Calculus II for Statistics Students

Peter Philip*

Lecture Notes

Originally Created for the Class of Spring Semester 2008 at LMU Munich,
Revised and Extended for Several Subsequent Classes

July 22, 2015

Contents

*E-Mail: philip@math.lmu.de

1 Introduction to Basic Topological Notions in Metric Spaces

Basic topological notions include open sets, closed sets, compactness, and convergence, all of which we already encountered in Calculus I is special situations, and all of which we will encounter once again below, now in a more general setting. In an abstract setting, the natural realm for such notions is the class of topological spaces. On the other hand, the special case we are most interested in is the Euclidean space $\mathbb{R}^n$. In a compromise between generality and concreteness, topological notions will be presented in the setting of metric spaces. Metric spaces are a special class of topological spaces that, on the one hand, is sufficiently general to include a wide range of examples and to give a flavor of abstract topological theory and thinking, but, on the other hand, shares many important and typical properties with the Euclidean space $\mathbb{R}^n$ (and its subsets).

Before proceeding to the introduction of a metric space in Sec. 1.2, we first consider our favorite example, the Euclidean space $\mathbb{R}^n$. Actually, it will usually be desirable to study the spaces $\mathbb{R}^n$ and $\mathbb{C}^n$ simultaneously, as this often provides additional useful results without any extra difficulty. Recall that we used the same approach in the, mostly, one-dimensional considerations in Calculus I. As in Calculus I, we will write $\mathbb{K}$ in situations, where we allow $\mathbb{K}$ to be $\mathbb{R}$ or $\mathbb{C}$ (cf. [Phi15a, Not. 6.1]).

1.1 The Space $\mathbb{K}^n$

In the case $\mathbb{K} = \mathbb{R}$, we obtain the so-called *Euclidean space* $\mathbb{R}^n$, which is of particular interest.

Definition 1.1. Let n be a natural number, i.e. $n \in \mathbb{N} = \{1, 2, \dots\}$. By $\mathbb{K}^n$, we mean the set of n-tuples $(z_1, \dots, z_n)$ with real or complex *coordinates* $z_1 \in \mathbb{K}$, $\dots$, $z_n \in \mathbb{K}$ (cf. [Phi15a, Ex. 2.15(c)]). Thus, the set $\mathbb{K}^n$ can be identified with the set of functions defined on the numbers $\{1, \dots, n\}$ with values in $\mathbb{K}$. The elements of $\mathbb{K}^n$ are referred to as *points* or *vectors* (cf. Rem. 1.2 below). For $z = (z_1, \dots, z_n) \in \mathbb{K}^n$, $w = (w_1, \dots, w_n) \in \mathbb{K}^n$, and $\lambda \in \mathbb{K}$, we define the componentwise *addition*

$$z + w := (z_1 + w_1, \dots, z_n + w_n), \tag{1.1a}$$

the componentwise *scalar multiplication*

$$\lambda z := (\lambda z_1, \dots, \lambda z_n), \tag{1.1b}$$

the (*Euclidean* for $\mathbb{K} = \mathbb{R}$) *inner product* or *scalar product*

$$z \cdot w := z_1 \bar{w}_1 + \cdots + z_n \bar{w}_n = \sum_{j=1}^{n} z_j \bar{w}_j \tag{1.1c}$$

(where we recall that, for a complex number w_j, the complex conjugate is denoted by $\bar{w}_j$), and the (*Euclidean* for $\mathbb{K} = \mathbb{R}$) *norm*, *length*, or *absolute value*

$$|z| := \sqrt{z \cdot z} = \sqrt{|z_1|^2 + \cdots + |z_n|^2}. \tag{1.1d}$$

Finally, one defines the (*Euclidean* for $\mathbb{K} = \mathbb{R}$) *distance* between z and w as

$$|z - w| = \sqrt{|z_1 - w_1|^2 + \cdots + |z_n - w_n|^2}. \tag{1.1e}$$

One also finds the notation $\langle z, w\rangle$ instead of $z \cdot w$ for the inner product, and $\|z\|$ instead of $|z|$ for the above norm.

Remark 1.2. The set $\mathbb{K}^n$ together with the componentwise addition and scalar multiplication as defined in Def. 1.1 constitutes a vector space over the field $\mathbb{K}$ (see Ex. A.2(d) of Appendix A – App. A reviews basic notions and results from Linear Algebra). In a slight abuse of notation, one often writes 0 instead of $(0, \ldots, 0)$.

Lemma 1.3. *The norm on $\mathbb{K}^n$ as defined in* (1.1d) *enjoys the following properties:*

(a) *It is* positive definite*, i.e.*

$$|z| \geq 0 \text{ and } \Big(|z| = 0 \;\Leftrightarrow\; z = 0\Big) \qquad \text{for each } z \in \mathbb{K}^n.$$

(b) *It is* homogeneous of degree 1*, i.e.*

$$|\lambda z| = |\lambda||z| \qquad \text{for each } \lambda \in \mathbb{K},\ z \in \mathbb{K}^n.$$

(c) *It satisfies the* triangle inequality*, i.e.*

$$|z + w| \leq |z| + |w| \qquad \text{for each } (z, w) \in \mathbb{K}^n \times \mathbb{K}^n.$$

Proof. (a): $|z|$ is nonnegative, as the square root is nonnegative. If $|z| = 0$, then $|z_1|^2 = \cdots = |z_n|^2 = 0$, i.e. $z_1 = \cdots = z_n = 0$.

(b): One calculates

$$|\lambda z| = \sqrt{\sum_{j=1}^{n} |\lambda z_j|^2} = \sqrt{|\lambda|^2 \sum_{j=1}^{n} |z_j|^2} = |\lambda| \sqrt{\sum_{j=1}^{n} |z_j|^2} = |\lambda||z|. \tag{1.2}$$

(c): The triangle inequality is a bit harder to prove. We will see two different proofs later. In Th. 1.83, we will prove the triangle inequality for general p-norms on $\mathbb{K}^n$ (the so-called Minkowski inequality). The special case $p = 2$ then yields the triangle inequality for the above norm. Moreover, in Sec. 1.7 we will show that, for each general inner product, the definition analogous to (1.1d) does always yield a general norm (see Prop. 1.88) – in particular, this definition always guarantees the triangle inequality. Once again, we obtain the above norm as a special case (see Ex. 1.90). ■

Remark 1.4. Lemma 1.3 shows that the norm on $\mathbb{K}^n$ as defined in (1.1d) is, indeed, a norm in the sense of Def. 1.19 below. This, in turn, implies that the distance on $\mathbb{K}^n$ as defined in (1.1e) is, indeed, a metric in the sense of Def. 1.17 below.

Definition 1.5. Elements $e \in \mathbb{K}^n$ of length 1 are called *unit vectors*. The n unit vectors

$$e_1 := (1, 0, \dots, 0), \quad e_2 := (0, 1, \dots, 0), \quad \dots, \quad e_n := (0, \dots, 0, 1) \tag{1.3}$$

are called the *standard unit vectors*. They form the standard basis of the vector space $\mathbb{K}^n$ over $\mathbb{K}$ (cf. Ex. A.15).

Remark 1.6. For every $z = (z_1, \dots, z_n) \in \mathbb{K}^n$, it holds that

$$z = \sum_{j=1}^{n} z_j e_j = \sum_{j=1}^{n} (z \cdot e_j)\, e_j. \tag{1.4}$$

Notation 1.7. For $x, y \in \mathbb{R}^n$, we write $x < y$ (resp. $x \leq y$) if, and only if, $x_j < y_j$ (resp. $x_j \leq y_j$) for each $j \in \{1, \dots, n\}$.

Remark 1.8. Note that, for each $n \geq 2$, given points $x, y \in \mathbb{R}^n$ might not be comparable. For example, if $x = (1, 0)$, $y = (0, 1)$, and $z = (2, 2)$, then $x < z$, $y < z$, but neither $x < y$ nor $y < x$.

Notation 1.9. A subset I of $\mathbb{R}^n$ is called an n-dimensional *interval* if, and only if, I has the form $I = I_1 \times \cdots \times I_n$, where $I_1, \dots, I_n$ are intervals in $\mathbb{R}$. The lengths $|I_1|, \dots, |I_n|$ are called the edge lengths of I. An interval I is called a *(hyper)cube* if, and only if, all its edge lengths are equal. If $x, y \in \mathbb{R}^n$, $x < y$, then we define the following intervals

$$]x, y[:= \{z \in \mathbb{R}^n : x < z < y\} =]x_1, y_1[\times \cdots \times]x_n, y_n[\qquad \textit{open interval}, \tag{1.5a}$$
$$[x, y] := \{z \in \mathbb{R}^n : x \leq z \leq y\} = [x_1, y_1] \times \cdots \times [x_n, y_n] \qquad \textit{closed interval}, \tag{1.5b}$$
$$[x, y[:= \{z \in \mathbb{R}^n : x \leq z < y\} = [x_1, y_1[\times \cdots \times [x_n, y_n[\qquad \textit{halfopen interval}, \tag{1.5c}$$
$$]x, y] := \{z \in \mathbb{R}^n : x < z \leq y\} =]x_1, y_1] \times \cdots \times]x_n, y_n] \qquad \textit{halfopen interval}. \tag{1.5d}$$

—

Recall the notion of a sequence from [Phi15a, Def. 2.14(b)], the notion of a convergent sequence in $\mathbb{K}$ from [Phi15a, Def. 7.1], and the notion of a Cauchy sequence in $\mathbb{K}$ from [Phi15a, Def. 7.28]. The introduction of the absolute value in $\mathbb{K}^n$ (the Euclidean norm for $\mathbb{K} = \mathbb{R}$) enables us to extend the notion of converging sequences and Cauchy sequences from $\mathbb{K}$ to $\mathbb{K}^n$. One merely has to replace the absolute value in $\mathbb{K}$ by the absolute value in $\mathbb{K}^n$:

Definition 1.10. Let $(z^k) = (z^k)_{k \in \mathbb{N}} = (z^1, z^2, \dots)$ be a sequence in $\mathbb{K}^n$. The sequence (z^k) is defined to be *convergent* with *limit* $a \in \mathbb{K}^n$ (notation: $\lim_{k \to \infty} z^k = a$ or $z^k \to a$ for $k \to \infty$) if, and only if, for each $\epsilon \in \mathbb{R}^+$, there is $N \in \mathbb{N}$ such that $|z^k - a| < \epsilon$ for each $k > N$. Similarly, (z^k) is defined to be a *Cauchy sequence* if, and only if, for each $\epsilon \in \mathbb{R}^+$, there is $N \in \mathbb{N}$ such that $|z^k - z^l| < \epsilon$ for each $k, l > N$.

Remark 1.11. For each $z = (z_1, \ldots, z_n) \in \mathbb{K}^n$, one has the following estimates:

$$\underset{j\in\{1,\ldots,n\}}{\forall} \quad |z_j| \leq \underbrace{|z|}_{\sqrt{|z_1|^2+\cdots+|z_n|^2}} \leq |z_1| + \cdots + |z_n|. \tag{1.6}$$

Theorem 1.12. *Let $(z^k) = (z^k)_{k\in\mathbb{N}}$ be a sequence in $\mathbb{K}^n$, where $z^k = (z_1^k, \ldots, z_n^k)$. The sequence (z^k) is convergent with limit $a = (a_1, \ldots, a_n) \in \mathbb{K}^n$ (resp. a Cauchy sequence) if, and only if, each of the scalar coordinate sequences $(z_j^k)_{k\in\mathbb{N}}$ in $\mathbb{K}$ is convergent with limit $a_j \in \mathbb{K}$ (resp. is a Cauchy sequence in $\mathbb{K}$), $j \in \{1, \ldots, n\}$.*

Proof. Suppose that $(z^k)_{k\in\mathbb{N}}$ is a convergent sequence with limit a. Then, according to Def. 1.10, given $\epsilon \in \mathbb{R}^+$, there is $N \in \mathbb{N}$ such that, for each $k > N$,

$$|z^k - a| < \epsilon. \tag{1.7}$$

Since, by (1.6), (1.7) implies

$$\underset{j\in\{1,\ldots,n\}}{\forall} \quad |z_j^k - a_j| \leq |z^k - a| < \epsilon, \tag{1.8}$$

according to [Phi15a, Def. 7.1], $(z_j^k)_{k\in\mathbb{N}}$ converges to a_j for each $j \in \{1, \ldots, n\}$. Conversely, if $(z_j^k)_{k\in\mathbb{N}}$ converges to a_j for each $j \in \{1, \ldots, n\}$, then, given $\epsilon \in \mathbb{R}^+$, [Phi15a, Def. 7.1] yields $N \in \mathbb{N}$ such that, for each $k > N$,

$$|z_j^k - a_j| < \frac{\epsilon}{n}. \tag{1.9}$$

Since, by (1.6), (1.9) implies

$$|z^k - a| \leq \sum_{j=1}^{n} |z_j^k - a_j| < n\,\frac{\epsilon}{n} = \epsilon, \tag{1.10}$$

$(z^k)_{k\in\mathbb{N}}$ converges to a. The claim regarding Cauchy sequences is proved analogously using [Phi15a, Def. 7.28] and is left as an exercise. ■

Definition 1.13. In generalization of [Phi15a, Def. 7.9], we define a sequence $(z^k)_{k\in\mathbb{N}}$ in $\mathbb{K}^n$, $n \in \mathbb{N}$, to be *bounded* if, and only if, the *set* $\{|z^k| : k \in \mathbb{N}\}$ is bounded in the sense of [Phi15a, Def. 2.24(a)], i.e. if, and only if,

$$\underset{M\in\mathbb{R}_0^+}{\exists} \quad \underset{k\in\mathbb{N}}{\forall} \quad 0 \leq |z^k| \leq M. \tag{1.11}$$

—

Recall the notion of subsequence and reordering from [Phi15a, Def. 7.21].

