

$$\lim_{\epsilon \searrow 0} \int_\epsilon^1 \frac{1}{x^{1-\delta}}\,\mathrm{d}x = \lim_{\epsilon \searrow 0} \left(\frac{1}{\delta} x^\delta \Big|_{x=\epsilon}^1 \right) = \lim_{\epsilon \searrow 0} \left(\frac{1}{\delta} - \frac{1}{\delta}\epsilon^\rho \right) = \frac{1}{\delta}.$$

Hence, we may have reason to argue that the area under the graph of $(0,1] \to \mathbb{R}$, $x \mapsto \frac{1}{x^{1-\delta}}$, is finite and has the value $1/\delta$, even though the function we wish to integrate is unbounded on the interval $(0,1]$.

Remark. In higher dimensions, even the cuboids we use have more boundary points than in one dimension (intervals have at most two boundary points). From this one may glean that things become more complicated and for the purpose of setting up a viable integration theory, there is no elegant one-page fit-all solution. We do not go into this further.

Call intervals $A_1 = [a_1, b_1] \subset \mathbb{R}^1$ "good". If $A_i \subset \mathbb{R}^i$ is good in $\mathbb{R}^i$, and

$$a_{i+1}, b_{i+1} \colon A_i \to \mathbb{R}$$

are continuous functions such that $a_{i+1} \leq b_{i+1}$ everywhere on A_i, then we also call the set

$$A_{i+1} := \{ (\vec{\xi}, x_{i+1}) \in A_i \times \mathbb{R} : a_{i+1}(\vec{\xi}) \leq x_{i+1} \leq b_{i+1}(\vec{\xi}) \}$$

good in $\mathbb{R}^{i+1}$.

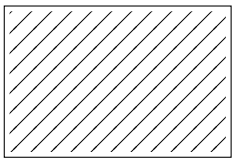
Proposition 7.1 (Fubini). *Suppose that $\Omega = A_n$ is a good set in $\mathbb{R}^n$ and $f \colon A_n \to \mathbb{C}$ is continuous. Then f is integrable and we have*

$$\int_\Omega f(\vec{x})\,\mathrm{d}^n\vec{x} = \int_{a_1}^{b_1} \int_{a_2(x_1)}^{b_2(x_1)} \cdots \int_{a_n(x_1,\ldots,x_{n-1})}^{b_n(x_1,\ldots,x_{n-1})} f(x_1,\ldots,x_n)\,\mathrm{d}x_n \ldots \mathrm{d}x_2\,\mathrm{d}x_1.$$

Note that our starting point with A_1 and then adding further coordinates "to the right" seems rather arbitrary. Indeed, it is, and Fubini's theorem also works when one starts with another coordinate and adds more coordinates to the left and to the right in any order that one chooses. To state this formally, we would have to re-number the indices in our above definition of 'good' sets. We omit this and instead simply

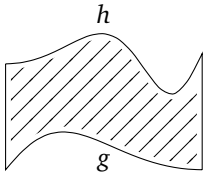
record some special cases, leaving it to the reader to imagine what the general case looks like.

(1) $\Omega = \{(x,y) \in \mathbb{R}^2 : a \le x \le b, A \le y \le B\}$:



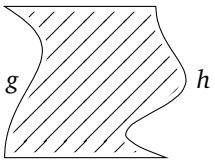
$$\int_\Omega f(\vec{x})\,\mathrm{d}^2\vec{x} = \int_a^b \int_A^B f(x,y)\,\mathrm{d}y\,\mathrm{d}x = \int_A^B \int_a^b f(x,y)\,\mathrm{d}x\,\mathrm{d}y.$$

(2) $\Omega = \{(x,y) \in \mathbb{R}^2 : a \le x \le b, g(x) \le y \le h(x)\}$:



$$\int_\Omega f(\vec{x})\,\mathrm{d}^2\vec{x} = \int_a^b \int_{g(x)}^{h(x)} f(x,y)\,\mathrm{d}y\,\mathrm{d}x.$$

(3) $\Omega = \{(x,y) \in \mathbb{R}^2 : a \le y \le b, g(y) \le x \le h(y)\}$:

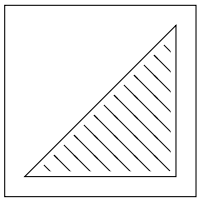


$$\int_\Omega f(\vec{x})\,\mathrm{d}^2\vec{x} = \int_a^b \int_{g(y)}^{h(y)} f(x,y)\,\mathrm{d}x\,\mathrm{d}y.$$

(In all of the above cases $g, h\colon [a,b] \to \mathbb{R}$ and $f\colon \Omega \to \mathbb{C}$ are assumed to be continuous.) Observe that we already know how to compute 1-dimensional integrals (from [2, 3] or from school). Hence, computing the above iterated integrals is also possible (computing the inner-most integral first and then moving on).

Example (Area of a triangle)**.** We shall compute the integral of $f\colon \Omega \to \mathbb{C}$, $x \mapsto 1$, defined on the triangle

$$\Omega = \{(x,y) \in \mathbb{R}^2 : 0 \le x \le 1, 0 \le y \le x\}.$$

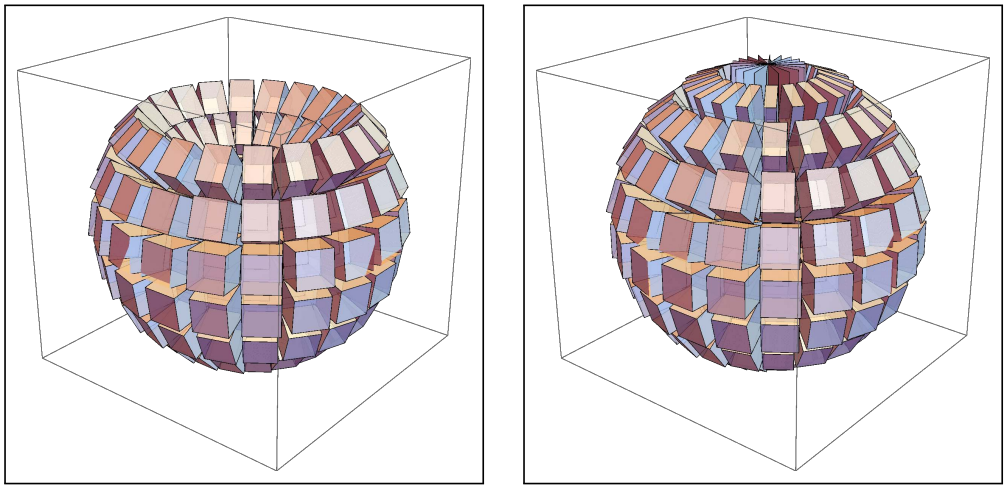


Clearly this turns out to be 1 times the area of Ω, i.e., $1/2$. Indeed, we have

$$\int_\Omega f(\vec{x})\,\mathrm{d}^2\vec{x} = \int_0^1 \int_0^x 1\,\mathrm{d}y\,\mathrm{d}x = \int_0^1 x\,\mathrm{d}x = \frac{1}{2}x^2\Big|_{x=0}^1 = \frac{1}{2} - 0 = \frac{1}{2}.$$

7.2. Transformation formula

7.2.1. The formula. We now wish to make the one-dimensional substitution rule available for higher-dimensional integration. Unfortunately, stronger assumptions are needed than one is used to from the one-dimensional case.



(a) Leaving some view inside. **(b)** Approximating a full ball.

Figure 62. Images of cuboids under $\mathrm{d}\vec{K}$, the differential of the spherical coordinate map $\vec{K}$, shifted to where $\vec{K}$ would place them.

A set $K \subseteq \mathbb{R}^n$ is called ***compact*** if it is ***bounded*** (i.e., $K \subseteq \{\,\vec{x} \in \mathbb{R}^n : \|\vec{x}\| < M\,\}$ for some suitably large $M > 0$) and ***closed***. The latter means that any ***boundary points*** of K are contained in K. Here a point is called a boundary point of K if *any* open ϵ-ball about that point contains elements of K and its complement $\mathbb{R}^n \setminus K$. The set of boundary points of K is often denoted by ∂K.

Example. Let $a < b$ be real numbers.

- Any closed interval $[a, b] \subseteq \mathbb{R}$ is compact. Its boundary points are a and b.
- Neither the open interval (a, b) nor any of the two half-open intervals $[a, b)$ or $(a, b]$ are closed.
- The "interval" $K = [a, b] \times \{0\} \subseteq \mathbb{R}^2$ is compact; in this particular case, every point of K is a boundary point. We have $\partial K = K$.
- The open ball $B = \{\,\vec{x} \in \mathbb{R}^n : \|\vec{x}\| < 1\}$ is not closed and, in particular, not compact.
- The closed ball $\overline{B} = \{\,\vec{x} \in \mathbb{R}^n : \|\vec{x}\| \leq 1\}$ is compact. Its boundary $\partial\overline{B} = \partial B$ is the sphere $S = \{\,\vec{x} \in \mathbb{R}^n : \|\vec{x}\| = 1\}$.

With the above notation, we are able to state the transformation formula.

Theorem 7.2 (Transformation formula)**.** *Let $U \subseteq \mathbb{R}^n$ be some region consisting only of inner points. Suppose further that $\vec{\Phi}\colon U \to \vec{\Phi}(U)$ is some C^1 function mapping U bijectively onto its image $\vec{\Phi}(U) \subseteq \mathbb{R}^n$ such that its inverse $\vec{\Phi}^{-1}\colon \vec{\Phi}(U) \to U$ is also C^1 and $\vec{\Phi}(U)$ is bounded. Let $f\colon \vec{\Phi}(U) \to \mathbb{C}$ be any function. Suppose that $K \subset U$ is compact. Then f is integrable on $\vec{\Phi}(K)$ if and only if $f \circ \vec{\Phi}$ is integrable on K and in that case we*

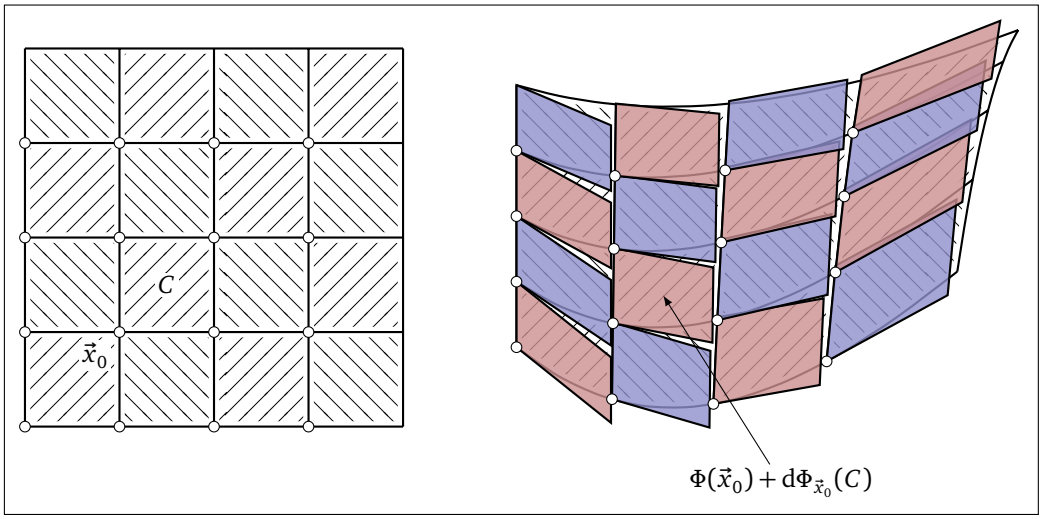


Figure 63. Mapping a square with a differentiable map $\Phi\colon \mathbb{R}^2 \to \mathbb{R}^2$ (hatched regions). The parallelograms are the images of the sub-squares under $\mathrm{d}\Phi_{\vec{x}_0}$ (shifted by $\Phi(\vec{x}_0)$), where $\vec{x}_0$ denotes the lower left corner of each sub-square.

have

$$\int_{\vec{\Phi}(K)} f(\vec{x})\,\mathrm{d}^n\vec{x} = \int_K f(\vec{\Phi}(\vec{u}))|\det J_{\vec{\Phi}}(\vec{u})|\,\mathrm{d}^n\vec{u}.$$

Proof (sketch). The details of this proof are well-beyond the scope of this course. Suffice it to say that one can argue that we may restrict to f being continuous and bounded and U being a cuboid. Suppose then the formula was wrong. By halving U in every dimension, we may split U into 2^n smaller cuboids and the formula must also be violated in *at least one* of them (in a sense that one must make precise in order for this argument to go through, but we gloss over this). By repeating this process, one may assume that the formula is violated on a very small cuboid C (again, "violated" in a precise sense). However, in this situation, for any point $\vec{x}_0 = \vec{\Phi}(\vec{c}_0) \in \vec{\Phi}(C)$,

$$\int_{\vec{\Phi}(C)} f(\vec{x})\,\mathrm{d}^n\vec{x} \approx f(\vec{x}_0)\int_{\vec{\Phi}(C)} \mathrm{d}^n\vec{x} =: f(\vec{x}_0)\,\mathrm{vol}(\vec{\Phi}(C)) = f(\vec{\Phi}(\vec{c}_0))\,\mathrm{vol}(\vec{\Phi}(C)),$$

and (using that $\vec{\Phi}$ is a C^1 function, thus making $\det J_\Phi(\,\cdot\,)$ continuous)

$$\begin{aligned}\int_C f(\vec{\Phi}(\vec{u}))|\det J_{\vec{\Phi}}(\vec{u})|\,\mathrm{d}^n\vec{u} &\approx f(\vec{\Phi}(\vec{c}_0))|\det J_{\vec{\Phi}}(\vec{c}_0)|\int_C \mathrm{d}^n\vec{u}\\ &= f(\vec{\Phi}(\vec{c}_0))|\det J_{\vec{\Phi}}(\vec{c}_0)|\,\mathrm{vol}(C).\end{aligned}$$

Now how do $\mathrm{vol}(\vec{\Phi}(C))$ and $\mathrm{vol}(C)$ relate? First, $\vec{\Phi}$ is well-approximated at every $\vec{c} \in C$ by its linearisation $\mathrm{d}\vec{\Phi}_{\vec{c}}$ at $\vec{c}$. Moreover, since $\vec{\Phi}$ is continuously differentiable

and C is very small, we may approximate $\mathrm{d}\vec{\Phi}_{\vec{c}}$ (and, thus, $\vec{\Phi}$) by $\mathrm{d}\vec{\Phi}_{\vec{c}_0}$ throughout C. Hence,

$$\mathrm{vol}(\vec{\Phi}(C)) \approx \mathrm{vol}(\mathrm{d}\vec{\Phi}_{\vec{c}_0}(C)) = \mathrm{vol}(C) \cdot |\det(\mathrm{d}\vec{\Phi}_{\vec{c}_0})| = \mathrm{vol}(C) \cdot |\det(J_{\vec{\Phi}}(\vec{c}_0))|.$$

Hence

$$\int_{\vec{\Phi}(C)} f(\vec{x})\,\mathrm{d}^n\vec{x} \approx \int_C f(\vec{\Phi}(\vec{u}))|\det J_{\vec{\Phi}}(\vec{u})|\,\mathrm{d}^n\vec{u}.$$

If, however, one sets up the above "violated in a precise sense" correctly, then this last approximation gives a contradiction. Hence, the transformation formula holds for arbitrary cuboids, as desired. □

Note that $\det J_{\vec{\Phi}}(\vec{u})$ makes sense only because $J_{\vec{\Phi}}(\vec{u})$ is a *square* matrix. This is ensured in the above, because Theorem 7.2 both U and its image $\vec{\Phi}(U)$ under $\vec{\Phi}$ are assumed to be contained in $\mathbb{R}^n$. However, we can also write

$$|\det J_{\vec{\Phi}}(\vec{u})| = \sqrt{\det(J_{\vec{\Phi}}(\vec{u})^\intercal J_{\vec{\Phi}}(\vec{u}))}.$$

The latter term makes sense even when $J_{\vec{\Phi}}(\vec{u})$ is non-square and also in this case it has an interpretation as a measure for the distortion of volume furnished by applying the linear map induced by $J_{\vec{\Phi}}(\vec{u})$ (see Proposition 3.4). This suggests that we introduce the following definition. Let $\vec{\Phi}\colon U \to \vec{\Phi}(U)$ be as in Theorem 7.2, apart from allowing now that $\vec{\Phi}(U) \subseteq \mathbb{R}^m$ for some $m \geq n$. Then

$$\int \cdots \int_{\vec{\Phi}(K)} f(\vec{x})\,\mathrm{d}V(\vec{x}) := \int_K f(\vec{\Phi}(\vec{u}))\sqrt{\det(J_{\vec{\Phi}}(\vec{u})^\intercal J_{\vec{\Phi}}(\vec{u}))}\,\mathrm{d}^n\vec{u}.$$

The symbol $\mathrm{d}V(\vec{x})$ is called the (n-dimensional) ***volume element***. In low dimensions, this has more specialised names.