Lemma 1.14. *Let $(z^k)_{k\in\mathbb{N}}$ be a sequence in $\mathbb{K}^n$ such that $\lim_{k\to\infty} z^k = a \in \mathbb{K}^n$. Then the following holds:*

(a) $(z^k)_{k\in\mathbb{N}}$ *is bounded.*

(b) $\lim_{l\to\infty} z^{k_l} = a$ *for every subsequence* $(z^{k_l})_{l\in\mathbb{N}}$ *of* $(z^k)_{k\in\mathbb{N}}$.

(c) $\lim_{k\to\infty} z^{\phi(k)} = a$ *for every reordering* $(z^{\phi(k)})_{k\in\mathbb{N}}$ *of* $(z^k)_{k\in\mathbb{N}}$.

(d) *If* $\lim_{k\to\infty} z^k = b \in \mathbb{K}^n$, *then* $a = b$. *Thus, the limit of a sequence in* $\mathbb{K}^n$ *is unique (given it exists at all).*

Proof. In each case, the key to the proof is Th. 1.12 and to apply the result that is already known for sequences in $\mathbb{K}$. From Th. 1.12, we know

$$\lim_{k\to\infty} z_j^k = a_j \quad \text{for each } j \in \{1,\ldots,n\}. \tag{1.12}$$

(a): Since convergent sequences in $\mathbb{K}$ are bounded according to [Phi15a, Prop. 7.10(b)], (1.12) implies the existence of $M_1,\ldots,M_n \in \mathbb{R}_0^+$ such that $0 \le |z_j^k| \le M_j$ for each $k \in \mathbb{N}$ and each $j \in \{1,\ldots,n\}$. Since $|z^k| \le |z_1^k| + \cdots + |z_n^k|$ by (1.6), one has

$$\bigwedge_{k\in\mathbb{N}} \quad 0 \le |z^k| \le \sum_{j=1}^{n} M_j \in \mathbb{R}_0^+, \tag{1.13}$$

showing that $(z^k)_{k\in\mathbb{N}}$ is bounded.

(b): If $(z^{k_l})_{l\in\mathbb{N}}$ is a subsequence of $(z^k)_{k\in\mathbb{N}}$, then $(z_j^{k_l})_{l\in\mathbb{N}}$ is a subsequence of $(z_j^k)_{k\in\mathbb{N}}$ for each $j \in \{1,\ldots,n\}$. Thus, (1.12) together with the result [Phi15a, Prop. 7.23] on sequences in $\mathbb{K}$ implies that $\lim_{l\to\infty} z_j^{k_l} = a_j$. We now apply Th. 1.12 in the opposite direction to get $\lim_{l\to\infty} z^{k_l} = a$ as claimed.

(c): If $(z^{\phi(k)})_{k\in\mathbb{N}}$ is a reordering of $(z^k)_{k\in\mathbb{N}}$, then $(z_j^{\phi(k)})_{k\in\mathbb{N}}$ is a reordering of $(z_j^k)_{k\in\mathbb{N}}$ for each $j \in \{1,\ldots,n\}$. Thus, (1.12) together with the result [Phi15a, Prop. 7.23] on sequences in $\mathbb{K}$ implies $\lim_{k\to\infty} z_j^{\phi(k)} = a_j$. As before, we now apply Th. 1.12 in the opposite direction to get $\lim_{k\to\infty} z^{\phi(k)} = a$ as claimed.

(d): Suppose $\lim_{k\to\infty} z^k = b \in \mathbb{K}^n$. Then, once more by Th. 1.12,

$$\lim_{k\to\infty} z_j^k = b_j \quad \text{for each } j \in \{1,\ldots,n\}. \tag{1.14}$$

Since limits of sequences in $\mathbb{K}$ are unique by [Phi15a, Prop. 7.10(a)], (1.14) together with (1.12) yields $a_j = b_j$ for each $j \in \{1,\ldots,n\}$, i.e. $a = b$. ■

Lemma 1.15. *Let* $(z^k)_{k\in\mathbb{N}}$, $(w^k)_{k\in\mathbb{N}}$ *be sequences in* $\mathbb{K}^n$ *such that* $\lim_{k\to\infty} z^k = a \in \mathbb{K}^n$, $\lim_{k\to\infty} w^k = b \in \mathbb{K}^n$. *Moreover, let* $\lambda, \mu \in \mathbb{K}$. *One then has the following convergences:*

(a) $\lim_{k\to\infty}(\lambda z^k + \mu w^k) = \lambda a + \mu b$.

(b) $\lim_{k\to\infty}(z^k \cdot w^k) = a \cdot b$.

(c) $\lim_{k\to\infty} |z^k| = |a|$.

Proof. As in the previous lemma, we employ Th. 1.12 to get

$$\lim_{k\to\infty} z_j^k = a_j \quad \text{for each } j \in \{1,\dots,n\}, \tag{1.15a}$$

$$\lim_{k\to\infty} w_j^k = b_j \quad \text{for each } j \in \{1,\dots,n\}. \tag{1.15b}$$

(a): As we already know from [Phi15a, (7.11a),(7.11b)] that the corresponding formula holds for sequences in $\mathbb{K}$, (1.15) implies $\lim_{k\to\infty}(\lambda z_j^k + \mu w_j^k) = \lambda a_j + \mu b_j$ for each $j \in \{1,\dots,n\}$. Once more applying Th. 1.12 provides $\lim_{k\to\infty}(\lambda z^k + \mu w^k) = \lambda a + \mu b$.

(b): Note that it suffices to consider the case $\mathbb{K} = \mathbb{C}$, as this includes the case $\mathbb{K} = \mathbb{R}$ as a special case. Due to (1.15) and identities for limits of complex sequences, we compute

$$\begin{aligned} a \cdot b &= \sum_{j=1}^n a_j\,\bar{b}_j = \sum_{j=1}^n \lim_{k\to\infty} z_j^k\,\overline{\lim_{k\to\infty} w_j^k} \\ &\overset{\text{[Phi15a, (7.11c),(7.11f),(7.16a)]}}{=} \lim_{k\to\infty} \sum_{j=1}^n z_j^k\,\overline{w_j^k} = \lim_{k\to\infty}(z^k \cdot w^k). \end{aligned} \tag{1.16}$$

(c) follows from (b) by making use of the continuity of the square root function (note that $(z^k \cdot z^k)$ is a sequence of *real* numbers):

$$\lim_{k\to\infty} |z^k| = \lim_{k\to\infty} \sqrt{z^k \cdot z^k} \overset{\text{[Phi15a, Th. 7.37, Th. 7.72(a)]}}{=} \sqrt{\lim_{k\to\infty}(z^k \cdot z^k)} = \sqrt{a \cdot a} = |a|. \tag{1.17}$$

This concludes the proof of the lemma. ■

Theorem 1.16. (a) *A sequence in $\mathbb{K}^n$ is convergent if, and only if, it is a Cauchy sequence.*

(b) *Bolzano-Weierstrass Theorem: Every bounded sequence in $\mathbb{K}^n$ has a convergent subsequence.*

Proof. (a): A sequence in $\mathbb{K}^n$ is convergent if, and only if, each of its coordinate sequences is convergent (Th. 1.12). As each coordinate sequence is a sequence in $\mathbb{K}$, we know from [Phi15a, Th. 7.29] that each coordinate sequence is convergent if, and only if, it is a Cauchy sequence. Finally, again by Th. 1.12, the coordinate sequences are all Cauchy sequences if, and only if, the original sequence in $\mathbb{K}^n$ is a Cauchy sequence, thereby establishing the case.

(b): If $(z^k)_{k\in\mathbb{N}}$ is bounded, then, due to (1.6), each coordinate sequence $(z_j^k)_{k\in\mathbb{N}}$, $j \in \{1,\dots,n\}$, is bounded. We prove by induction over $\{1,\dots,n\}$ that, for each $j \in \{1,\dots,n\}$, there is a subsequence $(y^{k,j})_{k\in\mathbb{N}}$ of $(x^k)_{k\in\mathbb{N}}$ such that the coordinate sequences $(y_\alpha^{k,j})_{k\in\mathbb{N}}$ converge for each $\alpha \in \{1,\dots,j\}$. Base Case ($j = 1$): Since $(z_1^k)_{k\in\mathbb{N}}$ is a bounded sequence in $\mathbb{K}$, the Bolzano-Weierstrass theorem for sequences in $\mathbb{K}$ (cf. [Phi15a, Prop. 7.26, Th. 7.27]) yields the existence of a convergent subsequence of $(z_1^k)_{k\in\mathbb{N}}$. This provides us with the needed subsequence $(y^{k,1})_{k\in\mathbb{N}}$ of $(z^k)_{k\in\mathbb{N}}$. Now suppose that $1 < j \leq n$. By induction, we already have a subsequence $(y^{k,j-1})_{k\in\mathbb{N}}$ of $(z^k)_{k\in\mathbb{N}}$ such that the coordinate sequences $(y_\alpha^{k,j-1})_{k\in\mathbb{N}}$ converge for each $\alpha \in \{1,\dots,j-1\}$. As $(y_\alpha^{k,j-1})_{k\in\mathbb{N}}$ is

a subsequence of the bounded $\mathbb{K}$-valued sequence $(z_\alpha^k)_{k\in\mathbb{N}}$, by the Bolzano-Weierstrass theorem for sequences in $\mathbb{K}$, it has a convergent subsequence. This provides us with the needed subsequence $(y^{k,j})_{k\in\mathbb{N}}$ of $(y^{k,j-1})_{k\in\mathbb{N}}$, which is then also a subsequence of $(z^k)_{k\in\mathbb{N}}$. Moreover, for each $\alpha \in \{1,\dots,j-1\}$, $(y_\alpha^{k,j})_{k\in\mathbb{N}}$ is a subsequence of the convergent sequence $(y_\alpha^{k,j-1})_{k\in\mathbb{N}}$, and, thus, also convergent. In consequence, $(y_\alpha^{k,j})_{k\in\mathbb{N}}$ converge for each $\alpha \in \{1,\dots,j\}$ as required. Finally, one observes that $(y^{k,n})_{k\in\mathbb{N}}$ is a subsequence of $(z^k)_{k\in\mathbb{N}}$ such that all coordinate sequences $(y_\alpha^{k,n})_{k\in\mathbb{N}}$, $\alpha \in \{1,\dots,n\}$, converge. Let $a_\alpha := \lim_{k\to\infty} y_\alpha^{k,n}$ for each $\alpha \in \{1,\dots,n\}$. Then, by Th. 1.12, $\lim_{k\to\infty} y^{k,n} = a$, thereby establishing the case. ■

1.2 Metrics and Norms

Definition 1.17. Let X be a set. A function $d : X \times X \longrightarrow \mathbb{R}_0^+$ is called a *metric* on X if, and only if, the following three conditions are satisfied:

(i) d is *positive definite*, i.e., for each $(x,y) \in X \times X$, $d(x,y) = 0$ if, and only if, $x = y$.

(ii) d is *symmetric*, i.e., for each $(x,y) \in X \times X$, $d(y,x) = d(x,y)$.

(iii) d satisfies the *triangle inequality*, i.e., for each $(x,y,z) \in X^3$, $d(x,z) \leq d(x,y) + d(y,z)$.

If d constitutes a metric on X, then the pair (X,d) is called a *metric space*. One then often refers to the elements of X as *points* and to the number $d(x,y)$ as the *d-distance* between the points x and y. If the metric d on X is understood, one also refers to X itself as a metric space.

Remark 1.18. The requirement that a metric be nonnegative is included in Def. 1.17 merely for emphasis. Nonnegativity actually follows from the remaining properties of a metric: For each $x, y \in X$, one computes

$$0 \overset{\text{Def. 1.17(i)}}{=} d(x,x) \overset{\text{Def. 1.17(iii)}}{\leq} d(x,y) + d(y,x) \overset{\text{Def. 1.17(ii)}}{=} 2\,d(x,y), \tag{1.18}$$

showing $d(x,y) \geq 0$.

Definition 1.19. Let X be a vector space over the field $\mathbb{K}$. Then a function $\|\cdot\| : X \longrightarrow \mathbb{R}_0^+$ is called a *norm* on X if, and only if, the following three conditions are satisfied:

(i) $\|\cdot\|$ is *positive definite*, i.e.

$$\Big(\|x\| = 0 \;\Leftrightarrow\; x = 0\Big) \qquad \text{for each } x \in X.$$

(ii) $\|\cdot\|$ is *homogeneous of degree* 1, i.e.

$$\|\lambda x\| = |\lambda|\|x\| \qquad \text{for each } \lambda \in \mathbb{K},\ x \in X.$$

(iii) $\|\cdot\|$ satisfies the *triangle inequality*, i.e.

$$\|x+y\| \leq \|x\| + \|y\| \qquad \text{for each } x, y \in X.$$

If $\|\cdot\|$ constitutes a norm on X, then the pair $(X, \|\cdot\|)$ is called a *normed vector space* or just *normed space*. If the norm $\|\cdot\|$ on X is understood, then one also refers to X itself as a normed space.