- $n = 1$: $\displaystyle\int_{\vec{\Phi}(K)} f(\vec{x})\,\mathrm{d}s(\vec{x})$ with the ***line element*** $\mathrm{d}s$.
- $n = 2$: $\displaystyle\iint_{\vec{\Phi}(K)} f(\vec{x})\,\mathrm{d}A(\vec{x})$ with the ***area element*** $\mathrm{d}A$.
- $n = 3$: $\displaystyle\iiint_{\vec{\Phi}(K)} f(\vec{x})\,\mathrm{d}V(\vec{x})$ with the ***volume element*** $\mathrm{d}V$ (as before).

The area element is also sometimes called ***surface element***. (In the literature one often also finds notation such as $\mathrm{d}S$ or $\mathrm{d}F$ for the area element. We shall not employ these.)

Sometimes we also omit the argument $\vec{x}$ in both the function being integrated and the respective volume element and write plainly

$$\int_{\vec{\Phi}(K)} f\,\mathrm{d}s, \quad \iint_{\vec{\Phi}(K)} f\,\mathrm{d}A \quad \text{or} \quad \iiint_{\vec{\Phi}(K)} f\,\mathrm{d}V \quad \text{respectively.}$$

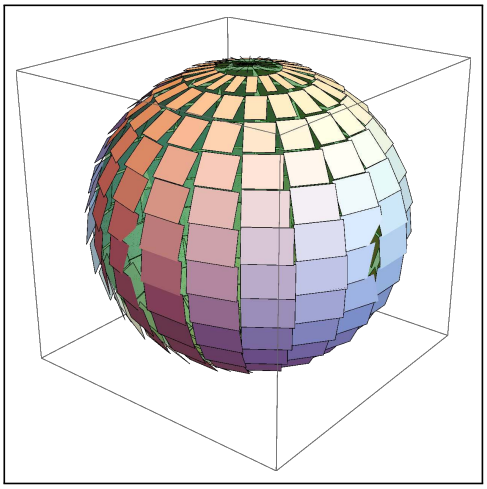


Figure 64. Images of rectangles under $\mathrm{d}\vec{K}(1,\cdot,\cdot)$, the differential of the spherical coordinate map $(\theta,\varphi)\mapsto\vec{K}(1,\theta,\varphi)$, shifted to where $\vec{K}$ would place them.

Remark (On varying notation)**.** Notation for integrals varies wildly throughout the literature: pick up any two books discussing integrals in dimensions greater than 1 and you are very likely to find utterly different notation.

Remark (On the relevance of having a parametrisation)**.** Note that the integrals we study here always come with a parametrisation of the object on which the integration is being performed. The question as to how much these integrals are intrinsic to the *object* itself and not just the chosen *parametrisation* thereof is a subtle one and we shall not venture into that territory.

Remark (On abuse of Theorem 7.2)**.** Below we shall frequently apply Theorem 7.2 with non-compact K and also work in situations where the $\vec{\Phi}$ used does not really satisfy all of the above requirements either. The prime example for this is when working with polar coordinates

$$\vec{P}\colon\mathbb{R}^2\to\mathbb{R}^2,\quad (r,\varphi)\mapsto\begin{pmatrix}r\cos\varphi\\ r\sin\varphi\end{pmatrix}$$

(or cylindrical or spherical coordinates). To parametrise the disk $\{\vec{x}\in\mathbb{R}^2:\|\vec{x}\|\leq 1\}$, one would restrict $\vec{P}$ to $K=[0,1]\times[0,2\pi)$. However, this set K is not compact (any point $(r,2\pi)$ with $0\leq r\leq 1$ belongs to ∂K but not to K, so K is not closed). Moreover, any point $(0,\varphi)$ gets mapped to $(0,0)$ by $\vec{P}$, so $\vec{P}$ is not injective on K. This is, however, not problematic, as this undesired behaviour happens on a set which is negligible with respect to integration. Formally, one could restrict $\vec{P}$ to $K_\epsilon=[\epsilon,1]\times[0,2\pi-\epsilon]$, compute the desired integral using $\vec{P}|_{K_\epsilon}$ and then take $\epsilon\searrow 0$, but we will never bother to explicitly write this like that.

Example 7.3. We wish to compute $I = \int_{-\infty}^{\infty} \exp(-x^2)\,\mathrm{d}x$. We have

$$I^2 = \int_{-\infty}^{\infty} \exp(-x_1^2)\,\mathrm{d}x_1 \int_{-\infty}^{\infty} \exp(-x_2^2)\,\mathrm{d}x_2 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp(-(x_1^2+x_2^2))\,\mathrm{d}x_1\,\mathrm{d}x_2.$$

This equals the integral $\int_{\mathbb{R}^2} \exp(-\|\vec{x}\|^2)\,\mathrm{d}^2\vec{x}$, and using polar coordinates (justified via Theorem 7.2), we find that

$$\begin{aligned} I^2 &= \int_{\mathbb{R}_+\times[0,2\pi)} \exp(-r^2) r\,\mathrm{d}^2(r,\varphi) = \int_0^{\infty}\int_0^{2\pi} \exp(-r^2) r\,\mathrm{d}\varphi\,\mathrm{d}r \\ &= -2\pi\frac{1}{2}\exp(-r^2)\Big|_{r=0}^{\infty}. \end{aligned}$$

As the last term equals π and clearly $I \geq 0$ (the integrand $\exp(-x^2)$ is positive), we find that $I = \sqrt{\pi}$.

The integrals just defined can be used to compute lengths of curves, areas etc. of parametrised objects. We shall illustrate this in the following subsections.

7.2.2. Lengths of curves. For instance, suppose that we have a parametrisation[1] $\vec{\gamma}\colon [0,1] \to \mathbb{R}^2$, $t \mapsto (\gamma_1(t), \gamma_2(t))$ of a curve in two dimensions. We assume here that γ is injective (the curve does not self-intersect) and γ is "sufficiently nice" (differentiable, etc.[2]). Then

$$\text{length of } \vec{\gamma}([0,1]) = \int_{\vec{\gamma}([0,1])} 1\,\mathrm{d}s = \int_0^1 \sqrt{\det(J_{\vec{\gamma}}(t)^{\mathsf{T}} J_{\vec{\gamma}}(t))}\,\mathrm{d}t.$$

Observe that

$$J_{\vec{\gamma}}(t) = \begin{pmatrix}\dot{\gamma}_1(t)\\ \dot{\gamma}_2(t)\end{pmatrix}, \quad \begin{pmatrix}\dot{\gamma}_1(t)\\ \dot{\gamma}_2(t)\end{pmatrix}^{\mathsf{T}} \begin{pmatrix}\dot{\gamma}_1(t)\\ \dot{\gamma}_2(t)\end{pmatrix} = \left(\dot{\gamma}_1(t)^2 + \dot{\gamma}_2(t)^2\right) \in \mathbb{R}^{1\times 1}.$$

Because the determinant of a 1×1-matrix is just the entry of that matrix we have

$$\text{length of } \vec{\gamma}([0,1]) = \int_0^1 \sqrt{\dot{\gamma}_1(t)^2 + \dot{\gamma}_2(t)^2}\,\mathrm{d}t.$$

The formula generalises easily to higher dimensions: just replace $\dot{\gamma}_1(t)^2 + \dot{\gamma}_2(t)^2$ by

$$\dot{\gamma}_1(t)^2 + \ldots + \dot{\gamma}_m(t)^2$$

[1] The attentive reader will note 0 and 1 are not inner points of $U = [0,1] \subset \mathbb{R}$, so strictly speaking Theorem 7.2 does not apply. There are several ways around this problem. We shall just ignore this, however.

[2] As we are trying to avoid introducing too many topological terms, stating all required assumptions is becoming increasingly bothersome. Our solution here: we just plainly omit them and trust that the reader will only apply the stated "results" in sufficiently well-behaved situations.

if $\vec{\gamma}$ maps into $\mathbb{R}^m$ via $t \mapsto (\gamma_1(t), \dots, \gamma_m(t))$.

Example (Length of a line segment)**.** We look at the curve $\vec{\gamma}\colon [0,1] \to \mathbb{R}^2$, $t \mapsto (3t, 2t)$. Its image is a straight line segment connecting the points $\vec{\gamma}(0) = (0,0)$ and $\vec{\gamma}(1) = (3,2)$. Therefore, its length is

$$\|(3,2) - (0,1)\| = \|(3,2)\| = \sqrt{3^2 + 2^2} = \sqrt{13}.$$

As ought to be expected, the formula we have derived reproduces this answer:

$$\text{length of } \vec{\gamma}([0,1]) = \int_0^1 \sqrt{3^2 + 2^2}\, \mathrm{d}t = \sqrt{13}.$$

Example (Length of a parabola segment)**.** We want to compute the length of the parabola $\vec{\gamma}\colon [-1,1] \to \mathbb{R}^2, t \mapsto (t, t^2)$. Here

$$\text{length of } \vec{\gamma}([-1,1]) = \int_{-1}^1 \sqrt{1^2 + (2t)^2}\, \mathrm{d}t = \int_{-1}^1 \sqrt{1 + 4t^2}\, \mathrm{d}t.$$

Computing the last integral takes a bit of effort. We sidestep this by noting that an antiderivative to the integrand is

$$F(t) = \tfrac{1}{2} t \sqrt{4t^2 + 1} + \tfrac{1}{4} \sinh^{-1}(2t),$$

where $\sinh^{-1}$ is the inverse function of $\sinh\colon \mathbb{R} \to \mathbb{R}$, $x \mapsto \frac{1}{2}(\exp(x) - \exp(-x))$. Using this, one can check that

$$\text{length of } \vec{\gamma}([-1,1]) = F(1) - F(-1) = \sqrt{5} + \tfrac{1}{2} \sinh^{-1}(2).$$

7.2.3. Areas of surfaces. We generalise our consideration of *lengths of curves* to *areas of surfaces*. Specifically, we shall assume that we are given a differentiable function $f : \mathbb{R}^2 \to \mathbb{R}$ and we ask for the area of its graph

$$\{ (u_1, u_2, f(u_1, u_2)) \in \mathbb{R}^3 : 0 \le u_1, u_2 \le 1 \}$$

above the square $[0,1]^2$. This graph is obviously parametrised by

$$\vec{\Phi}\colon [0,1]^2 \to \mathbb{R}^3, \quad (u_1, u_2) \mapsto (u_1, u_2, f(u_1, u_2)).$$

We have

$$J_{\vec{\Phi}}(u_1, u_2) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ \partial_1 f(u_1, u_2) & \partial_2 f(u_1, u_2) \end{pmatrix} =: \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ a & b \end{pmatrix}.$$

Then a quick calculation shows that

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \\ a & b \end{pmatrix}^{\mathsf{T}} \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ a & b \end{pmatrix} = \begin{pmatrix} a^2 + 1 & ab \\ ab & b^2 + 1 \end{pmatrix},$$

and the determinant of this turns out to be $1 + a^2 + b^2$. Hence, the area of the graph of $f|_{[0,1]^2}$ is given by the formula

$$\iint_{\Phi([0,1]^2)} 1 \, \mathrm{d}A = \int_{[0,1]^2} \sqrt{1 + (\partial_1 f(\vec{u}))^2 + (\partial_2 f(\vec{u}))^2} \, \mathrm{d}^2\vec{u}.$$

Instead of applying just this formula, we give an example of a more involved area computation.

Example (Surface area of a sphere)**.** We want to compute the surface area of a sphere. Recall the "spherical coordinates map":

$$\vec{K}\colon \mathbb{R}^3 \to \mathbb{R}^3, \quad (r, \theta, \varphi) \mapsto \begin{pmatrix} r\cos(\varphi)\sin\theta \\ r\sin(\varphi)\sin\theta \\ r\cos\theta \end{pmatrix}.$$

To parametrise the sphere with radius 1, we shall look at

$$\vec{\Phi}\colon (0,\pi) \times [0, 2\pi) \to \{\, \vec{x} \in \mathbb{R}^3 : \|\vec{x}\| = 1,\, x_3 \neq \pm 1 \,\}, \quad (\theta, \varphi) \mapsto \vec{K}(1, \theta, \varphi).$$

(Once more, ignore any problems with (non-)inner points of the domain of definition and the fact that our image "sphere" is missing the points $(0, 0, \pm 1)$.) Then the area of our "sphere" is given by

$$a = \int_{(0,\pi)\times[0,2\pi)} \sqrt{\det(J_{\vec{\Phi}}(\vec{u})^\intercal J_{\vec{\Phi}}(\vec{u}))} \, \mathrm{d}^2\vec{u}.$$

We have

$$J_{\vec{\Phi}}(\theta, \varphi) = \begin{pmatrix} \cos(\varphi)\cos\theta & -\sin(\varphi)\sin\theta \\ \sin(\varphi)\cos\theta & \cos(\varphi)\sin\theta \\ -\sin\theta & 0 \end{pmatrix},$$

and (after applying $\cos(x)^2 + \sin(x)^2 = 1$ a couple of times)

$$J_{\vec{\Phi}}(\theta, \varphi)^\intercal J_{\vec{\Phi}}(\theta, \varphi) = \begin{pmatrix} 1 & 0 \\ 0 & \sin(\theta)^2 \end{pmatrix}.$$

Hence,

$$\begin{aligned} a &= \int_{(0,\pi)\times[0,2\pi)} \sqrt{\sin(u_1)^2} \, \mathrm{d}^2\vec{u} = \int_0^\pi \int_0^{2\pi} |\sin(u_1)| \, \mathrm{d}u_2 \, \mathrm{d}u_1 = 2\pi \int_0^\pi \sin(\theta) \, \mathrm{d}\theta \\ &= 2\pi(-\cos(\theta))\Big|_{\theta=0}^{\pi} = 2\pi(-\cos(\pi) + \cos(0)) = 4\pi. \end{aligned}$$

Hence, the area of a sphere with radius 1 is 4π. (We already know this from geometry class in school, but it is good to see that our formulas provide the same answer.)

7.3. Integrating against vector fields

7.3.1. Review of integration so far. The sort of integrals of lines, surfaces and volumes in (potentially higher-dimensional) space are of the shape

$$\int_{\text{curve}} f(\vec{x})\,\mathrm{d}s(\vec{x}), \quad \iint_{\text{surface}} g(\vec{x})\,\mathrm{d}A(\vec{x}), \quad \iiint_{\text{solid}} h(\vec{x})\,\mathrm{d}V(\vec{x}),$$

where the domain of integration comes equipped with a parametrisation. Integration here works by "pushing" the known n-dimensional (Darboux-)integral on n-dimensional space onto our object of interest by means of the parametrisation. Up until now, the above functions f, g and h that we wish to integrate came *defined on* the curve, surface or solid of interest and may be thought of as attaching some "density" to the points of the objects under consideration. (Incidentally, in all of the above examples, these functions were chosen to be constant, so our objects so-far were considered to be "uniformly dense".) In this vein, one may think of

$$\int_{\text{curve}} f(\vec{x})\,\mathrm{d}s(\vec{x})$$

as computing the "mass" of the curve under consideration:

$$\int_{\text{curve}} f(\vec{x})\,\mathrm{d}s(\vec{x}) \triangleq \lim_{\substack{\text{chunk-}\\ \text{size}\searrow 0}} \sum_{\substack{\text{chunks } \mathscr{C}\\ \text{of the curve}}} \underbrace{(\text{density of the chunk } \mathscr{C})}_{\text{unit: mass per length}} \times \underbrace{(\text{length of chunk } \mathscr{C})}_{\text{unit: length}}.$$

(Similar interpretations can be attached to the integrals $\iint$ and $\iiint$.)

7.3.2. Surface integral over a vector field. In physics, one often faces a different situation. One is working in space (say, $\mathbb{R}^3$ for simplicity) and this space is interspersed by "fields". These could be electric, magnetic or other force fields. They are modelled by attaching to every point $\vec{x} \in \mathbb{R}^3$ a vector $\vec{K}(\vec{x}) \in \mathbb{R}^3$. We give an interpretation related to Chemistry. Imagine $\mathbb{R}^3$ (or a suitable subset thereof) to be filled with molecules and suppose that all of these molecules are moving at constant speed in the direction of $\vec{e}_1 = (1,0,0)$. Thinking of the molecules as bullets being fired from the $\vec{e}_2\vec{e}_3$-plane towards the $\vec{e}_1$-direction and holding a sheet of paper (which we imagine to be infinitely thin) in space, we wonder how many bullet will hit the sheet.