Lemma 1.20. *If $(X, \|\cdot\|)$ is a normed space, then the function*

$$d : X \times X \longrightarrow \mathbb{R}_0^+, \quad d(x,y) := \|x-y\|, \tag{1.19}$$

constitutes a metric on X: One also calls d the metric induced by the norm $\|\cdot\|$. Thus, the induced metric d makes X into a metric space.

Proof. Exercise. ■

Lemma 1.21. (a) *The following law holds in every metric space (X, d):*

$$|d(x,y) - d(x',y')| \leq d(x,x') + d(y,y') \quad \textit{for each } x, x', y, y' \in X. \tag{1.20a}$$

(b) *The following law holds in every normed vector space $(X, \|\cdot\|)$:*

$$\big|\|x\| - \|y\|\big| \leq \|x-y\| \quad \textit{for each } x, y \in X. \tag{1.20b}$$

This law is sometimes referred to as the inverse triangle inequality*.*

Proof. (a): First, note $d(x,y) \leq d(x,x') + d(x',y') + d(y',y)$, i.e.

$$d(x,y) - d(x',y') \leq d(x,x') + d(y',y). \tag{1.21a}$$

Second, $d(x',y') \leq d(x',x) + d(x,y) + d(y,y')$, i.e.

$$d(x',y') - d(x,y) \leq d(x',x) + d(y,y'). \tag{1.21b}$$

Taken together, (1.21a) and (1.21b) complete the proof of (1.20a).

(b): Let $d(x,y) := \|x-y\|$ be the induced metric on X. Applying (a) to d yields the estimate

$$\big|\|x\| - \|y\|\big| = |d(x,0) - d(y,0)| \leq d(x,y) + d(0,0) = \|x-y\|, \tag{1.22}$$

which establishes the case. ■

Example 1.22. (a) As noted before, the length of (1.1d) is, indeed, a norm on $\mathbb{K}^n$ (see Lem. 1.3 and Rem. 1.4), called Euclidean norm for $\mathbb{K} = \mathbb{R}$. It induces the metric (1.1e) on $\mathbb{K}^n$, called Euclidean metric for $\mathbb{K} = \mathbb{R}$. In particular, the absolute value/modulus constitutes a norm on $\mathbb{K}$ and $(z,w) \mapsto |z-w|$ is a metric on $\mathbb{K}$.

(b) Consider the set $\overline{\mathbb{R}} := \mathbb{R} \cup \{\infty\} \cup \{-\infty\}$ and

$$f : \overline{\mathbb{R}} \longrightarrow \mathbb{R}, \quad f(x) := \begin{cases} -1 & \text{for } x = -\infty, \\ \frac{x}{1+|x|} & \text{for } x \in \mathbb{R}, \\ 1 & \text{for } x = \infty. \end{cases} \tag{1.23}$$

We verify that

$$d : \overline{\mathbb{R}} \times \overline{\mathbb{R}} \longrightarrow \mathbb{R}_0^+, \quad d(x,y) := |f(x) - f(y)|, \tag{1.24}$$

defines a metric on $\overline{\mathbb{R}}$:

As, for each $x, y, z \in \overline{\mathbb{R}}$, we have $d(x,y) = |f(x) - f(y)| \leq |f(x) - f(z)| + |f(z) - f(y)| = d(x,z) + d(z,y)$, d satisfies the triangle inequality. Moreover, d is symmetric, due to $d(x,y) = |f(x) - f(y)| = d(y,x)$. Next, if $x = y$, then $d(x,y) = |f(x) - f(x)| = 0$. Conversely, if $d(x,y) = 0$, then $f(x) = f(y)$. To be able to conclude that $x = y$, we show that f is one-to-one. We can actually show that f is strictly increasing, which implies one-to-one by [Phi15a, Prop. 2.29(b)]. First, consider f on $]-\infty, 0[$, where $f(x) = x/(1-x)$. Thus, f is differentiable on this interval with $f'(x) = 1/(1-x)^2 > 0$, showing that f is strictly increasing on $]-\infty, 0[$ according to [Phi15a, Cor. 9.18(a)]. That f is strictly increasing on $]0, \infty[$ follows analogously. Finally, we have, for each $x \in]-\infty, 0[$,

$$f(-\infty) = -1 < \frac{x}{1-x} < 0 = f(0),$$

and, for each $x \in]0, \infty[$,

$$f(0) = 0 < \frac{x}{1+x} < 1 = f(\infty),$$

such that f is strictly increasing on the entire set $\overline{\mathbb{R}}$.

(c) Let $S \neq \emptyset$ be an otherwise arbitrary set. According to Ex. A.2(c), the set $\mathcal{F}(S, \mathbb{K})$ of all $\mathbb{K}$-valued functions on S is a vector space over $\mathbb{K}$ if vector addition and scalar multiplication are defined pointwise as in (A.5). Now consider the subset $B(S, \mathbb{K})$ of $\mathcal{F}(S, \mathbb{K})$, consisting of all bounded $\mathbb{K}$-valued functions on S, where we call a $\mathbb{K}$-valued function f *bounded* if, and only if, the *set* $\{|f(s)| : s \in S\} \subseteq \mathbb{R}_0^+$ is a bounded subset of $\mathbb{R}$. Define

$$\|f\|_{\sup} := \sup\{|f(s)| : s \in S\} \in \mathbb{R}_0^+ \quad \text{for each } f \in B(S, \mathbb{K}). \tag{1.25}$$

We will show that $B(S, \mathbb{K})$ constitutes a vector space over $\mathbb{K}$ and $\|\cdot\|_{\sup}$ provides a norm on $B(S, \mathbb{K})$ (i.e. $\big(B(S, \mathbb{K}), \|\cdot\|_{\sup}\big)$ is a normed vector space). To verify that $B(S, \mathbb{K})$ constitutes a vector space over $\mathbb{K}$, it suffices to show it is a subspace of the vector space $\mathcal{F}(S, \mathbb{K})$, which, according to Def. and Rem. A.3 is equivalent to showing $f, g \in B(S, \mathbb{K})$ and $\lambda \in \mathbb{K}$ imply $f + g \in B(S, \mathbb{K})$ and $\lambda f \in B(S, \mathbb{K})$.

If $f, g \in B(S, \mathbb{K})$, then

$$\underset{s \in S}{\forall} \quad |f(s) + g(s)| \leq |f(s)| + |g(s)| \leq \|f\|_{\sup} + \|g\|_{\sup} \in \mathbb{R}_0^+, \tag{1.26a}$$

showing $f+g \in B(S,\mathbb{K})$ and that $\|\cdot\|_{\sup}$ satisfies the triangle inequality

$$\underset{f,g\in B(S,\mathbb{K})}{\forall} \quad \|f+g\|_{\sup} \leq \|f\|_{\sup} + \|g\|_{\sup}. \tag{1.26b}$$

If $f \in B(S,\mathbb{K})$, $\lambda \in \mathbb{K}$, then,

$$\underset{s\in S}{\forall} \quad |\lambda\, f(s)| = |\lambda|\,|f(s)| \leq |\lambda|\,\|f\|_{\sup} \in \mathbb{R}_0^+ \tag{1.27a}$$

implies $\lambda f \in B(S,\mathbb{K})$, completing the proof that $B(S,\mathbb{K})$ is a subspace of $\mathcal{F}(S,\mathbb{K})$. Moreover,

$$\begin{aligned} \|\lambda f\|_{\sup} &= \sup\{|\lambda f(s)| : s \in S\} \\ &= \sup\{|\lambda||f(s)| : s \in S\} \\ &\overset{\text{[Phi15a, (4.9c)]}}{=} |\lambda| \sup\{|f(s)| : s \in S\} = |\lambda|\|f\|_{\sup}, \end{aligned} \tag{1.27b}$$

proving $\|\cdot\|_{\sup}$ is homogeneous of degree 1. To see that $\|\cdot\|_{\sup}$ constitutes a norm on $B(S,\mathbb{K})$, it merely remains to show positive definiteness. To this end, we notice that the zero element $f=0$ of the vector space $B(S,\mathbb{K})$ is the function $f \equiv 0$, which vanishes identically. Thus, $f=0$ if, and only if, $\|f\|_{\sup} := \sup\{|f(s)| : s \in S\} = 0$, showing $\|\cdot\|_{\sup}$ is positive definite, and completing the proof that $\|\cdot\|_{\sup}$ is a norm, making $B(S,\mathbb{K})$ into a normed vector space.

1.3 Open Sets, Closed Sets, and Related Notions

Remark 1.23. In the following, a multitude of notions will be introduced for metric spaces (X,d), for example open sets, closed sets, convergence of sequences, etc. Subsequently, we will then also use these notions in normed spaces $(X,\|\cdot\|)$, always implicitly assuming that they are meant with respect to the metric space (X,d), where d is the metric induced by the norm $\|\cdot\|$, i.e. where d is given by (1.19).

Definition 1.24. Let (X,d) be a metric space. Given $x \in X$ and $r \in \mathbb{R}^+$, define

$$B_r(x) := \{y \in X : d(x,y) < r\}, \tag{1.28a}$$

$$\overline{B}_r(x) := \{y \in X : d(x,y) \leq r\}, \tag{1.28b}$$

$$S_r(x) := \{y \in X : d(x,y) = r\}. \tag{1.28c}$$

The set $B_r(x)$ is called the *open ball* with center x and radius r, also known as the r-ball with center x. The set $\overline{B}_r(x)$ is called the *closed ball* with center x and radius r. The set $S_r(x)$ is called the *sphere* with center x and radius r. A set $U \subseteq X$ is called a *neighborhood* of x if, and only if, there is $\epsilon \in \mathbb{R}^+$ such that $B_\epsilon(x) \subseteq U$.

Definition 1.25. Let (X,d) be a metric space, $A \subseteq X$, and $x \in X$.

(a) The point x is called an *interior point* of A if, and only if, there is $\epsilon \in \mathbb{R}^+$ such that the ϵ-ball with center x is entirely contained in A, i.e. $B_\epsilon(x) \subseteq A$. Note: An interior point of A is always in A.

(b) The point x is called a *boundary point* of A if, and only if, each ϵ-ball with center x, $\epsilon \in \mathbb{R}^+$, contains at least one point from A and at least one point from A^c ($A \cap B_\epsilon(x) \neq \emptyset$ and $A^c \cap B_\epsilon(x) \neq \emptyset$), where $A^c = X \setminus A$ denotes the complement of A (cf. [Phi15a, Def. 1.22(c)]). Note: A boundary point of A is not necessarily in A.

(c) The point x is called a *cluster point* or *accumulation point* of A if, and only if, each ϵ-ball with center x, $\epsilon \in \mathbb{R}^+$, contains infinitely many points of A (cf. [Phi15a, Def. 7.33(a)]). Note: A cluster point of A is not necessarily in A.

(d) The point x is called an *isolated point* of A if, and only if, there is $\epsilon \in \mathbb{R}^+$ such that $B_\epsilon(x) \cap A = \{x\}$ (cf. [Phi15a, Def. 7.33(b)]). Note: An isolated point of A is always in A.

(e) The set of all interior points of A is called the *interior* of A. It is denoted by A° or by $\operatorname{int} A$.

(f) The set of all boundary points of A is called the *boundary* of A. It is denoted by ∂A.

(g) The set $A \cup \partial A$ is called the *closure* of A. It is denoted by $\overline{A}$ or by $\operatorname{cl} A$.

(h) A is called *open* if, and only if, every point of A is an interior point, i.e. if, and only if, $A = A^\circ$.

(i) A is called *closed* if, and only if, A^c is open.

Remark 1.26. If (X, d) is a metric space, $A \subseteq X$, then Def. 1.25(i) immediately implies that A is open if, and only if, A^c is closed: According to Def. 1.25(i), A^c is closed if, and only if, $(A^c)^c$ is open. However, $(A^c)^c = X \setminus A^c = X \setminus (X \setminus A) = A$.

Lemma 1.27. *Let (X, d) be a metric space.*

(a) *Given $x \in X$ and $r \in \mathbb{R}^+$, the open ball $B_r(x)$ is an open set and the closed ball $\overline{B}_r(x)$ is a closed set.*

(b) *The empty set $\emptyset$ and the entire space X are both open and closed. Such sets are sometimes called* clopen.

(c) *Points are always closed. More precisely, for each $x \in X$, the singleton set $\{x\}$ is closed.*

Proof. (a): Exercise.

(b): It suffices to show that $\emptyset$ and X are both open. To show that $\emptyset$ is open, we need to verify that every point we find in $\emptyset$ is an interior point. As we do not find any points in $\emptyset$, we are already done. Since, for each $x \in X$ and each $\epsilon > 0$, $B_\epsilon(x) \subseteq X$, every $x \in X$ is an interior point, showing that X is open.