If $\vec{v}$ and $\vec{w}$ are vectors whose length and direction matches the orientation of two adjacent sides of our sheet, then $\vec{v} \times \vec{w}$ is a vector whose length equals the area of our sheet and its direction is normal to the surface of the sheet. (Recall also, that the orientation of that normal can be inferred using the "right hand rule" provided that our coordinate system is oriented suitably.) A geometric consideration reveals that the number of molecules (bullets) passing through is roughly

$$\vec{e}_1 \cdot (\vec{v} \times \vec{w}) = \vec{K}(\vec{x}) \cdot (\vec{v} \times \vec{w}).$$

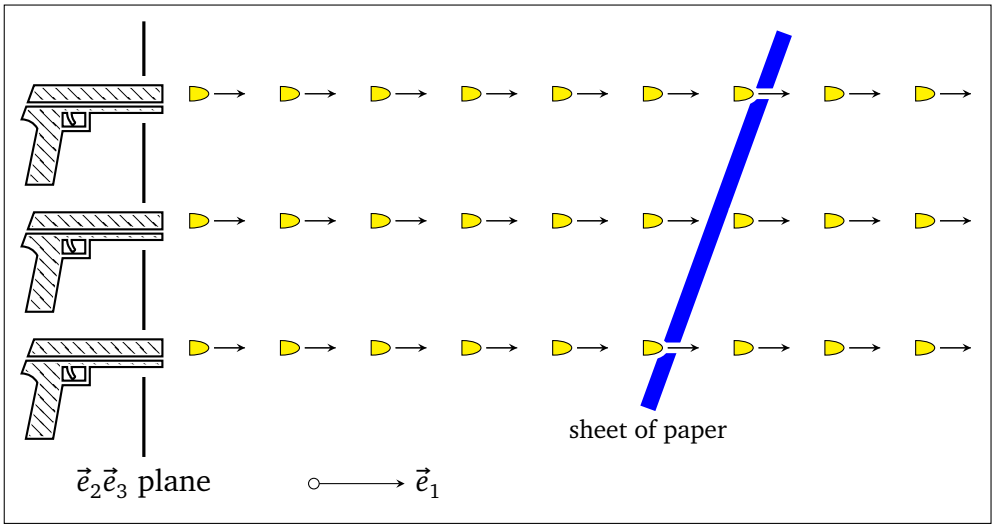


Figure 65. Illustration of bullets hitting a sheet of paper. Note that two of three spacial dimensions have been collapsed into one.

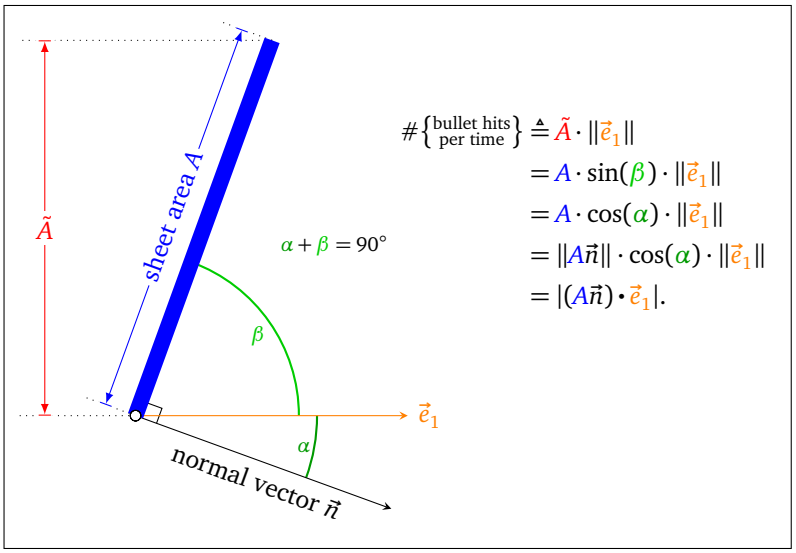


Figure 66. Illustration showing how the orientation of the sheet of paper affects the number of bullets hitting it in our thought experiment.

Now let $\vec{x}\colon U \to \vec{x}(U)$ parametrise a surface $S = \vec{x}(U) \subset \mathbb{R}^3$, where $U \subseteq \mathbb{R}^2$ is supposed to be some parameter space and the usual assumption that our parametrisation be "sufficiently nice" are always in place. The above considerations lead us to define

$$\iint_S \vec{K}\cdot\mathrm{d}\vec{A} := \int_U (\vec{K}\circ\vec{x})\cdot\left(\frac{\partial\vec{x}}{\partial u_1}\times\frac{\partial\vec{x}}{\partial u_2}\right)\mathrm{d}^2\vec{u}.$$

Note that this can be understood as an ordinary surface integral. Indeed,[3]

$$\left\| \frac{\partial \vec{x}}{\partial u_1} \times \frac{\partial \vec{x}}{\partial u_2} \right\| = |\det J_{\vec{x}}|.$$

We now consider the normal vector field[4]

$$\vec{n}_U = \left\| \frac{\partial \vec{x}}{\partial u_1} \times \frac{\partial \vec{x}}{\partial u_2} \right\|^{-1} \left(\frac{\partial \vec{x}}{\partial u_1} \times \frac{\partial \vec{x}}{\partial u_2} \right).$$

This attaches to every point $\vec{u} \in U$ of the parameter space a vector $\vec{n}_U(\vec{u})$ normal to surface S at the point $\vec{x}(\vec{u})$. Similarly, $\vec{n} = \vec{n}_U \circ x^{-1}$ does the same thing, although now the function argument is given by points on S. We may write

$$(\vec{K} \circ \vec{x}) \cdot \left(\frac{\partial \vec{x}}{\partial u_1} \times \frac{\partial \vec{x}}{\partial u_2} \right) = ((\vec{K} \circ \vec{x}) \cdot \vec{n}_U) |\det J_{\vec{x}}| = ((\vec{K} \circ \vec{x}) \cdot (\vec{n} \circ \vec{x})) |\det J_{\vec{x}}|.$$

Hence,

$$\iint_S \vec{K} \cdot \mathrm{d}\vec{A} = \int_U ((\vec{K} \circ \vec{x}) \cdot (\vec{n} \circ \vec{x})) |\det J_{\vec{x}}| \, \mathrm{d}^2\vec{u} = \iint_S (\vec{K} \cdot \vec{n}) \, \mathrm{d}A,$$

where close attention ought to be paid to the notation $\mathrm{d}\vec{A}$ and $\mathrm{d}A$. Here $\mathrm{d}A$ is the usual area element and $\mathrm{d}\vec{A}$ is called the ***vectorial area element***.

7.3.3. Orientation dependence. We would like to draw the reader's attention to one curious phenomenon: the value of

$$\vec{K} \cdot \vec{n} = \|\vec{K}\| \, \|\vec{n}\| \cos \sphericalangle(\vec{K}, \vec{n})$$

does *not only* depend on $\vec{K}$ and the length $\|\vec{n}\|$ of the normal, but also its *orientation*. Suppose, for instance, that we switch the roles of u_1 and u_2 in the parametrisation. That is, we consider

$$\circlearrowleft \colon \mathbb{R}^2 \to \mathbb{R}^2, \quad (u_2, u_1) \mapsto (u_1, u_2),$$

as well as

$$\vec{x}_{\circlearrowleft} \colon \circlearrowleft(U) \to \vec{x}(U), \quad (u_2, u_1) \mapsto \vec{x}(u_1, u_2) = (\vec{x} \circ \circlearrowleft)(u_2, u_1).$$

Then the "new" integral[5] of $\vec{K}$ over S turns out to be

$$\iint_S^{(\mathrm{new})} \vec{K} \cdot \mathrm{d}\vec{A} := \int_{U_{\circlearrowleft}} (\vec{K} \circ \vec{x}_{\circlearrowleft}) \cdot (\ldots) \, \mathrm{d}^2\vec{u},$$

[3] Attention: this is an equality of *functions*.

[4] Here we are tacitly assuming that $\frac{\partial \vec{x}}{\partial u_1} \times \frac{\partial \vec{x}}{\partial u_2}$ does not vanish. A famous theorem of Sard–Morse could be used to see that this assumption is always "sufficiently" guaranteed by our above assumptions as to cause no problems with the integration. Roughly: the vanishing of the vector in question is a sufficiently seldom event and would not affect the integration if we had used a sufficiently robust integral. We do not discuss this further and just ignore this point.

[5] That is, the integral computed according to the definition, but using the parametrisation $\vec{x}_{\circlearrowleft}$ instead of $\vec{x}$.

where the "…" term is given by

$$\partial_1\vec{x}_{\circlearrowleft} \times \partial_2\vec{x}_{\circlearrowleft} = (\partial_2\vec{x} \times \partial_1\vec{x}) \circ \circlearrowleft = -(\partial_1\vec{x} \times \partial_2\vec{x}) \circ \circlearrowleft.$$

(Here we have used the chain rule.) Hence, by the transformation formula (Theorem 7.2),

$$\begin{aligned}\iint_S^{(\text{new})} \vec{K} \cdot \mathrm{d}\vec{A} &= \int_{\circlearrowleft(U)} (\vec{K} \circ \vec{x} \circ \circlearrowleft) \cdot (-(\partial_1\vec{x} \times \partial_2\vec{x}) \circ \circlearrowleft)\, \mathrm{d}^2\vec{u} \\ &= \int_U (\vec{K} \circ \vec{x}) \cdot (-(\partial_1\vec{x} \times \partial_2\vec{x}))\, \mathrm{d}^2\vec{u} \\ &= -\iint_S^{(\text{old})} \vec{K} \cdot \mathrm{d}\vec{A}.\end{aligned}$$

We come to a big **WARNING**: the surface integral over a vector field depends on the orientation of the surface. Here "orientation" is understood as the direction in which the normal vector field points; more precisely, any two parametrisations of the same surface are said to induce the same orientation on that surface if the induced normal vector fields ($\vec{n}$ above as a function on the surface) differ only by a *positive* factor (which is allowed to depend on the points of the surface).

7.3.4. Line integral over a vector field. Now let $\vec{\gamma}\colon [a,b] \to \vec{\gamma}([a,b])$ parametrise a curve $C = \vec{\gamma}([a,b]) \subset \mathbb{R}^m$ (suitably differentiable, bounded etc.). We define

$$\int_C \vec{K} \cdot \mathrm{d}\vec{s} := \int_a^b (\vec{K} \circ \vec{\gamma})(t) \cdot (\dot{\vec{\gamma}}(t))\, \mathrm{d}t.$$

Observe that $\dot{\vec{\gamma}} = J_{\vec{\gamma}}$

$$\dot{\vec{\gamma}} = \frac{\dot{\vec{\gamma}}}{\|\dot{\vec{\gamma}}\|}\|\dot{\vec{\gamma}}\| = \frac{\dot{\vec{\gamma}}}{\|\dot{\vec{\gamma}}\|}\sqrt{\dot{\vec{\gamma}}^{\mathsf{T}}\dot{\vec{\gamma}}} = \frac{\dot{\vec{\gamma}}}{\|\dot{\vec{\gamma}}\|}\sqrt{J_{\vec{\gamma}}^{\mathsf{T}} J_{\vec{\gamma}}}.$$

Therefore, analogously to what we did with surfaces, one finds that

$$\int_C \vec{K} \cdot \mathrm{d}\vec{s} = \int_C \vec{K} \cdot (\ldots)\, \mathrm{d}s,$$

where the term "…" is given by a normalised version of $\dot{\vec{\gamma}}$ lifted up to be defined on C rather than the parameter interval $[a,b]$. We leave the remaining details to the reader.

The integral

$$\int_C \vec{K} \cdot \mathrm{d}\vec{s}$$

also admits a physical interpretation as computing the work done by moving an object from $\vec{\gamma}(a)$ to $\vec{\gamma}(b)$ along the curve $C = \vec{\gamma}(U)$. (Work is computed as force times length.)

One can reverse the parametrisation by letting

$$\vec{\gamma}_-\colon [a,b] \to C, \quad t \mapsto \vec{\gamma}(a+b-t).$$

Intuitively, if one imagines $\vec{\gamma}(t)$ as the position of someone travelling on the curve C at time t, then this person travels from $\vec{\gamma}(a)$ to $\vec{\gamma}(b)$. On the other hand, $\vec{\gamma}_-$ describes the path of someone starting at $\vec{\gamma}(b)$ and travelling "backwards" to $\vec{\gamma}(a)$. Then one can show that the integral

$$\int_C \vec{K} \cdot \mathrm{d}\vec{s}$$

computed via $\vec{\gamma}$ and computed via $\vec{\gamma}_-$ flips its sign. (Incidentally, this also matches the physical interpretation of this integral: moving an object along the opposite direction of the same curve in the same force field takes the same work, yet with opposite sign.)

If $\vec{K}$ is obtained by taking the gradient of a function, then line integrals along $\vec{K}$ are easy to compute. They depend only on the end points of the line:

Theorem 7.4. *Let f be a differentiable function defined on a set containing the curve $C = \vec{\gamma}([a,b]) \subset \mathbb{R}^m$. Then*

$$\int_C (\operatorname{grad} f) \cdot \mathrm{d}\vec{s} = f(\vec{\gamma}(b)) - f(\vec{\gamma}(a)).$$

Proof. We let $h \coloneqq f \circ \vec{\gamma}$. Then

$$\int_C (\operatorname{grad} f) \cdot \mathrm{d}\vec{s} = \int_a^b ((\operatorname{grad} f) \circ \vec{\gamma})(t) \cdot \dot{\vec{\gamma}}(t)\,\mathrm{d}t = \int_a^b h'(t)\,\mathrm{d}t.$$

The formula claimed by the theorem now follows via an application of the fundamental theorem of calculus. □

Assuming that one is interested in computing

$$\int_C \vec{K} \cdot \mathrm{d}\vec{s}$$

to begin with, then Theorem 7.4 would suggest that one should be interested in finding some f such that $\vec{K} = \operatorname{grad} f$, for then the above integral is easy to compute by evaluating f at the end points of C. In physics, such an f (or $-f$ due to some different sign convention) is usually called a ***potential*** of the vector field $\vec{K}$. Unfortunately, finding such an f may not always be possible. If $\vec{K}$ is a C^1 vector field and $\vec{K} = \operatorname{grad} f$, then

$$\operatorname{rot} \vec{K} = \operatorname{rot} \operatorname{grad} f = \text{constant zero vector field.}$$

Hence, $\operatorname{rot}\vec{K}$ being the constant zero vector field is certainly *necessary* for a C^1 vector field $\vec{K}$ to admit a potential f. Under additional hypotheses on the topology (read: shape) of the domain of definition of $\vec{K}$, this necessary condition can also be turned into a sufficient condition. We omit the details.

7.4. Integral theorems

In this section we continue our tradition of shamelessly glossing over some major technical issues. (Well, perhaps some shame was felt.)

7.4.1. The 1-dimensional case. Recall the fundamental theorem of calculus in one dimension:

$$f(b)-f(a)=\int_a^b f'(x)\,\mathrm{d}x\,.$$

Note that the total differential $\mathrm{d}f_{x_0}$ is the linear map $\mathbb{R}^1\to\mathbb{R}^1$, $v\mapsto f(x_0)v$. Similarly, viewing x as the function $[a,b]\to[a,b]$, $\xi\mapsto\xi$, the differential $\mathrm{d}x_{x_0}$ is the linear map $\mathbb{R}^1\to\mathbb{R}^1$, $v\mapsto v$. Hence,

$$\mathrm{d}f_{x_0}=f'(x_0)\,\mathrm{d}x_{x_0},$$

or, more plainly,

$$\mathrm{d}f=f'\,\mathrm{d}x.$$

In particular, we may write the fundamental theorem as

$$\underbrace{\int_{\partial[a,b]} f}_{\text{0-dimensional integral}} \quad = \quad \underbrace{\int_{[a,b]} \mathrm{d}f}_{\text{1-dimensional integral}},$$

where we think of $\int_{\partial[a,b]} f := f(b)-f(a)$ as the integral of f over the boundary $\partial[a,b]=\{a,b\}$ of $[a,b]$. Note that the values $f(b)$ and $f(a)$ have signs attached to them ("+" and "−"); this may be viewed as noting that the boundary $\partial[a,b]$ has some sort of "orientation": the point a "knows" that it is to the left ("−") of the thing of which it is a boundary point and, similarly, the point b "knows" that it bounding from the right ("+").

7.4.2. Orientation on the boundary. The previous considerations suggest that if we are to generalise the fundamental theorem to some result relating

$$\int_{\partial\ldots}\ldots \quad\text{with}\quad \iint_{\ldots}\ldots, \qquad\text{and}\qquad \iint_{\partial\ldots}\ldots \quad\text{with}\quad \iiint_{\ldots}\ldots,$$

then we ought to pay attention to the orientation. We shall sketch one way of doing this which works in practice.

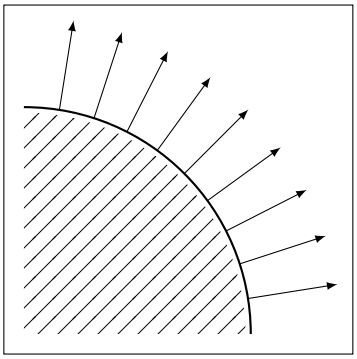


Figure 67. Two-dimensional depiction of vectors pointing *outside* of some solid. (To get a three-dimensional version, imagine the picture as part of a cross-section of a filled-in cylinder.)

Suppose you have some solid with boundary that you can parametrise. Then you ought to parametrise the boundary such that the corresponding normal field points outside of the solid (cf. Figure 67).

Suppose now that you have some surface in $\mathbb{R}^3$ that bounds some solid. Then you ought to parametrise that surface such that the normal vector field produced by that parametrisation points *outside* of the solid.

In practice, proceed as follows: start by picking *any* parametrisation $\vec{x}\colon (u_1, u_2) \mapsto \ldots$ of your surface. Now check if the produced normal vector field points in the correct direction (see the description below). If it does, you are done. If it does not, simply reverse the rôles of u_1 and u_2.

To check if your chosen parametrisation produces a correctly oriented normal vector field, imagine your surface and solid being drawn in a *right-handed* three-dimensional coordinate system. Then, at any point of your surface, compute the vectors $\partial_1\vec{x}$ and $\partial_2\vec{x}$. Point your thumb and index finger in the direction of those two vectors (thumb in the $\partial_1\vec{x}$-direction, index finger in the $\partial_2\vec{x}$-direction). Now extend your middle finger perpendicularly to your thumb and index finger. As discussed in Chapter 3, your middle finger now points in the direction of the cross product $\partial_1\vec{x} \times \partial_2\vec{x}$. If this also points outside of the bounded solid, you have chosen a correctly oriented parametrisation $\vec{x}$.

Remark 7.5. We give yet another warning: the boundary of a surface in $\mathbb{R}^3$ is to be understood as the points where the surface "does not locally look like $\mathbb{R}^2$ but like $\mathbb{R}_{\leq 0} \times \mathbb{R}$". For instance: the boundary of a disk ● is the circle ○ even if they are viewed as embedded in $\mathbb{R}^3$.

7.4.3. The theorems. In this section we give two classical integral theorems: those of Kelvin–Stokes and Gauß. There are other theorems of a similar flavour: Green, Green–Riemann, etc. Incidentally, all of them have the same underlying principle and this realisation was brought to its pinnacle by Élie Cartan who formulated

a modern version called the ***generalised Stokes theorem*** that works in arbitrary dimensions. The actual details are well beyond what can be discussed in this course, but we do sketch some aspects of this in § 7.6 below.

Theorem 7.6 (Kelvin–Stokes theorem)**.** *Suppose that $S \subseteq \mathbb{R}^3$ is some compact, parametrised surface with parametrised boundary ∂S and suppose that both parametrisations are chosen to match the orientation convention. (The boundary ∂S is supposed to be understood in the sense of Remark 7.5.) Then*

$$\oint_{\partial S} \vec{K} \cdot \mathrm{d}\vec{s} = \iint_S (\mathrm{rot}\,\vec{K}) \cdot \mathrm{d}\vec{A}$$

for any C^1 vector field $\vec{K}$ defined on a neighbourhood of S.

Theorem 7.7 (Gauß's divergence theorem)**.** *Suppose that $B \subseteq \mathbb{R}^3$ is some compact parametrised solid with boundary ∂B admitting a parametrisation with normal vector field pointing outwards of B. Then*

$$\oiint_{\partial B} \vec{K} \cdot \mathrm{d}\vec{A} = \iiint_B (\mathrm{div}\,\vec{K})\,\mathrm{d}V$$

for any C^1 vector field $\vec{K}$ defined on a neighbourhood of B.

Remark. The circular marks in the symbols

$$\oint_C \quad \text{and} \quad \oiint_S$$

are supposed to indicate that the curve C and surface S being integrated over do not contain boundary points in the sense of Remark 7.5. Readers who are confused by this can safely ignore this.

7.4.4. An example. We intend to verify the validity of Gauß's divergence theorem for a concrete example by computing both integrals by hand.

Example 7.8. Let

$$\vec{K}(x_1, x_2, x_3) = (x_1, x_2, x_3),$$

and take B to be the half-ball

$$B = \{\, \vec{x} \in \mathbb{R}^3 : \|\vec{x}\| \leq 1,\ x_3 \geq 0 \,\}.$$

Then

$$\iiint_B (\mathrm{div}\,\vec{K})\,\mathrm{d}V = \iiint_B 3\,\mathrm{d}V.$$

Clearly this is 3 times the volume of the half-ball B. One might remember from geometry class in school that the volume of a ball of radius r is $\frac{4}{3}\pi r^3$. Hence, we

expect our integral to be 2π. Indeed, we can see this using polar coordinates (recall Proposition 6.5)

$$\begin{aligned}\iiint_B 3\,\mathrm{d}V &= \int_{[0,1]\times[0,\pi/2]\times[0,2\pi)} 3r^2\sin(\theta)\,\mathrm{d}^3(r,\theta,\varphi)\\ &= \int_0^1\int_0^{\pi/2}\int_0^{2\pi} 3r^2\sin(\theta)\,\mathrm{d}\varphi\,\mathrm{d}\theta\,\mathrm{d}r\\ &= 6\pi\int_0^1\int_0^{\pi/2} r^2\sin(\theta)\,\mathrm{d}\theta\,\mathrm{d}r\\ &= 6\pi\int_0^1 \big(-r^2\cos(\theta)\big)\Big|_{\theta=0}^{\pi/2}\mathrm{d}r\\ &= 6\pi\int_0^1 r^2\,\mathrm{d}r = 6\pi\frac{1}{3}r^3\Big|_{r=0}^1 = 2\pi.\end{aligned}$$

Next, we compute

$$\oiint_{\partial B}\vec{K}\cdot\mathrm{d}\vec{A} = \iint_S \vec{K}\cdot\mathrm{d}\vec{A} + \iint_D \vec{K}\cdot\mathrm{d}\vec{A}.$$

Note that the boundary ∂B of B consists of a half-sphere

$$S = \{\,\vec{x}\in\mathbb{R}^3 : \|\vec{x}\| = 1,\ x_3 \geq 0\,\}$$

and the disk

$$D = \{\,\vec{x}\in\mathbb{R}^3 : \|\vec{x}\| \leq 1,\ x_3 = 0\,\}.$$

We parametrise S using spherical coordinates (with radius fixed to one), getting

$$\iint_S \vec{K}\cdot\mathrm{d}\vec{A} = \int_{[0,\pi/2]\times[0,2\pi)} \vec{K}(\cos(\varphi)\sin\theta,\sin(\varphi)\sin\theta,\cos\theta)\cdot(\ldots)\,\mathrm{d}^2(\theta,\varphi),$$

where "…" stands for

$$\begin{pmatrix}\cos(\varphi)\cos\theta\\ \sin(\varphi)\cos\theta\\ -\sin\theta\end{pmatrix}\times\begin{pmatrix}-\sin(\varphi)\sin\theta\\ \cos(\varphi)\sin\theta\\ 0\end{pmatrix} = \begin{pmatrix}\sin(\theta)^2\cos\varphi\\ \sin(\theta)^2\sin\varphi\\ \sin(\theta)\cos\theta\end{pmatrix}.$$

Since $\vec{K}$ is just the identity map on $\mathbb{R}^3$, it simply reproduces the vector that is plugged into it. Consequently, the dot product to be computed in the above integral equals

$$\begin{aligned}\begin{pmatrix}\cos(\varphi)\sin\theta\\ \sin(\varphi)\sin\theta\\ \cos\theta\end{pmatrix}\cdot\begin{pmatrix}\sin(\theta)^2\cos\varphi\\ \sin(\theta)^2\sin\varphi\\ \sin(\theta)\cos\theta\end{pmatrix} &= \cos(\varphi)^2\sin(\theta)^3 + \sin(\varphi)^2\sin(\theta)^3 +\\ &\quad + \cos(\theta)^2\sin\theta\\ &= \sin(\theta)^3 + \cos(\theta)^2\sin\theta = \sin(\theta).\end{aligned}$$

Therefore,

$$\iint_S \vec{K}\cdot \mathrm{d}\vec{A} = \int_{[0,\pi/2]\times[0,2\pi)} \sin(\theta)\,\mathrm{d}^2(\theta,\varphi) = \int_0^{\pi/2}\int_0^{2\pi} \sin(\theta)\,\mathrm{d}\varphi\,\mathrm{d}\theta = 2\pi.$$

For D we choose the parametrisation

$$\vec{x}\colon [0,2\pi]\times[0,1], \quad \begin{pmatrix}\varphi\\ r\end{pmatrix} \mapsto \begin{pmatrix} r\cos\varphi\\ r\sin\varphi\\ 0\end{pmatrix}.$$

Observe closely the order of the arguments.[6] For that reason we have

$$(\partial_1\vec{x})\times(\partial_2\vec{x}) = \begin{pmatrix}-r\sin\varphi\\ r\cos\varphi\\ 0\end{pmatrix}\times\begin{pmatrix}\cos\varphi\\ \sin\varphi\\ 0\end{pmatrix} = \begin{pmatrix}0\\ 0\\ -r\cos(\varphi)^2 - r\sin(\varphi)^2\end{pmatrix} = \begin{pmatrix}0\\ 0\\ -r\end{pmatrix},$$

which points outwards of B, as it should. Now

$$\begin{aligned}\iint_D \vec{K}\cdot\mathrm{d}\vec{A} &= \int_{[0,\pi/2]\times[0,2\pi)} \vec{K}(r\cos\varphi, r\sin\varphi, 0)\cdot\begin{pmatrix}0\\0\\-r\end{pmatrix}\mathrm{d}^2(\theta,\varphi)\\ &= \int_{[0,\pi/2]\times[0,2\pi)} \begin{pmatrix} r\cos\varphi\\ r\sin\varphi\\ 0\end{pmatrix}\cdot\begin{pmatrix}0\\0\\-r\end{pmatrix}\mathrm{d}^2(\theta,\varphi).\end{aligned}$$

One sees that the dot product inside the integral is zero. Consequently, the integral of $\vec{K}$ over D vanishes. Finally,

$$\oiint_{\partial B}\vec{K}\cdot\mathrm{d}\vec{A} = \iint_S \vec{K}\cdot\mathrm{d}\vec{A} + \iint_D \vec{K}\cdot\mathrm{d}\vec{A} = 2\pi + 0 = 2\pi = \iiint_B (\operatorname{div}\vec{K})\,\mathrm{d}V,$$

as expected.

7.5. Plausibility of Gauß's theorem

In this section we sketch some of the ingredients that usually go into the proof of Gauß's divergence theorem (Theorem 7.7).

7.5.1. The strategy. The basic idea is to prove the theorem for a box (which we do in § 7.5.2 below) and then decompose the parameter space into little boxes. When applying the box version of Gauß's theorem, one sees that the contribution from the sides of adjacent boxes cancel, due to opposite signs. However, where the boxes meet the boundary, there is no adjacent box cancelling the contribution of the relevant side of the box and all these non-cancelled contributions sum up to give the integral over the boundary of the surface.

[6] Incidentally, as we shall see below, in this particular example the orientation of D does not matter, because the relevant integral of $\vec{K}$ over D is zero anyway (independent of the chosen orientation).

For justifying the box version of Gauß's theorem, we do start with the formula, but actually try to *derive* it: we try to *find* the correct integrand on the right hand side of the formula in Theorem 7.7 and find it to be the divergence $\operatorname{div}\vec{K}$. The reader should get two insights from this:

(1) Gauß's theorem is natural;
(2) Other integral theorems such as the Kelvin–Stokes theorem (Theorem 7.6) can be obtained similarly—in principle—even *without* knowing the correct formula in the first place.

Remark. The approach below has several deficiencies:

(1) First, many questions with regard to the boundary and its parametrisation remain open. In serious treatments, one would define objects called oriented manifolds with boundary (things that locally "look like" $\mathbb{R}^n$ or $\mathbb{R}_{\leq 0} \times \mathbb{R}^{n-1}$ and have some fixed notion of orientation) and derive the orientation on their boundaries from the orientation of the ambient object (to make things fit together).
(2) Second, our whole argument of decomposing the parameter space is a huge lie, because there are major technical issues with this. (For the purpose of building some intuition, this process seems adequate though.)
(3) Moreover, one should take into account that "locally looking like something" does not necessitate "globally" looking like something; we have already encountered such subtleties when dealing with spherical coordinates where we had to ignore some hypotheses of our theorems as to make the results applicable (or pretend to). With care and luck, this may work, yet for building the theory on a sound footing, one cannot get away with such sloppiness and must take the possibility of parametrising some global object using many parametrisations of "local pieces". This gives rise to further difficulties with making these many parametrisations patch together nicely and even making sense of the generalised integrals needed. The various limiting processes that are involved even make the use of the Darboux integral quite cumbersome and call for more advanced machinery altogether.

We refrain from providing further elaboration.

7.5.2. Gauß's theorem for a box in parameter space. Let $U \subseteq \mathbb{R}^3$ be some compact parameter space. Let $\vec{x}\colon U \to \mathbb{R}^3$ be an orientation-preserving parametrisation[7] of some solid $B = \vec{x}(U)$. Given any point $\vec{u}_0 \in U$ and positive real numbers $\Delta_1, \Delta_2, \Delta_3 > 0$ we consider the solid

$$Q = \vec{x}(\vec{u}_0 + (-\Delta_1, \Delta_1) \times (-\Delta_2, \Delta_2) \times (-\Delta_2, \Delta_2)).$$

It's boundary is given by

$$\partial Q = F_1^{\pm} \cup F_2^{\pm} \cup F_3^{\pm},$$

[7]This means $\det J_{\vec{x}}(\vec{u}) > 0$ for every $\vec{u} \in U$. Up to some technical difficulties with vanishing determinant this can be achieved by potentially switching the roles of the coordinates of U.

where

$$F_1^{\pm} = \vec{x}(\vec{u}_0 + \{\pm\Delta_1\} \times [-\Delta_2, \Delta_2] \times [-\Delta_3, \Delta_3]),$$
$$F_2^{\pm} = \vec{x}(\vec{u}_0 + [-\Delta_1, \Delta_1] \times \{\pm\Delta_2\} \times [-\Delta_3, \Delta_3]),$$
$$F_3^{\pm} = \vec{x}(\vec{u}_0 + [-\Delta_1, \Delta_1] \times [-\Delta_2, \Delta_2] \times \{\pm\Delta_3\}).$$

Write

$$P_1 := [-\Delta_2, \Delta_2] \times [-\Delta_3, \Delta_3],$$
$$P_2 := [-\Delta_1, \Delta_1] \times [-\Delta_3, \Delta_3],$$
$$P_3 := [-\Delta_1, \Delta_1] \times [-\Delta_2, \Delta_2].$$

We now wish to choose an orientation on ∂Q. Clearly there is an obvious parametrisation of F_1^+ given by

$$\vec{x}_1^+ \colon P_1 \to \mathbb{R}^3, \quad (t_1, t_2) \mapsto \vec{x}(\vec{u}_0 + (\Delta, t_1, t_2)),$$

one for F_2^+

$$\vec{x}_2^+ \colon P_2 \to \mathbb{R}^3, \quad (t_1, t_2) \mapsto \vec{y}(\vec{u}_0 + (t_1, \Delta, t_2)),$$

and similarly for the other "faces" $F_i^{\pm}$ $(i = 1, 2, 3)$ of the "cube" Q. These parametrisations induce normal vector fields

$$\frac{\partial \vec{x}_1^+}{\partial t_1} \times \frac{\partial \vec{x}_1^+}{\partial t_2} \quad \text{for} \quad F_1^+, \qquad \frac{\partial \vec{x}_2^+}{\partial t_1} \times \frac{\partial \vec{x}_2^+}{\partial t_2} \quad \text{for} \quad F_2^+ \qquad \text{etc.}$$

Note that the first normal vector field points outwards of Q, whereas the second points inwards.[8] To fix this, we add appropriate signs. In hopefully self-explanatory notation, we summarise these sign conventions via the formula

$$\partial Q = \sum_{i=1,2,3} (-1)^{i-1}(F_{+i} - F_{-i}) = (F_{+1} - F_{-1}) - (F_{+2} - F_{-2}) + (F_{+3} - F_{-3}).$$

(We are wilfully ignorant about the points on the boundary of the "faces" $F_i^{\pm}$ here.)