(c): To see that $\{x\}$ is closed, we have to show that $X \setminus \{x\}$ is open. Let $y \in X \setminus \{x\}$. Due to the positive definiteness of d, it is $r := d(x, y) > 0$. Since $z \in B_r(y)$ implies

$d(z,y) < r$, it is $x \notin B_r(y)$, showing $B_r(y) \subseteq X \setminus \{x\}$. Thus, y is an interior point of $X \setminus \{x\}$. Since y was arbitrary, $X \setminus \{x\}$ is open. ■

Example 1.28. Let $X = \mathbb{K}$ and $d(z,w) := |z-w|$ for each $z, w \in \mathbb{K}$. Then (X,d) is a metric space.

(a) If $\mathbb{K} = \mathbb{R}$, $x \in \mathbb{R}$ and $r > 0$, then $B_r(x) =]x-r, x+r[$ (open interval with center x and length $2r$) and $\overline{B}_r(x) = [x-r, x+r]$ (closed interval with center x and length $2r$). If $\mathbb{K} = \mathbb{C}$, $z \in \mathbb{C}$ and $r > 0$, then $B_r(z) = \{w \in \mathbb{C} : |z-w| < r\}$ (open disk with center z and radius r) and $\overline{B}_r(z) := \{w \in \mathbb{C} : |z-w| \leq r\}$ (closed disk with center z and radius r). Thus, we see that the notation for the open ball is consistent with the one introduced in [Phi15a, Def. 7.7(a)], and the notation for the closed ball is consistent with the one introduced in [Phi15a, Ex. 7.47(a)]. For subsets A of $\mathbb{K}$, the new notion of A being closed (i.e. A being the complement of an open set) is also consistent with the notion of closedness of [Phi15a, Def. 7.42(b)]: For the proof, we will have to wait until Cor. 1.44, where it is contained in the equivalence between statements (i) and (iv) of Cor. 1.44.

(b) Let $A :=]0,1]$ and $\mathbb{K} = \mathbb{R}$. Then $A^\circ =]0,1[$, $\partial A = \{0,1\}$, $\overline{A} = [0,1]$.

(c) Let $A :=]0,1]$ and $\mathbb{K} = \mathbb{C}$. Then $A^\circ = \emptyset$, $\partial A = \overline{A} = [0,1]$.

(d) Let $A := \mathbb{Q}$. In this case, there is no difference between $\mathbb{K} = \mathbb{R}$ and $\mathbb{K} = \mathbb{C}$: $A^\circ = \emptyset$, $\partial A = \overline{A} = \mathbb{R}$.

(e) Let $A := \{1/n : n \in \mathbb{N}\}$. Once again, there is no difference between $\mathbb{K} = \mathbb{R}$ and $\mathbb{K} = \mathbb{C}$: Every element of A is an isolated point. In particular $A^\circ = \emptyset$. The unique cluster point of A is 0, and $\partial A = \overline{A} = A \cup \{0\}$.

Theorem 1.29. *Let (X,d) be a metric space.*

(a) *Unions of arbitrarily many (i.e. finitely or infinitely many) open sets are open. The intersection of finitely many open sets is open.*

(b) *Intersections of arbitrarily many closed sets are closed (cf.* [Phi15a, Prop. 7.44(b)]*). The union of finitely many closed sets is closed (cf.* [Phi15a, Prop. 7.44(a)]*).*

Proof. (a): Let I be a (finite or infinite) index set. For each $j \in I$, let $O_j \subseteq X$ be open. We have to verify that $O := \bigcup_{j \in I} O_j$ is open. Let $x \in O$. Then there is $j \in I$ such that $x \in O_j$. Since O_j is open, there is $\epsilon > 0$ such that $B_\epsilon(x) \subseteq O_j \subseteq O$. Thus, we have shown that x is an interior point of O. Since x was arbitrary, O is open. Now consider finitely many open sets $O_1, \ldots, O_N$, $N \in \mathbb{N}$, and let $O := \bigcap_{j=1}^N O_j$. Again, we have to prove that O is open. Hence, once more, let $x \in O$. Then $x \in O_j$ for each $j \in \{1, \ldots, N\}$. Since each O_j is open, for each $j \in \{1, \ldots, N\}$, there is $\epsilon_j > 0$ such that $B_{\epsilon_j}(x) \subseteq O_j$. If we let $\epsilon := \min\{\epsilon_j : j \in \{1, \ldots, N\}\}$, then $\epsilon > 0$ and $B_\epsilon(x) \subseteq B_{\epsilon_j}(x) \subseteq O_j$ for each $j \in \{1, \ldots, N\}$, i.e. $B_\epsilon(x) \subseteq O$, showing that x is an interior point of O. Since x was arbitrary, O is open.

(b): Let $I \neq \emptyset$ be a (finite or infinite) index set. For each $j \in I$, let $C_j \subseteq X$ be closed. We have to verify that $C := \bigcap_{j\in I} C_j$ is closed. According to the set-theoretic law [Phi15a, Prop. 1.38(e)]

$$C^{\mathrm{c}} = \left(\bigcap_{j\in I} C_j\right)^{\mathrm{c}} \overset{\text{[Phi15a, Prop. 1.38(e)]}}{=} \bigcup_{j\in I} C_j^{\mathrm{c}}.$$

Now, as we know that C_j is closed, we know that C_j^{c} is open. According to (a), that means that C^{c} is open, showing that C is closed. Similarly, if we consider finitely many closed sets $C_1, \ldots, C_N$, $N \in \mathbb{N}$, and letting $C := \bigcup_{j=1}^{N} C_j$, then the set-theoretic law [Phi15a, Prop. 1.38(f)] yields

$$C^{\mathrm{c}} = \left(\bigcup_{j=1}^{N} C_j\right)^{\mathrm{c}} \overset{\text{[Phi15a, Prop. 1.38(f)]}}{=} \bigcap_{j=1}^{N} C_j^{\mathrm{c}}.$$

Since C_j is closed, C_j^{c} is open, and, by (a), C^{c} is open, hence C closed. ■

Example 1.30. Consider $\mathbb{R}$ with its usual metric (as in Ex. 1.28 for $\mathbb{K} = \mathbb{R}$). Then the relation $\bigcap_{k=1}^{\infty}]-\frac{1}{k}, \frac{1}{k}[= \{0\}$ shows that, in general, an infinite intersection of open sets is not open, and $\bigcup_{k=1}^{\infty} [\frac{1}{k}, 1] =]0, 1]$, shows that, in general, an infinite union of closed sets is not closed.

Lemma 1.31. *Let (X, d) be a metric space, $A \subseteq X$. Then X is the disjoint union of A°, ∂A, and $(X \setminus A)^\circ$.*

Proof. One has to show four parts: $X = A^\circ \cup \partial A \cup (X \setminus A)^\circ$, $A^\circ \cap \partial A = \emptyset$, $\partial A \cap (X \setminus A)^\circ = \emptyset$, and $A^\circ \cap (X \setminus A)^\circ = \emptyset$.

Suppose $x \in X \setminus (A^\circ \cup \partial A)$. Since $x \notin \partial A$, there exists $\epsilon > 0$ such that $B_\epsilon(x) \subseteq A$ or $B_\epsilon(x) \subseteq X \setminus A$. As $x \notin A^\circ$, it must be $B_\epsilon(x) \subseteq X \setminus A$, i.e. $x \in (X \setminus A)^\circ$.

$A^\circ \cap \partial A = \emptyset$: If $x \in A^\circ$, then there is $\epsilon > 0$ such that $B_\epsilon(x) \subseteq A$, thus, $x \notin \partial A$.

$\partial A \cap (X \setminus A)^\circ = \emptyset$: Since $\partial A = \partial(X \setminus A)$, this follows from $A^\circ \cap \partial A = \emptyset$.

$A^\circ \cap (X \setminus A)^\circ = \emptyset$ holds as $A^\circ \subseteq A$, $(X \setminus A)^\circ \subseteq X \setminus A$, and $A \cap (X \setminus A) = \emptyset$. ■

Theorem 1.32. *Let (X, d) be a metric space, $A \subseteq X$.*

(a) *The boundary ∂A is closed.*

(b) *The interior A° is the union of all open subsets of A. In particular, A° is open. In other words, A° is the largest open set contained in A.*

(c) *The closure $\overline{A}$ is the intersection of all closed supersets of A. In particular, $\overline{A}$ is closed. In other words, $\overline{A}$ is the smallest closed set containing A.*

Proof. (a): According to Lem. 1.31, it is $\partial A = X \setminus \big(A^\circ \cup (X \setminus A)^\circ\big)$. Since A° and $(X \setminus A)^\circ$ are open, ∂A is closed.

(b): Let O be the union of all open subsets of A. Then O is open by Th. 1.29(a). If $x \in A^\circ$, then x is an interior point of A, i.e. there is $\epsilon > 0$ such that $B_\epsilon(x) \subseteq A$. Since $B_\epsilon(x)$ is open due to Lem. 1.27(a), $x \in O$. Conversely, if $x \in O$, then, as O is open, there is $\epsilon > 0$ such that $B_\epsilon(x) \subseteq O \subseteq A$, showing that x is an interior point of A, i.e. $x \in A^\circ$.

(c): According to (b), $(A^\mathrm{c})^\circ$ is the union of all open subsets of A^c, i.e.

$$(A^\mathrm{c})^\circ = \bigcup_{O \in \{S \subseteq A^\mathrm{c} \colon S \text{ open}\}} O, \tag{1.29}$$

then

$$\big((A^\mathrm{c})^\circ\big)^\mathrm{c} \overset{\text{[Phi15a, Prop. 1.38(f)]}}{=} \bigcap_{O \in \{S \subseteq A^\mathrm{c} \colon S \text{ open}\}} O^\mathrm{c} = \bigcap_{C \in \{S \supseteq A \colon S \text{ closed}\}} C \tag{1.30}$$

is the intersection of all closed supersets of A (note that C is a closed superset of A if, and only if, C^c is an open subset of A^c). As, by Lem. 1.31,

$$\big((A^\mathrm{c})^\circ\big)^\mathrm{c} = \partial(A^\mathrm{c}) \cup A^\circ = \partial A \cup A^\circ = \partial A \cup A = \overline{A}, \tag{1.31}$$

$\overline{A}$ is the intersection of all closed supersets of A as claimed. ■

Definition 1.33. Let (X, d) be a metric space. Then $A \subseteq X$ is called *bounded* if, and only if, $A = \emptyset$ or $A \neq \emptyset$ and the set $\{d(x, y) : x, y \in A\}$ is bounded in $\mathbb{R}$; $A \subseteq X$ is called *unbounded* if, and only if, A is not bounded. For each $A \subseteq X$, the number

$$\operatorname{diam} A := \begin{cases} 0 & \text{for } A = \emptyset, \\ \sup\big\{d(x, y) : x, y \in A\big\} & \text{for } \emptyset \neq A \text{ bounded}, \\ \infty & \text{for } A \text{ unbounded}, \end{cases} \tag{1.32}$$

is called the *diameter* of A. Thus, $\operatorname{diam} A \in [0, \infty] := \mathbb{R}_0^+ \cup \{\infty\}$ and A is bounded if, and only if, $\operatorname{diam} A < \infty$.

Lemma 1.34. *If (X, d) is a metric space, then $A \subseteq X$ is bounded if, and only if, there is $r > 0$ and $x \in X$ such that $A \subseteq B_r(x)$ (in particular, Def.* 1.33 *is consistent with* [Phi15a, Def. 7.42(a)]*).*

Proof. If A is bounded, then $\operatorname{diam} A < \infty$. Let r be any real number bigger than $\operatorname{diam} A$, e.g. $1 + \operatorname{diam} A$. Choose any point $x \in A$. Then, by the definition of $\operatorname{diam} A$, for each $y \in A$, it is $d(x, y) \leq \operatorname{diam} A < r$, showing that $A \subseteq B_r(x)$. Conversely, if $r > 0$ and $x \in X$ such that $A \subseteq B_r(x)$, then, by the definition of $B_r(x)$, one has $d(x, y) < r$ for each $y \in A$. Now, if $y, z \in A$, then $d(z, y) \leq d(z, x) + d(x, y) < 2r$, showing $\operatorname{diam} A \leq 2r < \infty$, i.e. A is bounded. ■

Lemma 1.35. *Let (X, d) be a metric space.*

(a) *Every finite subset of X is bounded.*

(b) *The union of two bounded subsets of X is bounded.*

Proof. (a): Let A be a finite subset of X and $a \in A$. Set $r := 1 + \max\{d(a,x) : x \in A\}$. Then $1 \leq r < \infty$, since A is finite. Moreover, $A \subseteq B_r(a)$, showing that A is bounded.

(b): Let A and B be bounded subsets of X. Then there are $x, y \in X$ and $r > 0$ such that $A \subseteq B_r(x)$ and $B \subseteq B_r(y)$. Define $\alpha := d(x,y)$ and $\epsilon := r + \alpha$. Then $A \subseteq B_r(x) \subseteq B_\epsilon(x)$. If $b \in B$, then $d(b,x) \leq d(b,y) + d(y,x) < r + \alpha = \epsilon$, showing $B \subseteq B_\epsilon(x)$, and, thus, $A \cup B \subseteq B_\epsilon(x)$, establishing that $A \cup B$ is bounded. ■

1.4 Convergence

Definition 1.36. Let (X,d) be a metric space and let $(x^k)_{k\in\mathbb{N}}$ be a sequence in X.