With this agreement, we now get a normal vector field $\vec{N}$ for ∂Q and can consider

$$\iint_{\partial Q} \vec{K} \cdot \mathrm{d}\vec{A}.$$

for any vector field $\vec{K} \colon B \to \mathbb{R}^3$. Precisely, the above integral equals

$$\int_{P_1} \left((\vec{K} \circ \vec{x}_1^+) \cdot \left(\frac{\partial \vec{x}_1^+}{\partial t_1} \times \frac{\partial \vec{x}_1^+}{\partial t_2} \right) - (\vec{K} \circ \vec{x}_1^-) \cdot \left(\frac{\partial \vec{x}_1^-}{\partial t_1} \times \frac{\partial \vec{x}_1^-}{\partial t_2} \right) \right) \mathrm{d}^2\vec{t} \pm \ldots \tag{7.2}$$

where the omitted terms hiding behind "$\pm\ldots$" arise from integration along the remaining "faces" $F_i^{\pm}$ in accordance with our sign convention.

[8] This is where we use that $\vec{y}$ preserves orientation.

We would now like to have some integrand $\boxed{??}$ such that

$$\iint_{\partial Q} \vec{K} \cdot \mathrm{d}\vec{A} \stackrel{!}{=} \iiint_{Q} \boxed{??}\, \mathrm{d}V = \int_{\prod_i [-\Delta_i, \Delta_i]} \Big(\boxed{??} \circ \vec{x}\Big)(\det J_{\vec{x}})\, \mathrm{d}^3\vec{u}. \tag{7.3}$$

We shall now assume that such an integrand $\boxed{??}$ exists and are concerned with computing it. (However, we shall not be concerned with proving that the integrand thus obtained *actually* satisfies (7.3).) Note that $\boxed{??}$ may be computed via the limit

$$\lim_{\Delta_1, \Delta_2, \Delta_3 \searrow 0} \frac{1}{2^3 \Delta_1 \Delta_2 \Delta_3} \iiint_{Q} \boxed{??}\, \mathrm{d}V = \Big(\boxed{??} \circ \vec{x}\Big) \det J_{\vec{x}} \Big|_{\vec{u}_0},$$

where for the duration of this section, the notation $h|_{\vec{u}_0}$ means 'the function h evaluated at $\vec{u}_0$'. However, if $\boxed{??}$ really does satisfy (7.3), then this limit must also be equal to

$$\lim_{\Delta_1, \Delta_2, \Delta_3 \searrow 0} \frac{1}{2^3 \Delta_1 \Delta_2 \Delta_3} \iint_{\partial Q} \vec{K} \cdot \mathrm{d}\vec{A}.$$

Recall that the integral herein equals (7.2). We shall only consider the contribution of the first term in (7.2) stemming from the integration along the faces $F_1^+ - F_1^-$, namely

$$I_1^+ = \lim_{\Delta_1, \Delta_2, \Delta_3 \searrow 0} \frac{1}{2^3 \Delta_1 \Delta_2 \Delta_3} \int_{P_1} \ldots \mathrm{d}^2\vec{t}.$$

with integrand

$$(\vec{K} \circ \vec{x}_1^+) \cdot \left(\frac{\partial \vec{x}_1^+}{\partial t_1} \times \frac{\partial \vec{x}_1^+}{\partial t_2} \right) - (\vec{K} \circ \vec{x}_1^-) \cdot \left(\frac{\partial \vec{x}_1^-}{\partial t_1} \times \frac{\partial \vec{x}_1^-}{\partial t_2} \right).$$

(The contribution of the other terms is handled similarly.) Upon splitting off $\lim_{\Delta_1 \searrow 0}$ and moving this inside, we obtain

$$I_1^+ = \lim_{\Delta_2, \Delta_3 \searrow 0} \frac{1}{2^2 \Delta_2 \Delta_3} \int_{P_1} \ldots \mathrm{d}^2\vec{t} \tag{7.4}$$

with integrand

$$\begin{aligned}
&\lim_{\Delta_1 \searrow 0} \frac{1}{2\Delta_1} \left((\vec{K} \circ \vec{x}_1^+) \cdot \left(\frac{\partial \vec{x}_1^+}{\partial t_1} \times \frac{\partial \vec{x}_1^+}{\partial t_2} \right) - (\vec{K} \circ \vec{x}_1^-) \cdot \left(\frac{\partial \vec{x}_1^-}{\partial t_1} \times \frac{\partial \vec{x}_1^-}{\partial t_2} \right) \right)(t_1, t_2) \\
&\quad = \lim_{\Delta_1 \searrow 0} \frac{1}{2\Delta_1} \left((\vec{K} \circ \vec{x}_1) \cdot \left(\frac{\partial \vec{x}}{\partial u_2} \times \frac{\partial \vec{x}}{\partial u_3} \right) \Big|_{\vec{u}_0 + (\Delta_1, t_1, t_2)} + \right. \\
&\qquad \left. - (\vec{K} \circ \vec{x}_1) \cdot \left(\frac{\partial \vec{x}}{\partial u_2} \times \frac{\partial \vec{x}}{\partial u_3} \right) \Big|_{\vec{u}_0 + (-\Delta_1, t_1, t_2)} \right).
\end{aligned}$$

This turns out to be

$$\frac{\partial}{\partial u_1}\left((\vec{K}\circ\vec{x})\cdot\left(\frac{\partial\vec{x}}{\partial u_2}\times\frac{\partial\vec{x}}{\partial u_3}\right)\right)\Bigg|_{\vec{u}_0+(0,t_1,t_2)}.$$

Upon using the chain rule for differentiation and recalling (3.7), this turns out to be

$$\left(\frac{\partial\vec{K}}{\partial x_1}\circ\vec{x}\right)\det J_{\vec{x}}\Bigg|_{\vec{u}_0+(0,t_1,t_2)}.$$

Plugging this into (7.4) and computing the limit there, we find that

$$I_1^+=\left(\frac{\partial\vec{K}}{\partial x_1}\circ\vec{x}\right)\det J_{\vec{x}}\Bigg|_{\vec{u}_0}.$$

Summing over the contribution of the remaining "faces", we deduce that

$$\boxed{??}=\frac{\partial\vec{K}}{\partial x_1}+\frac{\partial\vec{K}}{\partial x_2}+\frac{\partial\vec{K}}{\partial x_3}=\operatorname{div}\vec{K}.$$

Therefore, finally,

$$\iint_{\partial Q}\vec{K}\cdot\mathrm{d}\vec{A}=\iiint_{Q}\operatorname{div}\vec{K}\,\mathrm{d}V.$$

7.5.3. Driving the argument home. After thinking of decomposing the full parameter space into boxes

$$U\approx\bigcup_{\text{boxes }Q}Q$$

(at least approximately, whatever that means), ignoring any difficulties with this, we are lead to

$$\iiint_U\operatorname{div}\vec{K}\,\mathrm{d}V\approx\sum_{\text{boxes }Q}\iiint_Q\operatorname{div}\vec{K}\,\mathrm{d}V=\sum_{\text{boxes }Q}\iint_{\partial Q}\vec{K}\cdot\mathrm{d}\vec{A}=\sum_F\iint_F\vec{K}\cdot\mathrm{d}\vec{A},$$

where the last sum is taken over the faces (suitably oriented) of boxes which were not cancelled by the face of an adjacent box's face. These are—at least morally—the faces on the boundary:

$$\bigcup_{\text{faces }F}F\approx\partial U.$$

Therefore,

$$\iiint_U\operatorname{div}\vec{K}\,\mathrm{d}V\approx\sum_F\iint_F\vec{K}\cdot\mathrm{d}\vec{A}\approx\iint_{\partial U}\vec{K}\cdot\mathrm{d}\vec{A}.$$

(Again, be warned that we have shamelessly brushed many (serious!) technical issues under the rug. The above argument should not be considered a "proof" of Gauß's divergence theorem.)

7.6. Cartan's calculus of differential forms

In our discussion so far, we have met a variety of formulas, some of which may be quite hard to remember. It turns out that a very elegant framework for organising all of the above has been conceived by the French mathematician Élie Cartan. Describing his calculus requires some work.

7.6.1. Forms. A map

$$\omega\colon \underbrace{\mathbb{R}^n \times \cdots \times \mathbb{R}^n}_{k \text{ terms}} \to \mathbb{R},$$

is called a ***form***. Such a form is called ***k-linear*** or ***k-multi-linear*** if it is linear in each of its k (vector-valued) arguments seperately. It is called ***alternating*** if it is k-linear and

$$\omega(\vec{v}_1, \ldots, \vec{v}_i, \ldots, \vec{v}_j, \ldots, \vec{v}_k) = -\omega(\vec{v}_1, \ldots, \underbrace{\vec{v}_j, \ldots, \vec{v}_i}_{i\text{-th and } j\text{-th argument swapped}}, \ldots, \vec{v}_k)$$

for all arguments $\vec{v}_1, \ldots, \vec{v}_k \in \mathbb{R}^n$ and any $1 \leq i < j \leq k$. We let $\mathrm{Alt}^k \mathbb{R}^n$ denote the set of ***alternating k-forms***.

For $k = 0$ we interpret

$$\underbrace{\mathbb{R}^n \times \cdots \times \mathbb{R}^n}_{0 \text{ terms}} = \{0\} \quad \text{(by convention)}$$

and *identify* functions $\omega\colon \{0\} \to \mathbb{R}$ with their values $\omega(0) \in \mathbb{R}$. In this sense, we have $\mathrm{Alt}^0 \mathbb{R}^n = \mathbb{R}$.

Example. The determinant function $\det\colon \mathbb{R}^{n\times n} \to \mathbb{R}$ is an alternating n-form if elements of $\mathbb{R}^{n\times n}$ are viewed as n-tuples of vectors in $\mathbb{R}^n$.

Example. Any linear map $f\colon \mathbb{R}^n \to \mathbb{R}$ is an alternating 1-form.

Example. Let $A = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \mathbb{R}^{2\times 2}$ be an arbitrary 2×2-matrix. Then the map $\omega_A\colon \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$ given by

$$\omega_A\left(\begin{pmatrix} v_1 \\ v_2 \end{pmatrix}, \begin{pmatrix} w_1 \\ w_2 \end{pmatrix}\right) \mapsto \begin{pmatrix} v_1 \\ v_2 \end{pmatrix}^{\intercal} \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \end{pmatrix} = v_1 a w_1 + v_1 b w_2 + v_2 c w_1 + v_2 d w_2.$$

is 2-linear (also called ***bilinear***). One can show that it is alternating if and only if $a = d = 0$ and $b = -c$. In this case, the formula for determinants of 2×2-matrices shows that

$$\omega_A\left(\begin{pmatrix} v_1 \\ v_2 \end{pmatrix}, \begin{pmatrix} w_1 \\ w_2 \end{pmatrix}\right) = v_1 b w_2 - v_2 b w_1 = b \det \begin{pmatrix} v_1 & w_1 \\ v_2 & w_2 \end{pmatrix}.$$

7.6.2. Wedge product. Let $U \subseteq \mathbb{R}^n$ be open. A function $\omega\colon U \to \mathrm{Alt}^k \mathbb{R}^n$ shall be called a ***differential k-form*** on U and one usually writes $\omega_{\vec{p}}$ for the k-form $\omega(\vec{p}) \in \mathrm{Alt}^k$ given by ω at a point $\vec{p} \in U$.

Given a differential k-form ω and a differential ℓ-form η, one can construct a differential $k+\ell$-form $\omega \wedge \eta$ (called the ***wedge product of ω and η***) using

$$\begin{aligned}(\omega \wedge \eta)_{\vec{p}}(\vec{v}_1, \ldots, \vec{v}_k, \vec{v}_{k+1}, \ldots, \vec{v}_{k+\ell}) \\ := \frac{1}{k!\,\ell!} \sum_{\sigma} (\pm 1)\, \omega(\vec{v}_{\sigma(1)}, \ldots, \vec{v}_{\sigma(k)})\, \eta(\vec{v}_{\sigma(k+1)}, \ldots, \vec{v}_{\sigma(k+\ell)}),\end{aligned}$$

where σ ranges over all permutations of the index set $\{1, \ldots, k+\ell\}$ and the sign ± 1 is chosen according to the 'sign of σ', the latter being a concept from group theory which we do not explain further. Naturally, this means that we have not properly defined the wedge product here. Anyway, despite this, we provide identities for computing with wedge products below and, in practice, these are all one needs. Addition of differential k-forms and multiplication by a scalar r is defined pointwise. Then, for differential k-forms ω, $\tilde{\omega}$, differential ℓ-forms η, $\tilde{\eta}$, any differential j-form ν and any real number r one has the following identities:

- $\omega \wedge (\eta \wedge \nu) = (\omega \wedge \eta) \wedge \nu$, (associativity)
- $(r\omega) \wedge \eta = \omega \wedge (r\eta) = r(\omega \wedge \eta)$, $\omega \wedge (\eta + \tilde{\eta}) = (\omega \wedge \eta) + (\omega \wedge \tilde{\eta})$, $(\omega + \tilde{\omega}) \wedge \eta = (\omega \wedge \eta) + (\tilde{\omega} \wedge \eta)$, (bilinearity)
- $\omega \wedge \eta = (-1)^{k\ell}(\eta \wedge \omega)$. (anti-commutativity)

For any smooth function $f\colon U \to \mathbb{R}$ (a differential 0-form) its differential $\mathrm{d}f$ is a differential 1-form. We do remind the reader of the formula

$$\mathrm{d}f = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}\, \mathrm{d}x_i$$

from § 5.2, where x_i is the function $U \to \mathbb{R}$ that maps a vector $\vec{v} = (v_1, \ldots, v_n)$ to its i-th coordinate v_i.

One can show that any differential k-form ω on U can be written uniquely as

$$\omega = \sum_{1 \le i_1 < \ldots < i_k \le n} \cdots \sum \omega_{i_1, \ldots, i_k}\, \mathrm{d}x_{i_1} \wedge \ldots \wedge \mathrm{d}x_{i_k}$$

with certain functions $\omega_{i_1,\ldots,i_k}\colon U \to \mathbb{R}$. The differential k-form ω is called ***continuous***, ***differentiable***, or ***smooth*** if the corresponding property holds for all the functions $\omega_{i_1,\ldots,i_k}\colon U \to \mathbb{R}$. The set of ***smooth differential k-forms*** on U is denoted by $\Omega^k(U)$.

7.6.3. Exterior derivative. There is a function d (called the ***Cartan derivative*** or ***exterior derivative***) that smooth maps differential k-forms to smooth differential $k+1$-forms and satisfies the following properties (for $\omega, \tilde{\omega} \in \Omega^k(U)$, $\lambda, \mu \in \mathbb{R}$, and $\eta \in \Omega^\ell(U)$):

- d is linear: $\mathrm{d}(\lambda\omega + \mu\tilde{\omega}) = \lambda\mathrm{d}\omega + \mu\mathrm{d}\tilde{\omega}$,

- d coincides with the total derivative on 0-forms (= differentiable functions $f\colon U \to \mathbb{R}$),
- $\mathrm{d}(\omega \wedge \eta) = (\mathrm{d}\omega) \wedge \eta + (-1)^k(\omega \wedge \mathrm{d}\eta)$,
- $\mathrm{d}(\mathrm{d}\omega)$ is the zero differential k+2-form.

In particular, for any smooth function $f\colon U \to \mathbb{R}$, we have

$$0 = \mathrm{d}(f\,\mathrm{d}f) = \mathrm{d}(f \wedge \mathrm{d}f) = (\mathrm{d}f \wedge \mathrm{d}f) - (f \wedge \mathrm{d}(\mathrm{d}f)) = (\mathrm{d}f \wedge \mathrm{d}f). \tag{7.5}$$

Let $\vec{x}\colon U \to W$ be any function from U to some open set $W \subseteq \mathbb{R}^m$. Then one obtains a map $\vec{x}^*$ from differential k-forms η on W to differential k-forms $\vec{x}^*\eta$ on U via

$$(\vec{x}^*\eta)_{\vec{q}}(\vec{v}_1, \ldots, \vec{v}_k) \coloneqq \eta_{\vec{x}(\vec{q})}(\mathrm{d}\vec{x}_{\vec{q}}(\vec{v}_1), \ldots, \mathrm{d}\vec{x}_{\vec{q}}(\vec{v}_k)).$$

This is called the ***pull-back of η by $\vec{x}$***. Note that the pull-back of k-forms by $\vec{x}$ is defined for all k, despite there being no reference to k in the notation $\vec{x}^*$. Below, we also deal with forms of varying k and pull-backs of those. If we want to highlight the k, then we write $\vec{x}^{*k}$. The pull-back is linear: for any two k-forms η and $\tilde{\eta}$ on W and any real numbers r and $\tilde{r}$ we have

$$\vec{x}^*(r\eta + \tilde{r}\tilde{\eta}) = r\vec{x}^*\eta + \tilde{r}\vec{x}^*\tilde{\eta}.$$

Given a function $g\colon W \to \mathbb{R}$ (a differential 0-form) and a differential k-form η on W, we have

$$\vec{x}^*(f\,\eta) = (f \circ \vec{x})\vec{x}^*\eta.$$

Moreover, the pull-back is compatible with the wedge product: for any differential k-form η and any differential ℓ-form $\tilde{\eta}$ on W, one has

$$\vec{x}^*(\eta \wedge \tilde{\eta}) = (\vec{x}^*\eta) \wedge (\vec{x}^*\tilde{\eta}).$$

If $\vec{x}$ is smooth, then smooth differential k-forms on W pull back to smooth differential k-forms on U, thus inducing a map

$$\vec{x}^*\colon \Omega^k(W) \to \Omega^k(U).$$

One can show that the Cartan derivative commutes with the pull-back, i.e., one has

$$\mathrm{d}(\vec{x}^*\eta) = \vec{x}^*(\mathrm{d}\eta)$$

for any smooth differential k-form η on W. Note that with the previously defined slightly more verbose notation for the pull-back, the above equation actually reads

$$\mathrm{d}(\vec{x}^{*k}\eta) = \vec{x}^{*k+1}(\mathrm{d}\eta).$$

This equation can also be illustrated using a commutative diagram:

$$\begin{array}{ccc} V\colon & \Omega^k(V) & \xrightarrow{\mathrm{d}} \Omega^{k+1}(V), \\ \big\downarrow{\scriptstyle \vec{x}} & \big\uparrow{\scriptstyle \vec{x}^{*k}} & \big\uparrow{\scriptstyle \vec{x}^{*k+1}} \\ W\colon & \Omega^k(W) & \xrightarrow{\mathrm{d}} \Omega^{k+1}(W). \end{array}$$

These diagrams, for $k = 1, 2, 3, \ldots$, fit together to form an even larger commutative diagram:

$$\begin{array}{lccccccc} V\colon & \Omega^0(V) & \xrightarrow{\mathrm{d}} & \Omega^1(V) & \xrightarrow{\mathrm{d}} & \Omega^2(V) & \xrightarrow{\mathrm{d}} & \Omega^3(V) \xrightarrow{\mathrm{d}} \ldots, \\ \downarrow \vec{x} & \uparrow \vec{x}^* & & \uparrow \vec{x}^* & & \uparrow \vec{x}^* & & \uparrow \vec{x}^* \\ W\colon & \Omega^0(W) & \xrightarrow{\mathrm{d}} & \Omega^1(W) & \xrightarrow{\mathrm{d}} & \Omega^2(W) & \xrightarrow{\mathrm{d}} & \Omega^3(W) \xrightarrow{\mathrm{d}} \ldots. \end{array}$$

The fact stated at the beginning of this subsection that $\mathrm{d}(\mathrm{d}\omega)$ is the zero differential $k+2$-form for any smooth differential k-form ω makes the above diagram especially interesting for mathematicians and physicists, as it can be exploited to study the geometry of spaces (so-called ***smooth manifolds***). Here we cannot say much more about this, but the interested reader is referred to [9, Chapter 13] or [6, pp. 250–252].

7.6.4. Integration of forms. Recall that $U \subseteq \mathbb{R}^n$ should be open. Any differential n-form ω on U has the form

$$\omega = f \, \mathrm{d}u_1 \wedge \ldots \wedge \mathrm{d}u_n$$

with some function $f\colon U \to \mathbb{R}$, where $u_i\colon U \to \mathbb{R}$ denotes the function mapping every vector $\vec{v} \in \mathbb{R}^n$ to its i-th coordinate v_i. Assuming that f is integrable on U we let

$$\int_U f \, \mathrm{d}u_1 \wedge \ldots \wedge \mathrm{d}u_n := \int_U f(\vec{u}) \, \mathrm{d}^n\vec{u}.$$

Now suppose that $\vec{x}\colon U \to \vec{x}(U) \subseteq \mathbb{R}^m$ (smoothly) parametrises some object $\vec{x}(U)$ in m-dimensional space and suppose that $\det J_{\vec{x}}(\vec{p}) > 0$ for every $\vec{p} \in U$. Let η be a smooth differential k-form on $\vec{x}(U)$. We then let

$$\int_{\vec{x}(U)} \eta := \int_U \vec{x}^*\eta,$$

provided that the last integral exists; note that $\omega = \vec{x}^*\eta$ is a differential n-form on $U \subseteq \mathbb{R}^n$.

7.6.5. Comparison with previous notions of integration. Suppose that $\vec{x}\colon U \to \vec{x}(U)$ parametrises a surface $S = \vec{x}(U) \subset \mathbb{R}^3$ and suppose that $\vec{K}$ is a vector field defined in S with component functions K_1, K_2 and K_3. Assume that $\vec{K}$ is smooth. Then we associate a smooth differential 2-form η to $\vec{K}$ via

$$\eta = K_1 \, \mathrm{d}\underline{x_2} \wedge \mathrm{d}\underline{x_3} + K_2 \, \mathrm{d}\underline{x_3} \wedge \mathrm{d}\underline{x_1} + K_3 \, \mathrm{d}\underline{x_1} \wedge \mathrm{d}\underline{x_2}, \tag{7.6}$$

where $\underline{x_i}\colon \vec{x}(U) \to \mathbb{R}$ is the map which maps any point in $\vec{x}(U)$ to its i-th coordinate; in particular, $\underline{x_i}$ *does not* denote the i-th component function x_i of our parametrisation $\vec{x}$, although this clash in notation is not entirely unintended (see (7.8) below). We aim to compute $\vec{x}^*\eta$. By linearity of the pull-back, we may focus on each of the three terms, occuring in the definition of η, seperately. We have

$$\vec{x}^*(K_1 \, \mathrm{d}\underline{x_2} \wedge \mathrm{d}\underline{x_3}) = (K_1 \circ \vec{x}) \, \vec{x}^*(\mathrm{d}\underline{x_2} \wedge \mathrm{d}\underline{x_3}). \tag{7.7}$$

Now, because pull-back commutes with the Cartan derivative,

$$\vec{x}^*(\mathrm{d}\underline{x_2}) = \mathrm{d}(\vec{x}^*\underline{x_2}) = \mathrm{d}(\underline{x_2} \circ \vec{x}).$$

The composition $\underline{x_i} \circ \vec{x}$ truly *is* x_i, the i-th component function of $\vec{x}$. Therefore,

$$\vec{x}^*(\mathrm{d}\underline{x_i}) = \mathrm{d}x_i. \tag{7.8}$$

Because the Cartan derivative coincides with the total differential on differential 0-forms, we have

$$\vec{x}^*(\mathrm{d}\underline{x_2}) = \frac{\partial x_2}{\partial u_1}\,\mathrm{d}u_1 + \frac{\partial x_2}{\partial u_2}\,\mathrm{d}u_2.$$

Therefore,

$$\begin{aligned}\vec{x}^*(\mathrm{d}\underline{x_2} \wedge \mathrm{d}\underline{x_3}) &= \left(\frac{\partial x_2}{\partial u_1}\,\mathrm{d}u_1 + \frac{\partial x_2}{\partial u_2}\,\mathrm{d}u_2\right) \wedge \left(\frac{\partial x_3}{\partial u_1}\,\mathrm{d}u_1 + \frac{\partial x_3}{\partial u_2}\,\mathrm{d}u_2\right) \\ &= x_{11}^{23}\,\mathrm{d}u_1 \wedge \mathrm{d}u_1 + x_{12}^{23}\,\mathrm{d}u_1 \wedge \mathrm{d}u_2 + x_{21}^{23}\,\mathrm{d}u_2 \wedge \mathrm{d}u_1 + x_{22}^{23}\,\mathrm{d}u_2 \wedge \mathrm{d}u_2,\end{aligned}$$

where we have written

$$x_{11}^{23} = \frac{\partial x_2}{\partial u_1}\frac{\partial x_3}{\partial u_1}, \quad x_{12}^{23} = \frac{\partial x_2}{\partial u_1}\frac{\partial x_3}{\partial u_2}, \quad x_{21}^{23} = \frac{\partial x_2}{\partial u_2}\frac{\partial x_3}{\partial u_1}, \quad x_{22}^{23} = \frac{\partial x_2}{\partial u_2}\frac{\partial x_3}{\partial u_2}.$$

By (7.5), the outer two differential 2-forms $\mathrm{d}u_1 \wedge \mathrm{d}u_1$ and $\mathrm{d}u_2 \wedge \mathrm{d}u_2$ vanish and can, therefore, be omitted from the equation along with the their coefficients x_{11}^{23} and x_{22}^{23}. Moreover, by using that $\mathrm{d}u_2 \wedge \mathrm{d}u_1 = -\mathrm{d}u_1 \wedge \mathrm{d}u_2$, the two middle terms can be combined to yield

$$\vec{x}^*(\mathrm{d}\underline{x_2} \wedge \mathrm{d}\underline{x_3}) = \left(\frac{\partial x_2}{\partial u_1}\frac{\partial x_3}{\partial u_2} - \frac{\partial x_2}{\partial u_2}\frac{\partial x_3}{\partial u_1}\right)\mathrm{d}u_1 \wedge \mathrm{d}u_2.$$

In view of (7.6) and (7.7) it transpires that

$$\vec{x}^*\eta = (\vec{K} \circ \vec{x}) \cdot \left(\frac{\partial \vec{x}}{\partial u_1} \times \frac{\partial \vec{x}}{\partial u_2}\right)\mathrm{d}u_1 \wedge \mathrm{d}u_2.$$

Hence,

$$\int_{\vec{x}(U)} \left(K_1\,\mathrm{d}\underline{x_2} \wedge \mathrm{d}\underline{x_3} + K_2\,\mathrm{d}\underline{x_3} \wedge \mathrm{d}\underline{x_1} + K_3\,\mathrm{d}\underline{x_1} \wedge \mathrm{d}\underline{x_2}\right) = \int_{\vec{x}(U)} \eta = \iint_{\vec{x}(U)} \vec{K} \cdot \mathrm{d}\vec{A}.$$

Next, suppose that $\vec{x}(U)$ equals the boundary ∂B of some solid B as in Gauß's theorem (Theorem 7.7).[9] Then the right hand side of the above equals

$$\iiint_B (\operatorname{div}\vec{K})\,\mathrm{d}V.$$

Incidentally, using (7.5) one easily verifies that the Cartan derivative of η is

$$\mathrm{d}\eta = (\operatorname{div}\vec{K})\,\mathrm{d}\underline{x_1} \wedge \mathrm{d}\underline{x_2} \wedge \mathrm{d}\underline{x_3},$$

[9] This does not really play well with our earlier supposition that U be open, but never mind.

so that

$$\int_B \mathrm{d}\eta = \iiint_B (\operatorname{div}\vec{K})\,\mathrm{d}V.$$

In particular, comparing with the above, Gauß's theorem (Theorem 7.7) takes the form

$$\boxed{\oint_{\partial B} \eta = \int_B \mathrm{d}\eta.}$$

This result holds under much more general assumptions for arbitrary smooth differential k-forms and generalises the fundamental theorem of calculus, Gauß's theorem, and also the Kelvin–Stokes theorem (Theorem 7.6); it is known as the ***generalised Stokes theorem***.

We now drop the underline from all x_i, because our parametrisation $\vec{x}$ from above has served its purpose. Using the notation from § 6.1, one can see that we get a commutative diagram

$$\begin{array}{ccccccc} C^\infty(U) & \xrightarrow{\text{grad}} & \mathscr{V}(U) & \xrightarrow{\text{rot}} & \mathscr{V}(U) & \xrightarrow{\text{div}} & C^\infty(U), \\ \| & & \wr\downarrow{\scriptstyle (1)} & & \wr\downarrow{\scriptstyle (2)} & & \wr\downarrow{\scriptstyle (3)} \\ \Omega^0(U) & \xrightarrow{\mathrm{d}} & \Omega^1(U) & \xrightarrow{\mathrm{d}} & \Omega^2(U) & \xrightarrow{\mathrm{d}} & \Omega^3(U), \end{array} \tag{7.9}$$

where the maps marked with (1), (2) and (3) mean

(1) $\vec{K} \mapsto \vec{K} \cdot \begin{pmatrix} \mathrm{d}x_1 \\ \mathrm{d}x_2 \\ \mathrm{d}x_3 \end{pmatrix} := K_1\,\mathrm{d}x_1 + K_2\,\mathrm{d}x_2 + K_3\,\mathrm{d}x_3,$

(2) $\vec{K} \mapsto \vec{K} \cdot \begin{pmatrix} \mathrm{d}x_2 \wedge \mathrm{d}x_3 \\ \mathrm{d}x_3 \wedge \mathrm{d}x_1 \\ \mathrm{d}x_1 \wedge \mathrm{d}x_2 \end{pmatrix} = K_1\,\mathrm{d}x_2 \wedge \mathrm{d}x_3 + K_2\,\mathrm{d}x_3 \wedge \mathrm{d}x_1 + K_3\,\mathrm{d}x_1 \wedge \mathrm{d}x_2,$
as in (7.6),

(3) $f \mapsto f\,\mathrm{d}x_1 \wedge \mathrm{d}x_2 \wedge \mathrm{d}x_3.$

Moreover, the vertical maps (1), (2) and (3) are compatible with our notions of curve, surface and volume integrals respectively. Namely, we have

(1) $\displaystyle\int_C \vec{K} \cdot \mathrm{d}\vec{s} = \int_C (K_1\,\mathrm{d}x_1 + K_2\,\mathrm{d}x_2 + K_3\,\mathrm{d}x_3),$

(2) $\displaystyle\iint_S \vec{K} \cdot \mathrm{d}\vec{A} = \iint_S (K_1\,\mathrm{d}x_2 \wedge \mathrm{d}x_3 + K_2\,\mathrm{d}x_3 \wedge \mathrm{d}x_1 + K_3\,\mathrm{d}x_1 \wedge \mathrm{d}x_2),$

(3) $\displaystyle\iiint_B f\,\mathrm{d}V = \iiint_B f\,\mathrm{d}x_1 \wedge \mathrm{d}x_2 \wedge \mathrm{d}x_3.$

CHAPTER 8

Differential equations

8.1. Crash course on differential equations

8.1.1. Motivation. In fluid dynamics, one often uses one of two possible specifications of the flow field in question (or both; whatever description is more convenient for the task at hand): the *Eulerian* or the *Lagrangian* specification.

- In the *Lagrangian* specification, one labels fluid parcels in some way, say by its centre of mass $\vec{x}_0 \in \mathbb{R}^n$ at some fixed initial time t_0. (Here n would reasonably be 3 for a physical problem, although we prefer to draw two-dimensional pictures and symmetries in the problems under consideration may also result in models with $n < 3$.) One then follows these parcels in space as time evolves. Mathematically, this is described by a number $\vec{X}(\vec{x}_0, t)$, giving the location of the centre of mass of the fluid parcel $\vec{x}_0$ at time t (see Figure 68). Altogether, $\vec{X}$ is a function

  $$\vec{X}\colon \{\text{fluid parcels}\} \times \{\text{time}\} \to \mathbb{R}^n.$$

  Our choice of indexing the fluid parcels by their position at time t_0 implies that $\vec{x}_0 = \vec{X}(\vec{x}_0, t_0)$ for all $\vec{x}_0$. Moreover, although we do allow for individual fluid parcels to propagate through space as the time t evolves, we assume that at no time t two fluid parcels occupy the same point in space. Hence, at any time t, the equation

  $$\vec{X}(\vec{x}_0, t) \stackrel{!}{=} \vec{X}(\vec{x}_1, t)$$

  is only satisfied if $\vec{x}_0 = \vec{x}_1$. Phrased slightly differently, one may say that the equation

  $$\vec{X}(\vec{x}_0, t) \stackrel{!}{=} \vec{y} \tag{8.1}$$

  with fixed time t and fixed position $\vec{y} \in \mathbb{R}^n$ has at most one solution $\vec{x}_0$.
- In the *Eulerian* specification, one attaches a velocity vector $\vec{u}(\vec{x}, t)$ to each point $\vec{x}$ in space at a given time t, getting a function

  $$\vec{u}\colon \Omega \to \mathbb{R}^n, \quad \text{where} \quad \Omega \subseteq \{\text{points in space}\} \times \{\text{time}\};$$

  see Figure 69.