(a) The sequence $(x^k)_{k\in\mathbb{N}}$ is called *bounded* if, and only if, the *set* $\{x^k : k \in \mathbb{N}\}$ is bounded in the sense of Def. 1.33.

(b) The sequence $(x^k)_{k\in\mathbb{N}}$ is said to be *convergent with limit $y \in X$* if, and only if, the real sequence of distances $\big(d(x^k,y)\big)_{k\in\mathbb{N}}$ converges to 0. As for sequences in $\mathbb{K}$ and $\mathbb{K}^n$, the notation for $(x^k)_{k\in\mathbb{N}}$ converging to y is $\lim_{k\to\infty} x^k = y$ or $x^k \to y$ for $k \to \infty$. Thus, by definition,

$$\lim_{k\to\infty} x^k = y \quad \Leftrightarrow \quad \lim_{k\to\infty} d(x^k,y) = 0. \tag{1.33}$$

(c) The sequence $(x^k)_{k\in\mathbb{N}}$ is called *divergent* if, and only if, it is not convergent.

(d) The sequence $(x^k)_{k\in\mathbb{N}}$ is said to be a Cauchy sequence if, and only if, for each $\epsilon > 0$, there is $N \in \mathbb{N}$ such that, $d(x^k, x^l) < \epsilon$ for each $k, l > N$.

(e) A point $y \in X$ is called a *cluster point* or an *accumulation point* of the sequence $(x^k)_{k\in\mathbb{N}}$ if, and only if, for each $\epsilon > 0$, $B_\epsilon(y)$ contains infinitely many members of the sequence (i.e. the cardinality of the set $\{k \in \mathbb{N} : x^k \in B_\epsilon(y)\}$ is ∞).

Lemma 1.37. *For a sequence $(x^k)_{k\in\mathbb{N}}$ in a metric space (X,d), the following two statements are equivalent:*

(i) *$(x^k)_{k\in\mathbb{N}}$ is convergent with limit $y \in X$.*

(ii) *For each $\epsilon > 0$, there is $N \in \mathbb{N}$ such that, for each $k > N$, $x^k \in B_\epsilon(y)$.*

In consequence, the analogous result also holds for a sequence in a normed vector space.

Proof. (i) is equivalent to $\lim_{k\to\infty} d(x^k,y) = 0$, which is equivalent to the statement that, for each $\epsilon > 0$, there is $N \in \mathbb{N}$ such that, for each $k > N$, $d(x^k,y) < \epsilon$, which is equivalent to (ii). ■

The following Prop. 1.38 shows that many of the properties we learned in Calculus I for sequences in $\mathbb{K}$ still hold in general metric spaces.

Proposition 1.38. *Let (X, d) be a metric space and let $(x^k)_{k\in\mathbb{N}}$ be a sequence in X.*

(a) *Limits are unique, that means if $a, b \in X$ such that $\lim_{k\to\infty} x^k = a$ and $\lim_{k\to\infty} x^k = b$, then $a = b$.*

(b) *If $(x^k)_{k\in\mathbb{N}}$ is convergent, then it is bounded.*

(c) *If $\lim_{k\to\infty} x^k = a \in X$, then every subsequence and every reordering of $(x^k)_{k\in\mathbb{N}}$ is also convergent with limit a.*

(d) *A point $y \in X$ is a cluster point of $(x^k)_{k\in\mathbb{N}}$ if, and only if, the sequence has a subsequence converging to y.*

(e) *If $(x^k)_{k\in\mathbb{N}}$ is convergent, then it is a Cauchy sequence.*

Proof. All the following proofs are conducted analogous to the respective proofs for sequences in $\mathbb{K}$.

(a): To carry out the proof via contraposition, show that $a \neq b$ and $\lim_{k\to\infty} x^k = a$ imply that b is not a limit of $(x^k)_{k\in\mathbb{N}}$. If $a \neq b$, then $d(a, b) > 0$. Let $\epsilon := d(a, b)/2$. If $x \in B_\epsilon(a)$, then $d(x, a) < \epsilon$. As $d(a, b) \leq d(a, x) + d(x, b)$, one gets

$$d(x, b) \geq d(a, b) - d(a, x) > 2\epsilon - \epsilon = \epsilon,$$

showing $x \notin B_\epsilon(b)$. Since there is $N \in \mathbb{N}$ such that $x^k \in B_\epsilon(a)$ for $k > N$, no such x^k can be in $B_\epsilon(b)$, i.e. b is not a limit of $(x^k)_{k\in\mathbb{N}}$.

(b) (cf. the proof for sequences in $\mathbb{K}$ in [Phi15a, Prop. 7.10(b)]): If $\lim_{k\to\infty} x^k = a$, then there is $N \in \mathbb{N}$ such that $x^k \in B_1(a)$ for each $k > N$, i.e. $\{x^k : k > N\}$ is bounded. Moreover, the finite set $\{x^k : k \leq N\}$ is bounded. Therefore, $\{x^k : k \in \mathbb{N}\}$ is the union of two bounded sets, and, hence, bounded.

(c) (cf. the proofs for sequences in $\mathbb{K}$ in [Phi15a, Prop. 7.23]): Let $(y^k)_{k\in\mathbb{N}}$ be a subsequence of $(x^k)_{k\in\mathbb{N}}$, i.e. there is a strictly increasing function $\phi : \mathbb{N} \longrightarrow \mathbb{N}$ such that $y^k = x^{\phi(k)}$. If $\lim_{k\to\infty} x^k = a$, then, given $\epsilon > 0$, there is $N \in \mathbb{N}$ such that $x^k \in B_\epsilon(a)$ for each $k > N$. For $\tilde{N}$ choose any number from $\mathbb{N}$ that is bigger than or equal to N and in the range of ϕ. Take $M := \phi^{-1}(\tilde{N})$. Then, for each $k > M$, one has $\phi(k) > \tilde{N} \geq N$, and, thus, $y^k = x^{\phi(k)} \in B_\epsilon(a)$, showing $\lim_{k\to\infty} y^k = a$.

Let $(y^k)_{k\in\mathbb{N}}$ be a reordering of $(x^k)_{k\in\mathbb{N}}$, i.e. there is a bijective function $\phi : \mathbb{N} \longrightarrow \mathbb{N}$ such that $y^k = x^{\phi(k)}$. Let ϵ and N be as before. Define

$$M := \max\{\phi^{-1}(k) : k \leq N\}.$$

As ϕ is bijective, it is $\phi(k) > N$ for each $k > M$. Then, for each $k > M$, one has $y^k = x^{\phi(k)} \in B_\epsilon(a)$, showing $\lim_{k\to\infty} y^k = a$.

(d) (cf. the proof for sequences in $\mathbb{K}$ in [Phi15a, Prop. 7.26]): If $(y^k)_{k\in\mathbb{N}}$ is a subsequence of $(x^k)_{k\in\mathbb{N}}$, $\lim_{k\to\infty} y^k = a$, then every $B_\epsilon(a)$, $\epsilon > 0$, contains infinitely many y^k, i.e. infinitely many x^k, i.e. a is a cluster point of $(x^k)_{k\in\mathbb{N}}$. Conversely, if a is a cluster point of $(x^k)_{k\in\mathbb{N}}$, then, inductively, define $\phi : \mathbb{N} \longrightarrow \mathbb{N}$ as follows: For $\phi(1)$, choose any point x^k in $B_1(a)$ (such a point exists, since a is a cluster point of the sequence). Now assume that $n > 1$ and that $\phi(l)$ have already been defined for each $l < n$. Let $M := \max\{\phi(l) : l < n\}$. Since $B_{\frac{1}{n}}(a)$ contains infinitely many x^k, there must be some $x^k \in B_{\frac{1}{n}}(a)$ such that $k > M$. Choose this k as $\phi(n)$. Thus, by construction, ϕ is strictly increasing, i.e. $(y^k)_{k\in\mathbb{N}}$ with $y^k := x^{\phi(k)}$ is a subsequence of $(x^k)_{k\in\mathbb{N}}$. Moreover, for each $\epsilon > 0$, there is $N \in \mathbb{N}$ such that $1/N < \epsilon$. Then, for each $k > N$, $y^k \in B_{\frac{1}{k}}(a) \subseteq B_{\frac{1}{N}}(a) \subseteq B_\epsilon(a)$, showing $\lim_{k\to\infty} y^k = a$

(e) (cf. the proof for sequences in $\mathbb{K}$ in [Phi15a, Th. 7.29]): If $\lim_{k\to\infty} x^k = a$, then, given $\epsilon > 0$, there is $N \in \mathbb{N}$ such that $x^k \in B_{\frac{\epsilon}{2}}(a)$ for each $k > N$. If $k, l > N$, then $d(x^k, x^l) \le d(x^k, a) + d(a, x^l) < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon$, establishing that $(x^k)_{k\to\infty}$ is a Cauchy sequence. ■

Example 1.39. (a) The sequence

$$\big((0,1),(1,1/2),(0,1/3),(1,1/4),\dots\big)$$

is not a Cauchy sequence. In particular, it does not converge. It has precisely two cluster points, namely $(0,0)$ and $(1,0)$. Moreover, $\big((0,1/(2k-1))\big)_{k\in\mathbb{N}}$ is a subsequence converging to $(0,0)$ and $\big((1,1/(2k))\big)_{k\in\mathbb{N}}$ is a subsequence converging to $(1,0)$.

(b) Let X be the vector space over $\mathbb{K}$ of sequences in $\mathbb{K}$ that are finally constant and equal to 0. Thus, the sequence $z = (z_n)_{n\in\mathbb{N}}$, $z_n \in \mathbb{K}$ for each $n \in \mathbb{N}$, is in X if, and only if, there exists $N \in \mathbb{N}$ such that $z_n = 0$ for each $n \ge N$. Clearly, X endowed with the norm $\|\cdot\|_{\sup}$ is a subspace of the normed vector space $B(S,\mathbb{K})$ of Example 1.22(c) with $S := \mathbb{N}$. Defining, for each $n, k \in \mathbb{N}$,

$$z_n^k := \begin{cases} 1/n & \text{for } 1 \le n \le k, \\ 0 & \text{for } n > k, \end{cases} \tag{1.34}$$

one sees that $(z^k)_{k\in\mathbb{N}}$ is a Cauchy sequence in X (i.e. with respect to $\|\cdot\|_{\sup}$), but it is not convergent in X (its limit, the sequence $(1/n)_{n\in\mathbb{N}}$ is not finally constant and, thus, not in X).

Lemma 1.40. (a) *In each metric space (X,d), the metric d is continuous in the following sense: If $(x^k)_{k\in\mathbb{N}}$ and $(y^k)_{k\in\mathbb{N}}$ are convergent sequences in X, $\lim_{k\to\infty} x^k = x$, $\lim_{k\to\infty} y^k = y$, then $\lim_{k\to\infty} d(x^k, y^k) = d(x,y)$.*

(b) *In each normed vector space $(X, \|\cdot\|)$, the norm is continuous is the sense that $\lim_{k\to\infty} x^k = x$ implies $\lim_{k\to\infty} \|x^k\| = \|x\|$ for each convergent sequence $(x^k)_{k\in\mathbb{N}}$ in X.*

Proof. (a): Given $\epsilon > 0$, there is $N \in \mathbb{N}$ such that, for each $k > N$, $x^k \in B_{\frac{\epsilon}{2}}(x)$ and $y^k \in B_{\frac{\epsilon}{2}}(y)$. Then, for each $k > N$, according to Lem. 1.21(a), $|d(x,y) - d(x^k,y^k)| \leq d(x,x^k) + d(y,y^k) < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon$, thereby establishing the case.

(b): Applying (a) to the induced metric d yields $\lim_{k\to\infty} \|x^k\| = \lim_{k\to\infty} d(x^k, 0) = d(x,0) = \|x\|$. ■

For $X = \mathbb{K}^n$ with d being the metric given by (1.1e), we know the converse of Lem. 1.38(e) is also true, namely every Cauchy sequence in $\mathbb{K}^n$ converges. However, this is not true for all metric spaces, as simple examples show. Take, e.g., $X = \mathbb{Q}$ or $X =]0,1]$ with d being given by the absolute value. A less trivial example is the sequence space X of Example 1.39(b). This gives rise to the following definition.

Definition 1.41. A metric space (X,d) and its metric d are both called *complete* if, and only if, every Cauchy sequence in X converges. A normed space is called a *Banach space* if, and only if, the metric induced by the norm is complete. In that case, one also says that the normed space and the norm itself are complete.

—

We will now proceed to study some relations between cluster points of a set A, the closure of a set A, and convergent sequences in A.