The Eulerian and Lagrangian specifications are related by the following equation, which plainly states that the velocity $\frac{\partial \vec{X}}{\partial t}(\vec{x}_0, t)$ of the fluid parcel $\vec{x}_0$ at time t is given

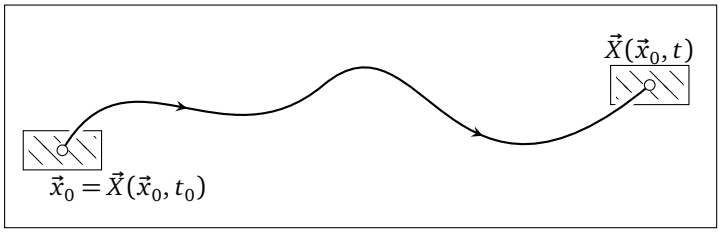


Figure 68. Lagrangian specification of a flow field.

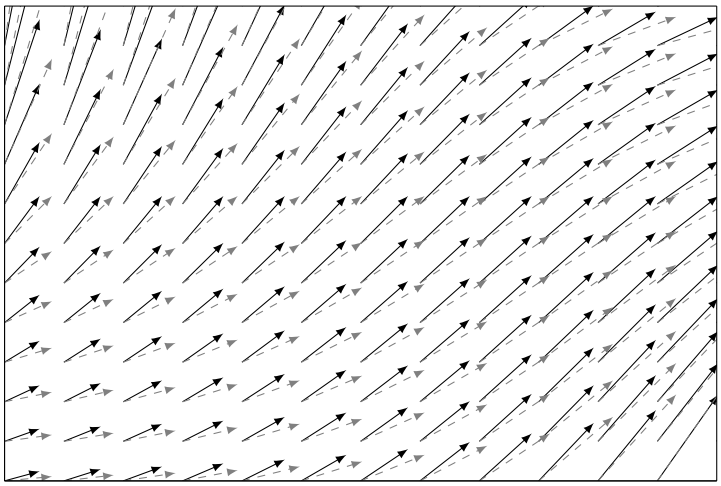


Figure 69. Eulerian specification of a flow field. The picture shows the field at two different points in time (solid vs dashed arrows).

by the velocity field $\vec{u}$ evaluated at the the current location $\vec{X}(\vec{x}_0, t)$ of (the centre of mass of) the fluid parcel at time t:

$$\frac{\partial \vec{X}}{\partial t}(\vec{x}_0, t) = \vec{u}(\vec{X}(\vec{x}_0, t), t). \tag{8.2}$$

Given $\vec{X}$ in the Lagrangian specification, we may compute $\vec{u}$ by doing the following: for any given time t, and any point $\vec{y}$ in space where we have fluid (elsewhere we need not consider the fluid velocity $\vec{u}$), we may solve the equation (8.1) for $\vec{x}_0$. Then $\vec{u}(\vec{y}, t)$ is given by the left hand side of (8.2).

Now suppose that we are given $\vec{u}$ in the Eulerian specification and intend to compute $\vec{X}$. To simplify, we now focus on a fixed fluid parcel $\vec{x}_0$. With this in mind, we now drop the reference to $\vec{x}_0$ from our notation, thus writing $\vec{X}(t)$ for $\vec{X}(\vec{x}_0, t)$. As we now just have one variable (time t), the partial derivative

$$\frac{\partial \vec{X}}{\partial t}$$

is just written in the usual form

$$\frac{\mathrm{d}\vec{X}}{\mathrm{d}t}, \quad \text{or simply} \quad \dot{\vec{X}}.$$

The equation (8.2) then takes the shape

$$\dot{\vec{X}}(t) = \vec{u}(\vec{X}(t), t). \tag{8.3}$$

By writing $\vec{X}_1, \ldots, \vec{X}_n$ for the components of $\vec{X}$ and similarly for the components of $\vec{u}$, we may write (8.3) in the form

$$\begin{cases} \dot{X}_1(t) = u_1(X_1(t), \ldots, X_n(t), t), \\ \qquad\vdots \\ \dot{X}_n(t) = u_n(X_1(t), \ldots, X_n(t), t). \end{cases}$$

This is known as a ***system of (explicit) ordinary first-order differential equations***.[1]

We note that systems of differential equations can also be used to handle higher-order differential equations:

Example. Consider the differential equation

$$\ddot{Y}(t) \stackrel{!}{=} u(Y(t), \dot{Y}(t), t).$$

Suppose that Y satisfies the above differential equation. We let $X_1 = Y$ and $X_2 = \dot{Y}$. Then $\ddot{Y} = \dot{X}_2$. We let $v(x_1, x_2, t) = x_2$. Then $\vec{X} = (X_1, X_2)$ satisfies the following system of differential equations:

$$\begin{cases} \dot{X}_1(t) \stackrel{!}{=} v(X_1(t), X_2(t), t) = X_2(t), \\ \dot{X}_2(t) \stackrel{!}{=} u(X_1(t), X_2(t), t). \end{cases}$$

Conversely, if (X_1, X_2) solves the above, then $Y = X_1$ solves our original second-order differential equation.

The above procedure for re-writing higher-order differential equations as systems of first-order differential equations, works in general (at least for *explicit* differential equations). Consequently, in the literature, methods for solving differential equations are often only phrased in terms of first-order systems. A reader who encounters higher-order differential equations (or even systems thereof) must then be aware of the fact that these cases are also covered.

We now simplify the situation by assuming that $n = 1$. As pointed out above, this should be viewed as an unpleasantly sharp restriction, because we lose the ability of treating higher-order differential equations. However, for the introductory nature

[1] To explain the many words used here: 'system' refers to there being multiple equations, 'explicit' means that the highest-order derivative is isolated on one side of each of the equations, 'first-order' means that only first derivatives, but not higher derivatives of our sought-after function appear in the equations. The word 'ordinary' is used to indicate that one seeks a function of one variable (t here), rather than a function in many variables.

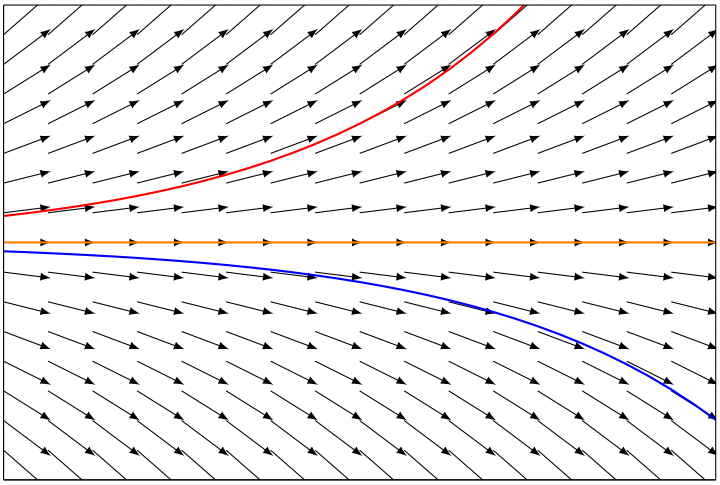


Figure 70. Illustration of Example 8.1 with $\lambda = 1/2$. The orange graph shows the constant zero solution. The red graph shows the solution $t \mapsto \exp(t/2)$ and the blue graph shows $t \mapsto -(1/3)\exp(t/2)$. The arrows drawn here have slope $u(x,t) = x/2$. Note that this figure, albeit similar looking, is morally different from Figure 69: the corresponding figure would have to be one-dimensional, because our spacial coordinate $\vec{x} = x$ lives in $\mathbb{R}$ for the purpose of Example 8.1. The second dimension in our present picture comes from also considering t, which is only barely visible in Figure 69 by means of the differently coloured arrows.

of the present discussion, this restriction seems appropriate. Our system (8.3) now takes the form

$$\dot{X}(t) = u(X(t), t). \tag{8.4}$$

8.1.2. Examples. We now consider some examples in an *ad-hoc* fashion. The main goal here is to get a basic feeling for what to expect. Our exposition is modelled on Jänich's beautiful arrangement [5, Kapitel 4], although it should be noted that he has more time and space for a much more vivid discussion. (Any readers who are capable of speaking German are highly encouraged to check out [5].)

Example 8.1. For fixed $\lambda \in \mathbb{R}$, we consider the differential equation

$$\dot{X} \stackrel{!}{=} \lambda X.$$

This corresponds to taking $u(x,t) = \lambda x$ (independent of t) in (8.4). Imagine this system to describe a flow and consider time $t = 0$. We may be interested in finding the path of every single fluid parcel in $\mathbb{R}$. Hence, we are interested in solving the above differential equation subject to the initial value condition

$$X(0) \stackrel{!}{=} x_0$$

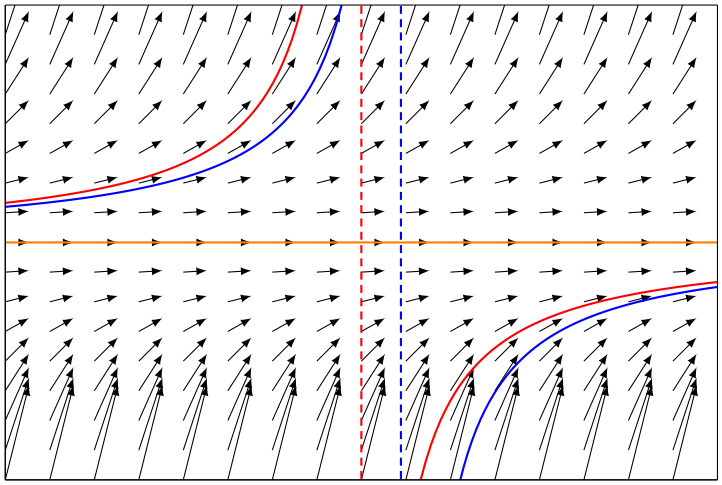


Figure 71. Illustration of Example 8.2. The orange graph shows the constant zero solution. The red graph shows the solution $t \mapsto -t^{-1}$ and the blue graph shows $t \mapsto -(t-1/3)^{-1}$.

for every fixed $x_0 \in \mathbb{R}$. If $x_0 = 0$, then we see that the constant zero function $t \mapsto 0$ solves out problem. On the other hand, if $x_0 = 1$, then $t \mapsto \exp(\lambda t)$ does the job. Actually, this generalises: for any x_0, the solution to our problem is $t \mapsto x_0 \exp(\lambda t)$. One can show that these are all the solutions there are.

Example 8.2. Consider the differential equation

$$\dot{X} \stackrel{!}{=} X^2.$$

This corresponds to taking $u(x,t) = x^2$ (independent of t) in (8.4). Again, we are also interested in finding solutions satisfying an initial value condition $X(0) \stackrel{!}{=} x_0$ for any $x_0 \in \mathbb{R}$ we may choose. Once more, we notice that for $x_0 = 0$, the constant zero function does the job. On the other hand, what are our solutions for $x_0 \neq 0$? We make the *ansatz* $X(t) = t^n$ for some n to be chosen later. Then

$$nt^{n-1} = \dot{X}(t) \stackrel{!}{=} X(t)^2 = (t^n)^2 = t^{2n}.$$

Upon comparing exponents, we see that we ought to have $n = -1$, but the resulting equation then yields

$$-t^{-2} \stackrel{!}{=} t^{2(-1)},$$

which does not work. However, this can be fixed by taking $X(t) = -t^{-1}$ instead. Then, indeed,

$$\dot{X}(t) = -t^{-1-1}(-1) = t^{-2} = (t^{-1})^2 = (-t^{-1})^2 = X(t)^2.$$

Hence, we have found a solution to our differential equation. However, this solution is not defined at $t = 0$ (division by zero!). To fix this, we use a feature of our differential

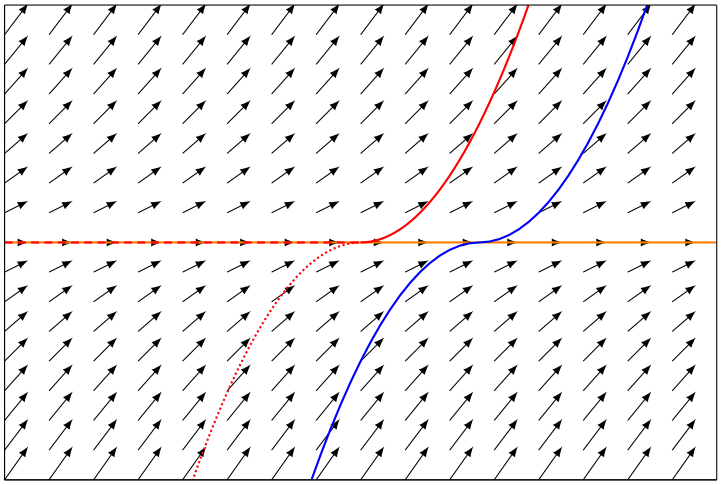


Figure 72. Illustration of Example 8.3. One sees that initial condition may not determine a unique solutions. (Look at the intersection of the red and orange graphs.)

equation: it does not depend on t in the sense that the function u defined above is independent of t. Such differential equations are called ***autonomous***. They have the following property: if X satisfies the autonomous differential equation

$$\dot{X}(t) \stackrel{!}{=} u(X(t), t),$$

then so does the (right-)shifted function $X_{\to t_0}\colon t \mapsto X(t - t_0)$ for any $t_0 \in \mathbb{R}$. Indeed,

$$\dot{X}_{\to t_0}(t) \stackrel{(\dagger)}{=} \dot{X}(t - t_0) = u(X(t - t_0), t - t_0) \stackrel{(\ddagger)}{=} u(X(t - t_0), t) = u(X_{\to t_0}(t), t),$$

where (†) follows by the chain rule and (‡) by autonomy of the differential equation. The upshot of this is, that $t \mapsto -(t - t_0)^{-1}$ solves our differential equation. Moreover, by choosing t_0 appropriately, we can arrange that this function takes an arbitrarily chosen value $x_0 \neq 0$ at $t = 0$. (Indeed, take $t_0 = x_0^{-1}$.) With more work, one could see that we have, in fact, found all solutions this way.

Example 8.3. Consider the differential equation

$$\dot{X} \stackrel{!}{=} \sqrt{|X|}.$$

This corresponds to taking $u(x, t) = \sqrt{|x|}$ (independent of t) in (8.4). Again, we find that the constant zero function is a solution, and by a similar *ansatz* as in the previous example we find that

$$t \mapsto \begin{cases} -(t - t_0)^2 & \text{if } t < t_0, \\ (t - t_0)^2 & \text{if } t \geq t_0 \end{cases}$$

is a solution. Perhaps surprisingly, though, also

$$t \mapsto \begin{cases} 0 & \text{if } t < t_0, \\ (t-t_0)^2 & \text{if } t \geq t_0 \end{cases}$$

is a solution. (One can find more, actually.) This solution coincides with the constant zero solution for times $t \leq t_0$ and then abruptly changes its behaviour.

The above examples show that, even in simple looking cases, solutions to differential equations of the type (8.4)

(1) may exist on all of $\mathbb{R}$; (Example 8.1)
(2) may exist only on subintervals of $\mathbb{R}$; (Example 8.2)
(3) may not be unique in general. (Example 8.3)

8.1.3. What to expect? Even innocently looking differential equations can be quite difficult. For instance, taking $u(x,t) = f(t)$ for some function f in (8.4) means finding some function X such that $\dot{X}(t) = f(t)$, which essentially means $X(t) = \int_{t_0}^{t} f(\tau)\,\mathrm{d}\tau$ for some $t_0 \in \mathbb{R}$. This means, that solving such simple differential equations is already equivalent to computing anti-derivatives. As we have seen in § 0.7, the latter is something which we know how to do for certain examples, but have by no means completely mastered.

When generalising to functions of many variables, thus considering ***partial differential equations***, the general picture looks quite dim. Here the understanding of non-linear partial differential equations is very unsatisfactory. In the case of ordinary differential equations, more can be said, but often one cannot solve the differential equations in question explicitly in the sense that one can write down a simple formula involving only 'well-known functions'. Therefore, much work in the subject concentrates on two aspects: (1) settle existence and uniqueness of solutions of the differential equations in question. If this works, then one can take the following viewpoint (2):

> Differential equations *define* functions. What can one learn about the functions thus obtained, *without* having to find an explicit formula for them?

Unfortunately, in this course, we lack the time to pursue these goals. Instead, we take a more pragmatic approach and show some tricks which, in certain very easy situations, lead to a closed-form expression for solutions of differential equations; the hope here being that a potential reader may gain some practical tools for following easy examples encountered during their studies. Any more advanced examples then require turning away from the current notes and plunging oneself into the literature.