Lemma 1.42. *Let (X,d) be a metric space, $A \subseteq X$. Then $x \in X$ is a cluster point of A if, and only if, there is a sequence $(a^k)_{k\in\mathbb{N}}$ in A such that $\lim_{k\to\infty} a^k = x$, however $a^k \neq x$, for each $k \in \mathbb{N}$.*

Proof. If $(a^k)_{k\to\infty}$ is a sequence in A such that $\lim_{k\to\infty} a^k = x$ and $a^k \neq x$ for each $k \in \mathbb{N}$, then define a subsequence $(b^k)_{k\to\infty}$ of $(a^k)_{k\to\infty}$ with the additional property that $b^k \neq b^l$ for each $k \neq l$. To that end, inductively, define a strictly increasing function $\phi : \mathbb{N} \longrightarrow \mathbb{N}$ as follows: Let $\phi(1) := 1$. For $N > 1$ assume that $\phi(k)$ for each $1 \leq k < N$ has already been defined such that $\phi(1) < \ldots \phi(N-1)$ and such that $a^{\phi(k)} \neq a^{\phi(l)}$ for each $1 \leq k, l \leq N-1$. Set

$$\epsilon := \min\big\{d(a^{\phi(k)}, x) : 1 \leq k \leq N-1\big\}.$$

Then $\epsilon > 0$ since all $a^k \neq x$. For $\phi(N)$ choose some natural number $M > \phi(N-1)$ such that $a^M \in B_\epsilon(x)$ (which exists as $\lim_{k\to\infty} a^k = x$). Due to the choice of ϵ, one has $a^{\phi(k)} \neq a^{\phi(N)}$ for each $1 \leq k \leq N-1$. Now, if one lets $b^k := a^{\phi(k)}$, then $(b^k)_{k\to\infty}$ is a subsequence of $(a^k)_{k\to\infty}$ (since ϕ is increasing) with the additional property that $b^k \neq b^l$ for each $k \neq l$. As $(b^k)_{k\to\infty}$ is a subsequence of $(a^k)_{k\to\infty}$, one has $\lim_{k\to\infty} b^k = \lim_{k\to\infty} a^k = x$ by Prop. 1.38(c). Finally, given $\epsilon > 0$, there is $N \in \mathbb{N}$ such that, for each $k > N$, $b^k \in B_\epsilon(x)$. Since the b^k are now all distinct, this constitutes an infinite number of elements from A, i.e. x is a cluster point of A.

Conversely, if x is a cluster point of A, for each $k \in \mathbb{N}$, choose $a^k \in (A \cap B_{\frac{1}{k}}(x)) \setminus \{x\}$ (such an element exists as x is a cluster point of A). If $\epsilon > 0$, then there is $N \in \mathbb{N}$, such that $1/N \leq \epsilon$. If $k > N$, then $a^k \in B_{\frac{1}{k}}(x) \subseteq B_\epsilon(x)$, showing that $\lim_{k\to\infty} a^k = x$. ■

Theorem 1.43. *Let (X,d) be a metric space, $A \subseteq X$. Let $H(A)$ denote the set of cluster points of A, and let $L(A)$ denote the set of limits of sequences in A, i.e. $L(A)$ consists of all $x \in X$ such that there is a sequence $(x^k)_{k\in\mathbb{N}}$ in A satisfying $\lim_{k\to\infty} x^k = x$.*

It then holds that $\overline{A} = L(A) = A \cup H(A)$.

Proof. It suffices to show that $L(A) \subseteq \overline{A} \subseteq A \cup H(A) \subseteq L(A)$.

"$L(A) \subseteq \overline{A}$": Suppose $x \notin \overline{A}$. Since $X \setminus \overline{A}$ is open, there is $\epsilon > 0$ such that $B_\epsilon(x) \subseteq X \setminus \overline{A} \subseteq X \setminus A$. Thus, $x \notin L(A)$.

"$\overline{A} \subseteq A \cup H(A)$": Let $x \in \overline{A} \setminus A$. We need to show that $x \in H(A)$. As $\overline{A} = A \cup \partial A$ and $x \notin A$, we have $x \in \partial A$. For each $k \in \mathbb{N}$, $x^k \in A$ is constructed inductively as follows: For x^1 choose any element of $B_1(x) \cap A$. Now let $n > 1$ and assume that, for each $1 \leq l < n$, x^l has already been constructed such that $x^l \in B_{1/l}(x) \cap A$ and, for $k \neq l$, $x^k \neq x^l$. Define

$$\delta := \min\left(\left\{d(x,x^l) : 1 \leq l < n\right\} \cup \left\{\frac{1}{n}\right\}\right)$$

and choose $x^n \in B_\delta(x) \cap A$. Then $x^n \in B_{1/n}(x) \cap A$ and $x^n \neq x^l$ for each $1 \leq l < n$. Now, for each $\epsilon > 0$, there is $N \in \mathbb{N}$ such that $1/N < \epsilon$. For each $k > N$, one has $x^k \in B_{1/k}(x) \subseteq B_{1/N}(x) \subseteq B_\epsilon(x)$, i.e. $B_\epsilon(x)$ contains infinitely many different $x^k \in A$ (note also that $\lim_{k\to\infty} x^k = x$), showing that x is a cluster point of A.

"$A \cup H(A) \subseteq L(A)$": If $a \in A$, then the constant sequence $(a, a, \dots)$ converges to a, implying $a \in L(A)$. If $a \in H(A)$, then $a \in L(A)$ according to Lem. 1.42. ■

Corollary 1.44. *Let (X,d) be a metric space, $A \subseteq X$. Then the following statements are equivalent:*

(i) *A is closed.*

(ii) *$A = \overline{A}$.*

(iii) *A contains all cluster points of A.*

(iv) *A contains all limits of sequences in A that are convergent in X (cf.* [Phi15a, Def. 7.42(b)]*).*

In particular, if A does not have any cluster points, then A is closed.

Proof. The equivalence of (i) and (ii) is due to Th. 1.32(c) ($\overline{A}$ is the smallest closed set containing A). The equivalences of (ii), (iii), and (iv) are due to Th. 1.43: Using the notation $L(A)$ and $H(A)$ from Th. 1.43, one has that $A = \overline{A}$ implies $A = A \cup H(A)$, i.e. $H(A) \subseteq A$, i.e. (ii) implies (iii). If $H(A) \subseteq A$, then $L(A) = A \cup H(A) = A$, i.e. (iii) implies (iv). If $A = L(A)$, then $A = \overline{A}$, i.e. (iv) implies (ii). ■

Example 1.45. Let $p, q \in \mathbb{N}$ and consider the metric spaces given by $\mathbb{K}^p$, $\mathbb{K}^q$, $\mathbb{K}^{p+q}$, each endowed with the metric given by (1.1e) (i.e. the Euclidean distance for $\mathbb{K} = \mathbb{R}$). Let $A \subseteq \mathbb{K}^p$, $B \subseteq \mathbb{K}^q$.

(a) If A and B are closed, then $A \times B$ is closed in $\mathbb{K}^{p+q} = \mathbb{K}^p \times \mathbb{K}^q$: Let $(c^k)_{k\in\mathbb{N}}$ be a convergent sequence in $A \times B$ with $\lim_{k\to\infty} c^k = c \in \mathbb{K}^{p+q}$. Then, for each $k \in \mathbb{N}$, $c^k = (a^k, b^k)$ with $a^k \in \mathbb{K}^p$, $b^k \in \mathbb{K}^q$. Moreover, $c = (a, b)$ with $a \in \mathbb{K}^p$ and $b \in \mathbb{K}^q$. According to Th. 1.12, one has $a = \lim_{k\to\infty} a^k$ and $b = \lim_{k\to\infty} b^k$. Since A and B are closed, from Cor. 1.44(iv), we know that $a \in A$ and $b \in B$, i.e. $c = (a, b) \in A\times B$, showing that $A \times B$ is closed.

(b) If A and B are open, then $A \times B$ is open in $\mathbb{K}^{p+q} = \mathbb{K}^p \times \mathbb{K}^q$: It suffices to show that $(A \times B)^{\mathrm{c}} = \mathbb{K}^{p+q} \setminus (A \times B)$ is closed. To that end, note

$$(A \times B)^{\mathrm{c}} = (A^{\mathrm{c}} \times \mathbb{K}^q) \cup (\mathbb{K}^p \times B^{\mathrm{c}}) : \tag{1.35}$$

For a point $(z, w) \in \mathbb{K}^p \times \mathbb{K}^q = \mathbb{K}^{p+q}$, one reasons as follows:

$$\begin{aligned}
(z, w) \in (A \times B)^{\mathrm{c}} \quad &\Leftrightarrow \quad (z, w) \notin A \times B \\
&\Leftrightarrow \quad \big(z \notin A \text{ and } w \in \mathbb{K}^q\big) \text{ or } \big(z \in \mathbb{K}^p \text{ and } w \notin B\big) \\
&\Leftrightarrow \quad (z, w) \in A^{\mathrm{c}} \times \mathbb{K}^q \text{ or } (z, w) \in \mathbb{K}^p \times B^{\mathrm{c}} \\
&\Leftrightarrow \quad (z, w) \in (A^{\mathrm{c}} \times \mathbb{K}^q) \cup (\mathbb{K}^p \times B^{\mathrm{c}}),
\end{aligned} \tag{1.36}$$

thereby proving (1.35). One now observes that A^{c} and B^{c} are closed, as A and B are open. As $\mathbb{K}^p$ and $\mathbb{K}^q$ are also closed, by (a), $A^{\mathrm{c}} \times \mathbb{K}^q$ and $\mathbb{K}^p \times B^{\mathrm{c}}$ are closed, and, thus, by (1.35), so is $(A \times B)^{\mathrm{c}}$. In consequence, $A \times B$ is open as claimed.

In particular, open intervals in $\mathbb{R}^n$ are open and closed intervals in $\mathbb{R}^n$ are closed.

1.5 Limits and Continuity of Functions

Definition 1.46. Let (X, d_X) and (Y, d_Y) be metric spaces, $M \subseteq X$. If $\xi \in X$ is a cluster point of M, then a function $f : M \longrightarrow Y$ is said to tend to $\eta \in Y$ (or to have the *limit* $\eta \in Y$) for $x \to \xi$ (denoted by $\lim_{x\to\xi} f(x) = \eta$) if, and only if, for each $\epsilon > 0$, there is $\delta > 0$ such that

$$d_Y\big(f(x), \eta\big) < \epsilon \quad \text{for each } \xi \neq x \in M \cap B_\delta(\xi). \tag{1.37}$$

Remark 1.47. The reason that $x = \xi$ is excluded in (1.37) is that one wants to allow the situation $f(\xi) \neq \lim_{x\to\xi} f(x)$, i.e. the value of a function in ξ is allowed to differ from the functions limit for $x \to \xi$. Thus, for a cluster point ξ of M with $\xi \in M$, one of three distinct cases will always occur: (i) $\lim_{x\to\xi} f(x)$ does not exist, (ii) $f(\xi) \neq \lim_{x\to\xi} f(x)$, (iii) $f(\xi) = \lim_{x\to\xi} f(x)$.

—

In the following Definitions 1.48 and 1.49, we will generalize the notions of continuity [Phi15a, Def. 7.31], uniform continuity [Phi15a, (10.38)], and Lipschitz continuity [Phi15a, Def. and Rem. 10.16] to metric spaces.

Definition 1.48. Let (X, d_X) and (Y, d_Y) be metric spaces, $M \subseteq X$. If $\xi \in M$, then a function $f : M \longrightarrow Y$ is said to be *continuous* in ξ if, and only if, for each $\epsilon > 0$, there is $\delta > 0$ such that

$$d_Y\big(f(x), f(\xi)\big) < \epsilon \quad \text{for each } x \in M \cap B_\delta(\xi). \tag{1.38}$$

Definition 1.49. Let (X, d_X) and (Y, d_Y) be metric spaces, $M \subseteq X$, $f : M \longrightarrow Y$.

(a) f is called *continuous* in M if, and only if, f is continuous in each $\xi \in M$. The set of all continuous functions from M into Y is denoted by $C(M, Y)$.

(b) f is called *uniformly continuous* in M if, and only if, for each $\epsilon > 0$, there is $\delta > 0$ such that:

$$\forall_{x,y \in M} \quad d_X(x, y) < \delta \;\Rightarrow\; d_Y\big(f(x), f(y)\big) < \epsilon. \tag{1.39}$$

The point here is that δ must not depend on x and y.

(c) f is called *Lipschitz continuous* in M with *Lipschitz constant* L if, and only if, there is $L \in \mathbb{R}_0^+$ such that:

$$\forall_{x,y \in M} \quad d_Y\big(f(x), f(y)\big) \leq L\, d_X(x, y). \tag{1.40}$$

The set of all Lipschitz continuous functions from M into Y is denoted by $\mathrm{Lip}(M, Y)$.

Remark 1.50. All the notions introduced above for metric spaces will also be used in normed vector spaces. They are then meant with respect to the metric induced by the norm.

Lemma 1.51. *Let (X, d_X) and (Y, d_Y) be metric spaces, $M \subseteq X$, $f : M \longrightarrow Y$. If f is Lipschitz continuous in M, then f is uniformly continuous in M. If f is uniformly continuous in M, then f is continuous in M.*

Proof. If f is Lipschitz continuous, then there is $L \in \mathbb{R}_0^+$ such that $d_Y\big(f(x), f(y)\big) \leq L\, d_X(x, y)$ for each $x, y \in M$. Thus, given $\epsilon > 0$, choose $\delta := \epsilon$ for $L = 0$ and $\delta := \epsilon / L$ for $L > 0$. Let $x, y \in M$ such that $d_X(x, y) < \delta$. If $L = 0$, then $d_Y\big(f(x), f(y)\big) = 0 < \epsilon$. If $L > 0$, then $d_Y\big(f(x), f(y)\big) \leq L\, d_X(x, y) < L\epsilon / L = \epsilon$, showing that f is uniformly continuous. If f is uniformly continuous, then, given $\epsilon > 0$, there is $\delta > 0$ such that, for each $x, y \in M$ with $d_X(x, y) < \delta$, it is $d_Y\big(f(x), f(y)\big) < \epsilon$. This already shows that f is continuous in x. As x is arbitrary, f is continuous. ∎

Example 1.52. Consider $X = \mathbb{R}$ with the usual metric given by the absolute value function, $M := \mathbb{R}^+$.