8.1.4. Tricks for solving some differential equations. The first trick we wish to discuss concerns the solution of so-called ***differential equations with separated***

variables. These are differential equations of the form (8.4) where $u(x,t)$ factors in the form $f(x)g(t)$. That is,

$$\dot{X}(t) \stackrel{!}{=} f(X(t))g(t).$$

First, suppose that x_0 is a point where f vanishes: $f(x_0) = 0$. Then the constant function $t \mapsto x_0$ constitutes a solution to the above differential equation. Next, consider a point x_0 where f does not vanish and we look for solutions X with $X(t_0) = x_0$ for some time t_0. By rearranging the above equation, we find that

$$\frac{\dot{X}(t)}{f(X(t))} = g(t).$$

We tacitly assuming that f is continuous, so that $f(X(t))$ is also non-zero for t close to t_0. Then we integrate (assuming also that g is suitably nice so that integration is possible):

$$\int_{t_0}^{\tau} \frac{\dot{X}(t)}{f(X(t))}\,\mathrm{d}t = \int_{t_0}^{\tau} g(t)\,\mathrm{d}t.$$

On the left hand side, we can use integration via substitution, substituting away X. This yields

$$\int_{X(t_0)}^{X(\tau)} \frac{1}{f(x)}\,\mathrm{d}x = \int_{t_0}^{\tau} g(t)\,\mathrm{d}t.$$

If F is an anti-derivative for $x \mapsto 1/f(x)$, then

$$F(X(\tau)) - F(x_0) = \int_{t_0}^{\tau} g(t)\,\mathrm{d}t.$$

Assuming that F is invertible in a neighbourhood of x_0, we can now recover $X(\tau)$ as

$$X(\tau) = F^{-1}\left(\int_{t_0}^{\tau} g(t)\,\mathrm{d}t + F(x_0)\right).$$

One can find suitable assumptions guaranteeing this invertibility, and then show that the above formula really yields a solution to the differential equation in question, but we refrain from doing this. Instead, we show some examples of this formula in action.

Example 8.4. Consider the differential equation

$$\dot{X}(t) \stackrel{!}{=} X(t)t.$$

Here $f(x) = x$ and $g(t) = t$. Since f vanishes at zero, we see that the constant zero function is a solution. Next, we look for non-zero solutions. We must find an anti-derivative F of $x \mapsto 1/f(x) = 1/x$. One such anti-derivative is $F\colon \mathbb{R}\setminus\{0\} \to \mathbb{R}$, $x \mapsto \log|x|$. (The others are obtained by adding arbitrary constants to this.) The function F, is not bijective (x and $-x$ get mapped to the same value, because of the absolute value sign). It does, however, become bijective when restricting the

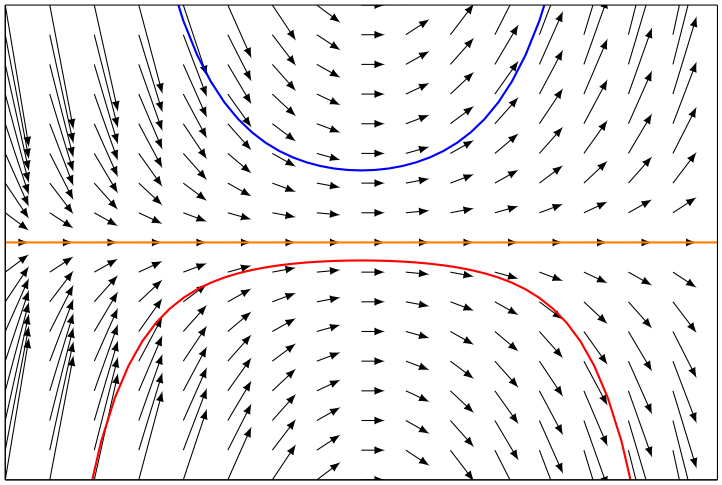


Figure 73. Illustration of Example 8.4.

domain of definition either to the negative numbers or to the positive numbers. The corresponding inverse function F^{-1} is either $\mathbb{R}_- \to \mathbb{R}$, $y \mapsto -\exp(y)$, or $\mathbb{R}_+ \to \mathbb{R}$, $y \mapsto \exp(y)$, respectively. Consequently,

$$X(\tau) = F^{-1}\left(\int_{t_0}^{\tau} g(t)\,\mathrm{d}t + F(x_0)\right) = \pm\exp\left(\frac{1}{2}\tau^2 - \frac{1}{2}t_0^2 + \log|x_0|\right).$$

Here our construction shows that the sign is chosen such that $\pm\exp(\log|x_0|) = x_0$. Hence,

$$X(t) = x_0 \exp\left(\frac{1}{2}(t^2 - t_0^2)\right).$$

Let us check that this has worked. First, $X(t_0) = x_0\exp(0) = x_0$, as desired. Second, by computing derivatives, we see that, indeed, $X'(t) = X(t)t$. Hence, we have solved the differential equation in question.

Example. The differential equations considered in Example 8.1, Example 8.2 and Example 8.3 are all given in separated form with $g(t) = 1$. The interested reader can try to solve them using the above procedure.

Example 8.5. We consider the problem of water flowing out of a cylindrical tank that has a small hole in its bottom (see Figure 74). Under some simplifying assumptions like the hole being small with respect to the diameter of the cylinder one can derive ***Toricelli's law*** from the ***Bernoulli equation*** which the reader might know from fluid dynamics. Toricelli's law states that the function $v(t)$ modelling the velocity of the leaking water at the hole satisfies $v(t) = \sqrt{2gH(t)}$, where $H(t)$ measures the level of the water ($H(t) = 0$ means zero level, the water having drained.), and g denotes Earth's gravity. (In practice, the out-flowing water jet curves inwards into itself, thus reducing the flow rate. For this reason, one often sees a correction factor of roughly

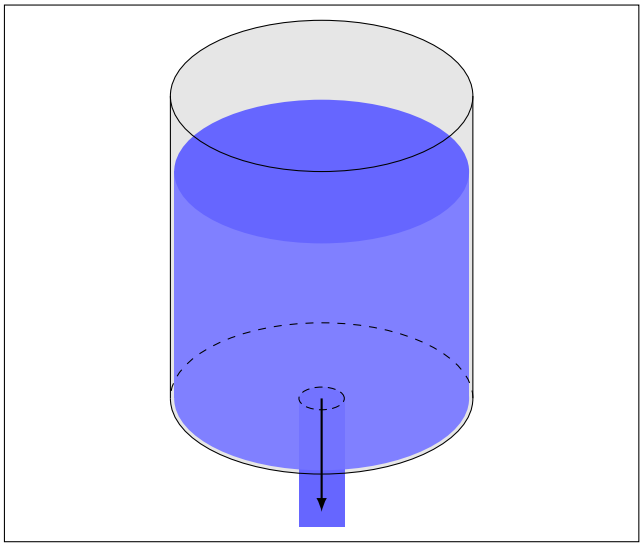


Figure 74. Water leaking out of a cylindrical tank; see Example 8.5.

the size 0.6 multiplied to the right hand side.) Since the level $H(t)$ is proportional to the amount of water in the tank and the velocity of the leaking water is proportional to the of the derivative of the amount of water leaving the tank, conservation of mass shows that

$$\dot{H}(t) = -(\text{const.})\sqrt{H(t)},$$

where the constant depends on the relative size of the cross-section of the tank and the size of the hole, as well as g. Closer inspection shows that we have essentially already solved such a differential equation in Example 8.3 (albeit with different sign choice). The interested reader can work out the details, or consult [7, § 1.3, page 16, Example 7].

As a last technique for this section, we discuss linear, first-order, ordinary differential equations. That is, we wish to solve

$$\dot{X}(t) \stackrel{!}{=} X(t)g(t) + b(t), \tag{8.5}$$

for certain functions g and b. If b is the zero function, then the above is a differential equation with separated variables, but in this case it is easy to guess the solution anyway. Indeed, if G is an anti-derivative of g (i.e., $\dot{G} = g$), then, for any constant $c \in \mathbb{R}$, the function $X_0\colon t \mapsto c\exp(G(t))$ satisfies

$$\dot{X}_0(t) = X_0(t)\dot{G}(t) = X_0(t)g(t).$$

Hence, X_0 solves (8.5) with $b(t)$ replaced by zero.

Now, in order to actually solve (8.5) one can use a trick called ***variation of constants***. The idea is to start with the 'almost'-solution X_0 and 'twist' it slightly in order to produce a solution to (8.5). More precisely, one makes the *ansatz* of taking

the constant c above to be a function of t. Hence, we let $X(t) = c(t)\exp(G(t))$ and try to choose $c(t)$ as to make X solve (8.5). This means that we ought to have

$$\begin{aligned} X(t)g(t) + b(t) \overset{!}{=} \dot{X}(t) &= \frac{\mathrm{d}}{\mathrm{d}t}(c(t)\exp(G(t))) = \dot{c}(t)\exp(G(t)) + c(t)\frac{\mathrm{d}}{\mathrm{d}t}\exp(G(t)) \\ &= \dot{c}(t)\exp(G(t)) + c(t)\exp(G(t))g(t) \\ &= \dot{c}(t)\exp(G(t)) + X(t)g(t). \end{aligned}$$

Rearranging now shows that, in order to succeed with our *ansatz*, we ought to have

$$b(t) \overset{!}{=} \dot{c}(t)\exp(G(t)),$$

or (equivalently),

$$\dot{c}(t) \overset{!}{=} b(t)\exp(-G(t)).$$

By integrating, we arrive at

$$c(t) = \int_{t_0}^{t} b(\tau)\exp(-G(\tau))\,\mathrm{d}\tau + \text{const.},$$

where t_0 is a basically arbitrary real number. (Here 'basically arbitrary' means that we may pick and value for it, but should ensure that the integrand can be integrated on the interval $[t_0, t]$.)

Now under sufficiently general assumptions, one can actually show (which we do not do here) that the solutions thus obtained,

$$X(t) = \left(\int_{t_0}^{t} b(\tau)\exp(-G(\tau))\,\mathrm{d}\tau + \text{const.}\right)\exp(G(t))$$

constitute all solutions to (8.5).

8.2. A glimpse at partial differential equations: the diffusion equation

In this section we speak of functions u of some "spacial coordinate" $\vec{x}$ (in $\mathbb{R}^n$ with $n \in \{1, 2, 3\}$) and some "time coordinate" t. We usually write this as $u(\vec{x}, t)$, but when using the nabla operator ∇ or the divergence, then we think of t as being *fixed* and u as the function $\vec{x} \mapsto u(\vec{x}, t)$. Hence,

$$\nabla u = \frac{\partial u}{\partial x_1} + \ldots + \frac{\partial u}{\partial x_n} \quad \text{and not} \quad = \frac{\partial u}{\partial x_1} + \ldots + \frac{\partial u}{\partial x_n} + \frac{\partial u}{\partial t}.$$

(Beware, though, that the above formula is a consequence of notational *convention* and not of logic.) Accordingly, the divergence of a vector field also only takes derivatives with respect to the coordinates of $\vec{x}$ into account and ignores the t-variable and the Laplace operator Δ is given by

$$\Delta u = \frac{\partial^2 u}{\partial x_1^2} + \ldots + \frac{\partial^2 u}{\partial x_n^2}.$$

We sometimes write

$$\frac{\partial u}{\partial t} = u_t, \quad \frac{\partial u}{\partial x_i} = u_{x_i} \quad \text{and} \quad \frac{\partial^2 u}{\partial x_i \partial x_j} = u_{x_i x_j} \quad \text{for } i, j = 1, \dots, n.$$

8.2.1. Deriving the diffusion equation. We sketch how one may obtain Fick's second law governing the diffusion of particles from Fick's first law and Gauß's theorem. Throughout we tacitly employ the physicist's code that all appearing functions are as nice as they need to be to make the argument work (for instance, being sufficiently smooth).

Suppose we have some region $U \subseteq \mathbb{R}^3$ in space which contains some fluid with some dye. We model this by some function $u\colon U \times \mathbb{R} \to \mathbb{R}$, where $u(\vec{x}, t)$ gives the concentration (dye particles per unit region in space) of the dye at the point $\vec{x} \in U$ and time $t \in \mathbb{R}$. Later we shall also introduce further restrictions, such as restricting to $t \geq 0$.

We now look at some nice compact "test region" $B \subseteq U$ (a ball or a box or something similar). The total amount of dye in B at a given point t in time is given by

$$\iiint_B u(\cdot, t)\, \mathrm{d}V.$$

The rate of change of this is

$$\frac{\mathrm{d}}{\mathrm{d}t} \iiint_B u(\cdot, t)\, \mathrm{d}V = \iiint_B \frac{\partial}{\partial t} u(\cdot, t)\, \mathrm{d}V. \tag{8.6}$$

It is physically obvious, that any change in concentration inside B must arise from dye particles entering (or leaving) the test region B. Let the diffusion flux be the vector field $\vec{J}\colon U \times \mathbb{R} \to \mathbb{R}^3$ measuring how many dye particles enter per unit area. More precisely, we require things to be set up in such a way that[2]

$$\frac{\mathrm{d}}{\mathrm{d}t} \iiint_B u(\cdot, t)\, \mathrm{d}V = -\oiint_{\partial B} \vec{J}(\cdot, t) \cdot \mathrm{d}\vec{A}.$$

On the other hand, ***Fick's first law*** states that this diffusive flux $\vec{J}$ is proportional to the negative gradient of the concentration:

$$\vec{J} \propto -\nabla u.$$

Note that the above coincides with our intuition: diffusion happens most quickly from the direction with respect to which the concentration difference is maximised;

[2]The confusing minus sign is an aritfact of our orientation convention. The normal points outwards, so without the minus sign, the right hand side counts the dye particles *leaving* the region B per time unit, yet the left hand side counts the particles *entering* B per time unit. The equation in question is a special case of so-called ***continuity equations*** that describe the transport of conserved quantities (dye particles, in the present case).

however, this direction is precisely given by minus the gradient (recall Proposition 5.5). In the following we shall assume that the above constant of proportionality equals one, so that the above becomes an equation. Hence,

$$\frac{\mathrm{d}}{\mathrm{d}t}\iiint_B u(\,\cdot\,,t)\,\mathrm{d}V = \oiint_{\partial B} \nabla u(\,\cdot\,,t)\cdot \mathrm{d}\vec{A}.$$

On applying Gauß's divergence theorem (Theorem 7.7) to the right hand side, we find that

$$\oiint_{\partial B} \nabla u(\,\cdot\,,t)\cdot \mathrm{d}\vec{A} = \iiint_B \operatorname{div}(\nabla u(\,\cdot\,,t))\,\mathrm{d}V = \iiint_B \Delta u(\,\cdot\,,t)\,\mathrm{d}V.$$

Upon pluggin this into the previous equation and recalling (8.6) from above, we infer

$$\iiint_B \frac{\partial}{\partial t} u(\,\cdot\,,t)\,\mathrm{d}V = \iiint_B \Delta u(\,\cdot\,,t)\,\mathrm{d}V.$$

By varying the test region B, it transpires that this is only possible if

$$\frac{\partial}{\partial t} u = \Delta u.$$

Writing u_t for the left hand side, we obtain the ***diffusion equation***

$$u_t = \Delta u.$$

By appealing lower-dimensional versions of Gauß's theorem one gets variants of this equation with $U \subseteq \mathbb{R}$ or $U \subseteq \mathbb{R}^2$.

Remark. The same equation also governs *heat* and the derivation is quite similar, replacing our appellation of Fick's first law by Fourier's law of thermal conduction. We omit the details.

8.2.2. Solving the diffusion equation. The Fourier analysis discussed in Chapter 4 can be employed to solve the diffusion equation and match certain boundary conditions. This is a topic we take up in the exercises.

Bibliography

[1] G. Bärwolff. *Höhere Mathematik für Naturwissenschaftler und Ingenieure*. Berlin: Springer, 2017.

[2] Chr. Elsholtz, J. Hatzl, C. Heuberger, and J. Pöschko. Mathematik 1 für ChemikerInnen. Lecture notes. Available online: `https://www.math.tugraz.at/~elsholtz/WWW/lectures/ws19/chemie1/vorlesung.html`, 2017.

[3] Chr. Elsholtz, C. Heuberger, and J. Pöschko. Mathematik 2 für ChemikerInnen. Lecture notes. Available online: `https://www.math.tugraz.at/~elsholtz/WWW/lectures/ss20/chemie2/vorlesung.html`, 2018.

[4] K. Jänich. *Analysis für Physiker und Ingenieure. Funktionentheorie, Differentialgleichungen, spezielle Funktionen*. Berlin: Springer, 2001.

[5] K. Jänich. *Mathematik 1. Geschrieben für Physiker*. Berlin: Springer, 2005.

[6] K. Jänich. *Mathematik 2. Geschrieben für Physiker*. Berlin: Springer, 2011.

[7] E. Kreyszig. *Advanced engineering mathematics*. New York: Wiley, 10th edition, 2020.

[8] R. B. Nelsen. *Proofs without words. II. More exercises in visual thinking*. Washington, DC: MAA, 2000.

[9] M. Stone and P. Goldbart. *Mathematics for physics. A guided tour for graduate students*. Cambridge: Cambridge University Press, 2009.

Index