(a) $f : M \longrightarrow \mathbb{R}$, $f(x) := 1/x$ is continuous, but not uniformly continuous: For each $\xi \in \mathbb{R}^+$ and each $\delta > 0$, one has

$$f(\xi) - f(\xi + \delta) = \frac{1}{\xi} - \frac{1}{\xi + \delta} = \frac{\delta}{\xi(\xi + \delta)}. \tag{1.41}$$

Thus, for a fixed $\delta > 0$ and $\epsilon > 0$, one has, for each $\xi \in \mathbb{R}^+$ that is chosen smaller than $\delta/2$ and also smaller than $1/(2\epsilon)$,

$$f(\xi) - f(\xi + \delta/2) = \frac{\delta}{2\xi(\xi + \delta/2)} \overset{\xi<\delta/2}{>} \frac{1}{2\xi} \overset{\xi<1/(2\epsilon)}{>} \epsilon,$$

i.e. $x := \xi$ and $y := \xi + \delta/2$ are points such that $|x - y| = \delta/2 < \delta$, but

$$|f(x) - f(y)| = \frac{\delta}{2\xi(\xi + \delta/2)} > \epsilon,$$

showing f is not uniformly continuous.

(b) $g : M \longrightarrow \mathbb{R}$, $g(x) := x^2$ is continuous, but not uniformly continuous: For each $\xi \in \mathbb{R}^+$ and each $\delta > 0$, one has

$$g(\xi + \delta) - g(\xi) = (\xi + \delta)^2 - \xi^2 = 2\xi\delta + \delta^2.$$

Thus, for a fixed $\delta > 0$ and $\epsilon > 0$, one has, for each $\xi \in \mathbb{R}^+$ that is chosen bigger than ϵ/δ,

$$g(\xi + \delta/2) - g(\xi) = \xi\delta + \delta^2/4 > \xi\delta > \epsilon,$$

i.e. $x := \xi$ and $y := \xi + \delta/2$ are points such that $|x - y| = \delta/2 < \delta$, but

$$|g(x) - g(y)| = \frac{\delta}{2\xi(\xi + \delta/2)} > \epsilon,$$

showing f is not uniformly continuous.

(c) $h : M \longrightarrow \mathbb{R}$, $h(x) := \sqrt{x}$ is uniformly continuous, but not Lipschitz continuous: To show that h is uniformly continuous is left as an exercise. If h were Lipschitz continuous, then there needed to be $L \geq 0$ such that

$$\sqrt{\xi + \delta} - \sqrt{\xi} \leq L\delta \tag{1.42}$$

for each $\xi \in \mathbb{R}^+$, $\delta > 0$. However, since

$$\frac{\sqrt{\xi + \delta} - \sqrt{\xi}}{\delta} = \frac{\delta}{\delta(\sqrt{\xi + \delta} + \sqrt{\xi})} = \frac{1}{\sqrt{\xi + \delta} + \sqrt{\xi}}, \tag{1.43}$$

by choosing ξ and δ sufficiently small, one can always make the expression in (1.43) larger than any given L, showing that h is not Lipschitz continuous.

Example 1.53. According to Lem. 1.21(b), the norm $\|\cdot\|$ on a normed vector space X satisfies the inverse triangle inequality

$$\big|\|x\| - \|y\|\big| \leq \|x - y\| \quad \text{for each } x, y \in X, \tag{1.44}$$

i.e. the norm is Lipschitz continuous with Lipschitz constant 1.

Theorem 1.54. *Let (X, d_X) and (Y, d_Y) be metric spaces (for example, normed spaces), $M \subseteq X$, $f : M \longrightarrow Y$. Then the following three statements are equivalent:*

(i) *f is continuous.*

(ii) *For each open set $O \subseteq Y$, the preimage $f^{-1}(O) = \{x \in M : f(x) \in O\}$ is open in M, i.e., for each open $O \subseteq Y$, there exists an open $U \subseteq X$ such that $U \cap M = f^{-1}(O)$.*

(iii) *For each closed set $C \subseteq Y$, the preimage $f^{-1}(C)$ is closed in M, i.e., for each closed $C \subseteq Y$, there exists a closed $A \subseteq X$ such that $A \cap M = f^{-1}(C)$ (cf.* [Phi15a, Rem. 7.46]*).*

Proof. "(i) $\Rightarrow$ (ii)": Assume f is continuous and consider $O \subseteq Y$ open. Let $\xi \in f^{-1}(O)$ and $\eta := f(\xi)$. As O is open, there exists $\epsilon(\xi) >$ such that $B_{\epsilon(\xi)}(\eta) \subseteq O$. Moreover, as f is continuous in ξ, there is $\delta(\xi) > 0$ such that $f\big(M \cap B_{\delta(\xi)}(\xi)\big) \subseteq B_{\epsilon(\xi)}(\eta) \subseteq O$. Set $U := \bigcup_{\xi \in f^{-1}(O)} B_{\delta(\xi)}(\xi)$. Then U is a union of open sets, i.e. U is open by Th. 1.29(a). If $x \in M \cap U$, then $x \in M \cap B_{\delta(\xi)}(\xi)$ for some $\xi \in f^{-1}(O)$ and $f(x) \in B_{\epsilon(\xi)}(f(\xi)) \subseteq O$, showing $x \in f^{-1}(O)$. Conversely, if $x \in f^{-1}(O)$, then $x \in M \cap B_{\delta(x)}(x)$, i.e. $x \in M \cap U$. Thus $U \cap M = f^{-1}(O)$.

"(ii) $\Rightarrow$ (i)": Assume that, for each open set $O \subseteq Y$, $f^{-1}(O)$ is open in M. Let $\xi \in M$ and $\epsilon > 0$. Once again, write $\eta := f(\xi)$. Since $B_\epsilon(\eta)$ is open, we know that $f^{-1}\big(B_\epsilon(\eta)\big)$ is open in M, i.e. there is an open $U \subseteq X$ such that $M \cap U = f^{-1}\big(B_\epsilon(\eta)\big)$. Since $\xi \in U$ and U is open, there is $\delta > 0$ satisfying $B_\delta(\xi) \subseteq U$. Thus, for each $x \in M \cap B_\delta(\xi)$, we have $f(x) \in B_\epsilon(\eta)$, showing the continuity of f in ξ. As ξ was arbitrary, f is continuous.

"(ii) $\Leftrightarrow$ (iii)": $C \subseteq Y$ is closed if, and only if, $Y \setminus C$ is open. Since $f^{-1}(Y \setminus C) = M \setminus f^{-1}(C)$, one has that $f^{-1}(C)$ is closed in M if, and only if, $f^{-1}(Y \setminus C)$ is open in M. Thus, the preimage of all closed subsets of Y is a closed subset in M if, and only if, the preimage of all open subsets of Y is an open subset in M. ■

As already remarked at the beginning of [Phi15a, Sec. 7.2.2] in the one-dimensional context, it is often more convenient to use sequences rather than ϵ- and δ-balls in order to check if functions have limits or are continuous. For functions between metric spaces (in particular, between normed spaces), it is possible to generalize [Phi15a, (8.31)] and [Phi15a, Th. 7.37] to use sequences in that way:

Theorem 1.55. *Let (X, d_X) and (Y, d_Y) be metric spaces (for example, normed spaces), $M \subseteq X$, $f : M \longrightarrow Y$.*

(a) *If $\xi \in X$ is a cluster point of M, then $\lim_{x \to \xi} f(x) = \eta$ exists if, and only if, for each sequence $(x^k)_{k \in \mathbb{N}}$ in $M \setminus \{\xi\}$ with $\lim_{k \to \infty} x^k = \xi$, the sequence $(f(x^k))_{k \in \mathbb{N}}$ converges to $\eta \in Y$, i.e.*

$$\lim_{k \to \infty} x^k = \xi \quad \Rightarrow \quad \lim_{k \to \infty} f(x^k) = \eta. \tag{1.45a}$$

(b) *If $\xi \in M$, f is continuous in ξ if, and only if, for each sequence $(x^k)_{k\in\mathbb{N}}$ in M with $\lim_{k\to\infty} x^k = \xi$, the sequence $(f(x^k))_{k\in\mathbb{N}}$ converges to $f(\xi)$, i.e.*

$$\lim_{k\to\infty} x^k = \xi \quad \Rightarrow \quad \lim_{k\to\infty} f(x^k) = f(\xi). \tag{1.45b}$$

Proof. (a): First assume that $\lim_{x\to\xi} f(x) = \eta$ exists. Moreover, assume that $(x^k)_{k\in\mathbb{N}}$ is a sequence in $M \setminus \{\xi\}$ with $\lim_{k\to\infty} x^k = \xi$. For each $\epsilon > 0$, there is $\delta > 0$ such that $\xi \neq x \in M$ and $d_X(x,\xi) < \delta$ implies $d_Y\big(f(x),\eta\big) < \epsilon$. Since $\lim_{k\to\infty} x^k = \xi$, there is also $N \in \mathbb{N}$ such that, for each $k > N$, $d_X(x^k,\xi) < \delta$. Thus, for each $k > N$, $d_Y\big(f(x^k),\eta\big) < \epsilon$, proving $\lim_{k\to\infty} f(x^k) = \eta$. For the converse, assume that $\lim_{x\to\xi} f(x) = \eta$ is not true. We have to construct a sequence $(x^k)_{k\in\mathbb{N}}$ in $M \setminus \{\xi\}$ with $\lim_{k\to\infty} x^k = \xi$, but $(f(x^k))_{k\in\mathbb{N}}$ does not converge to η. Since $\lim_{x\to\xi} f(x) = \eta$ is not true, there must be some $\epsilon_0 > 0$ such that, for each $1/k$, $k \in \mathbb{N}$, there is at least one $\xi \neq x^k \in M$ satisfying $d_X(x^k,\xi) < 1/k$ and $d_Y\big(f(x^k),\eta\big) \geq \epsilon_0$. Then $(x^k)_{k\in\mathbb{N}}$ is a sequence in $M \setminus \{\xi\}$ with $\lim_{k\to\infty} x^k = \xi$ and $(f(x^k))_{k\in\mathbb{N}}$ does not converge to η.

(b) (the proof is analogous to the proof of [Phi15a, Th. 7.37]): If $\xi \in M$ is not a cluster point of M, then there is $\delta > 0$ such that $M \cap B_\delta(\xi) = \{\xi\}$ (i.e. ξ is an isolated point of M). Then every $f : M \longrightarrow Y$ is continuous in ξ. On the other hand, every sequence in M converging to ξ must be finally equal to ξ, so that (1.45b) is trivially valid at ξ. Thus, the assertion of the theorem holds if $\xi \in M$ is not a cluster point of M. If $\xi \in M$ is a cluster point of M, then one proceeds analogous to the argument in (a): Assume that f is continuous in ξ and $(x^k)_{k\in\mathbb{N}}$ is a sequence in M with $\lim_{k\to\infty} x^k = \xi$. For each $\epsilon > 0$, there is $\delta > 0$ such that $x \in M$ and $d_X(x,\xi) < \delta$ implies $d_Y\big(f(x),f(\xi)\big) < \epsilon$. Since $\lim_{k\to\infty} x^k = \xi$, there is also $N \in \mathbb{N}$ such that, for each $k > N$, $d_X(x^k,\xi) < \delta$. Thus, for each $k > N$, $d_Y\big(f(x^k),f(\xi)\big) < \epsilon$, proving $\lim_{k\to\infty} f(x^k) = f(\xi)$. Conversely, assume that f is not continuous in ξ. We have to construct a sequence $(x^k)_{k\in\mathbb{N}}$ in M with $\lim_{k\to\infty} x^k = \xi$, but $(f(x^k))_{k\in\mathbb{N}}$ does not converge to $f(\xi)$. Since f is not continuous in ξ, there must be some $\epsilon_0 > 0$ such that, for each $1/k$, $k \in \mathbb{N}$, there is at least one $x^k \in M$ satisfying $d_X(x^k,\xi) < 1/k$ and $d_Y\big(f(x^k),f(\xi)\big) \geq \epsilon_0$. Then $(x^k)_{k\in\mathbb{N}}$ is a sequence in M with $\lim_{k\to\infty} x^k = \xi$ and $(f(x^k))_{k\in\mathbb{N}}$ does not converge to $f(\xi)$. ■

Example 1.56. (a) Constant functions are always continuous.

(b) If $(z^k)_{k\in\mathbb{N}}$ is a sequence in $\mathbb{K}^n$, $n \in \mathbb{N}$, such that $\lim_{k\to\infty} z^k = z \in \mathbb{K}^n$, then Th. 1.12 implies that $\lim_{k\to\infty} z_j^k = z_j$ for each $j \in \{1,\ldots,n\}$. Thus, according to Th. 1.55(b), all the *projections* $\pi_j : \mathbb{K}^n \longrightarrow \mathbb{K}$, $\pi_j(z_1,\ldots,z_n) := z_j$ are continuous.

Definition 1.57. If X is a metric space, $M \subseteq X$, and $f : M \longrightarrow \mathbb{K}^n$, $n \in \mathbb{N}$. then the functions $f_1 : M \longrightarrow \mathbb{K}$, …, $f_n : M \longrightarrow \mathbb{K}$, such that $f(x) = (f_1(x),\ldots,f_n(x))$ are called the *coordinate functions* of f.

Theorem 1.58. *If X is a metric space, $M \subseteq X$, and $f : M \longrightarrow \mathbb{K}^n$, $n \in \mathbb{N}$, then the following statements* (i) – (iii) *are equivalent:*

(i) *The function f is continuous.*

(ii) *All the coordinate functions $f_1,\ldots,f_n$ are continuous.*

(iii) *Both the real and the imaginary part of each coordinate function are continuous, i.e. the real-valued functions* $\operatorname{Re} f_1, \operatorname{Im} f_1, \ldots, \operatorname{Re} f_n, \operatorname{Im} f_n$ *all are continuous.*

Proof. The equivalences

$$
\begin{aligned}
\text{(i)} \quad &\overset{\text{Th. 1.55(b)}}{\Leftrightarrow} \quad \forall_{x\in M} \quad \forall_{(x^k)_{k\in\mathbb{N}} \text{ in } M} \quad \Big(\lim_{k\to\infty} x^k = x \quad \Rightarrow \quad \lim_{k\to\infty} f(x^k) = f(x)\Big) \\
&\overset{\text{Th. 1.12}}{\Leftrightarrow} \quad \forall_{x\in M} \quad \forall_{(x^k)_{k\in\mathbb{N}} \text{ in } M} \quad \Big(\lim_{k\to\infty} x^k = x \quad \Rightarrow \quad \forall_{j\in\{1,\ldots,n\}} \quad \lim_{k\to\infty} f_j(x^k) = f_j(x)\Big) \\
&\overset{\text{Th. 1.55(b)}}{\Leftrightarrow} \quad \text{(ii)} \\
&\overset{\text{[Phi15a, (7.2)]}}{\Leftrightarrow} \quad \forall_{x\in M} \; \forall_{(x^k)_{k\in\mathbb{N}} \text{ in } M} \Big(\lim_{k\to\infty} x^k = x \;\Rightarrow\; \forall_{j\in\{1,\ldots,n\}} \Big(\lim_{k\to\infty} \operatorname{Re} f_j(x^k) = \operatorname{Re} f_j(x) \\
&\qquad\qquad \wedge \lim_{k\to\infty} \operatorname{Im} f_j(x^k) = \operatorname{Im} f_j(x)\Big)\Big) \\
&\overset{\text{Th. 1.55(b)}}{\Leftrightarrow} \quad \text{(iii)}
\end{aligned}
$$

prove the theorem. ■

Remark 1.59. Let $X \neq \emptyset$ be an arbitrary nonempty set, $f, g : X \longrightarrow \mathbb{K}$, and $\lambda \in \mathbb{K}$. In [Phi15a, Not. 6.2], we defined the functions $f+g$, λf, fg, f/g, $|f|$, and, for $\mathbb{K} = \mathbb{R}$, also $\max(f,g)$, $\min(f,g)$, f^+, f^-. If Y is an arbitrary vector space over $\mathbb{K}$ and $f, g : X \longrightarrow Y$, then we can generalize the definition of $f+g$ and λf by letting

$$
\begin{aligned}
(f+g) : X \longrightarrow Y, \qquad & (f+g)(x) := f(x) + g(x), && \text{(1.46a)} \\
(\lambda f) : X \longrightarrow Y, \qquad & (\lambda f)(x) := \lambda f(x). && \text{(1.46b)}
\end{aligned}
$$

It turns out that this makes the set of functions from X into Y, $\mathcal{F}(X,Y)$, into a vector space over $\mathbb{K}$ with zero element $f \equiv 0$ (cf. Ex. A.2(c)). Finally, for $f : X \longrightarrow \mathbb{C}^n$, $f(x) = (f_1(x), \ldots, f_n(x))$, $n \in \mathbb{N}$, we define

$$
\begin{aligned}
\operatorname{Re} f : X \longrightarrow \mathbb{R}^n, \qquad & \operatorname{Re} f(x) := (\operatorname{Re} f_1(x), \ldots, \operatorname{Re} f_n(x)), && \text{(1.47a)} \\
\operatorname{Im} f : X \longrightarrow \mathbb{R}^n, \qquad & \operatorname{Im} f(x) := (\operatorname{Im} f_1(x), \ldots, \operatorname{Im} f_n(x)), && \text{(1.47b)} \\
\bar{f} : X \longrightarrow \mathbb{C}^n, \qquad & \bar{f}(x) := (\overline{f_1(x)}, \ldots, \overline{f_n(x)}), && \text{(1.47c)}
\end{aligned}
$$

such that

$$
\begin{aligned}
f &= \operatorname{Re} f + i \operatorname{Im} f, && \text{(1.48a)} \\
\bar{f} &= \operatorname{Re} f - i \operatorname{Im} f. && \text{(1.48b)}
\end{aligned}
$$

Lemma 1.60. *Let* $(X, \|\cdot\|)$ *be a normed vector space, and let* $(x^k)_{k\in\mathbb{N}}$ *and* $(y^k)_{k\in\mathbb{N}}$ *be sequences in* X *with* $\lim_{k\to\infty} x^k = x \in X$ *and* $\lim_{k\to\infty} y^k = y \in X$*. Then the following holds:*

$$
\begin{aligned}
\lim_{k\to\infty} (x^k + y^k) &= x + y, && \text{(1.49a)} \\
\lim_{k\to\infty} (\lambda x^k) &= \lambda x \quad \textit{for each } \lambda \in \mathbb{K}. && \text{(1.49b)}
\end{aligned}
$$

Proof. Since $\lim_{k\to\infty}\|x^k - x\| = 0$ and $\lim_{k\to\infty}\|y^k - y\| = 0$, it follows from $\|x^k + y^k - x - y\| \le \|x^k - x\| + \|y^k - y\|$ that also $\lim_{k\to\infty}\|x^k + y^k - x - y\| = 0$. For each $\lambda \in \mathbb{K}$, one has $\lim_{k\to\infty}\|\lambda x^k - \lambda x\| = \lim_{k\to\infty}(|\lambda|\,\|x^k - x\|) = |\lambda| \lim_{k\to\infty}\|x^k - x\| = 0$. ■

Theorem 1.61. *Let X be a metric space (e.g. a normed space), Y is a normed vector space, and assume that $f, g : X \longrightarrow Y$ are continuous in $\xi \in X$. Then $f + g$ and λf are continuous in ξ for each $\lambda \in \mathbb{K}$ (in particular, $C(X, Y)$ constitutes a subspace of the vector space $\mathcal{F}(X, Y)$ over $\mathbb{K}$). Moreover, if $Y = \mathbb{C}^n$, $n \in \mathbb{N}$, then $\operatorname{Re} f$, $\operatorname{Im} f$, and $\bar{f}$ are all continuous in ξ; if $Y = \mathbb{K}$, then fg, f/g for $g(\xi) \neq 0$, and $|f|$ are all continuous in ξ; if $Y = \mathbb{R}$, then $\max(f, g)$, $\min(f, g)$, f^+, and f^- are all continuous in ξ as well.*

Proof. Let $(x^k)_{k\in\mathbb{N}}$ be a sequence in X such that $\lim_{k\to\infty} x^k = \xi$. Then the continuity of f and g in ξ yields $\lim_{k\to\infty} f(x^k) = f(\xi)$ and $\lim_{k\to\infty} g(x^k) = g(\xi)$. Lemma 1.60 then yields $\lim_{k\to\infty}(f + g)(x^k) = (f + g)(\xi)$ and $\lim_{k\to\infty}(\lambda f)(x^k) = (\lambda f)(\xi)$. For $Y = \mathbb{C}^n$, $n \in \mathbb{N}$, Th. 1.12 together with [Phi15a, (7.2)] and [Phi15a, (7.11f)] shows $\lim_{k\to\infty} \operatorname{Re} f(x^k) = \operatorname{Re} f(\xi)$, $\lim_{k\to\infty} \operatorname{Im} f(x^k) = \operatorname{Im} f(\xi)$, and $\lim_{k\to\infty} \bar{f}(x^k) = \bar{f}(\xi)$, providing the continuity of $\operatorname{Re} f$, $\operatorname{Im} f$, and $\bar{f}$ at ξ. For $Y = \mathbb{K}$, the rules for the limits of sequences in $\mathbb{K}$ [Phi15a, Th. 7.13(a)] yield $\lim_{k\to\infty}(fg)(x^k) = (fg)(\xi)$, $\lim_{k\to\infty}(f/g)(x^k) = (f/g)(\xi)$ for $g(\xi) \neq 0$ (here, one might need to discard some initial part of the sequence $((f/g)(x^k))_{k\in\mathbb{N}}$ to make sure that all the $g(x^k) \neq 0$), and $\lim_{k\to\infty} |f|(x^k) = |f|(\xi)$. This provides the continuity of $f + g$, λf, fg, f/g, and $|f|$ at ξ. Moreover, for $Y = \mathbb{R}$, [Phi15a, Th. 7.13(b)] implies $\lim_{k\to\infty} \max(f, g)(x^k) = \max(f, g)(\xi)$ and $\lim_{k\to\infty} \min(f, g)(x^k) = \min(f, g)(\xi)$, proving the continuity of $\max(f, g)$, $\min(f, g)$, f^+, and f^- at ξ. ■

Example 1.62. Each $\mathbb{K}$-linear function $A : \mathbb{K}^n \longrightarrow \mathbb{K}^m$, $(n, m) \in \mathbb{N}^2$, is continuous: Using the standard unit vectors e_j, for each $z \in \mathbb{K}^n$, one has $A(z) = A(\sum_{j=1}^n z_j e_j) = \sum_{j=1}^n z_j A(e_j)$. Thus, one can build A by summing the functions $A_j : \mathbb{K}^n \longrightarrow \mathbb{K}^m$, $A_j(z) := z_j A(e_j)$ for each $j \in \{1, \ldots, n\}$. Since $\lim_{k\to\infty} z^k = z$ implies $\lim_{k\to\infty} z_j^k = z_j$, which implies $\lim_{k\to\infty} z_j^k A(e_j) = z_j A(e_j)$, all A_j are continuous, and, thus A is continuous by Th. 1.61.

Theorem 1.63. *Let (X, d_X), (Y, d_Y), (Z, d_Z) be metric spaces (for example, normed spaces), $D_f \subseteq X$, $f : D_f \longrightarrow Y$, $D_g \subseteq Y$, $g : D_g \longrightarrow Z$, $f(D_f) \subseteq D_g$. If f is continuous in $\xi \in D_f$ and g is continuous in $f(\xi) \in D_g$, then $g \circ f : D_f \longrightarrow Z$ is continuous in ξ. In consequence, if f and g are both continuous, then the composition $g \circ f$ is also continuous.*

Proof. Let $\xi \in D_f$ and assume that f is continuous in ξ and g is continuous in $f(\xi)$. If $(x^k)_{k\in\mathbb{N}}$ is a sequence in D_f such that $\lim_{k\to\infty} x^k = \xi$, then the continuity of f in ξ implies that $\lim_{k\to\infty} f(x^k) = f(\xi)$. Then the continuity of g in $f(\xi)$ implies that $\lim_{k\to\infty} g(f(x^k)) = g(f(\xi))$, thereby establishing the continuity of $g \circ f$ in ξ. ■

Example 1.64. The function $f : \mathbb{R}^+ \times \mathbb{C} \longrightarrow \mathbb{C}$, $f(x, z) := x^z = \exp(z \ln x)$, is continuous: With the projections $\pi_1, \pi_2 : \mathbb{C}^2 \longrightarrow \mathbb{C}$, we can write $f = \exp \circ(\pi_2(\ln \circ \pi_1))$ (note π_1 is $\mathbb{R}^+$-valued on $\mathbb{R}^+ \times \mathbb{C}$). Since π_1 and π_2 are continuous by Example 1.56(b), $\ln \circ \pi_1$ is continuous by Th. 1.63 and $\pi_2(\ln \circ \pi_1)$ is continuous by Th. 1.61. Finally, $f = \exp \circ(\pi_2(\ln \circ \pi_1))$ is continuous by Th. 1.63.

—

In [Phi15a, Ex. 7.40(b),(c)], we had shown that 1-dimensional polynomials and rational functions are continuous (where they are defined), and we had already briefly considered n-dimensional polynomials and rational functions in [Phi15a, Sec. 6.3]. We will now extend [Phi15a, Ex. 7.40(b),(c)] to n-dimensional polynomials and rational functions:

Theorem 1.65. *Each polynomial $P : \mathbb{K}^n \longrightarrow \mathbb{K}$, $n \in \mathbb{N}$, is continuous and each rational function P/Q is continuous at each $z \in \mathbb{K}^n$ such that $Q(z) \neq 0$.*

Proof. Let

$$\begin{gathered} P : \mathbb{K}^n \longrightarrow \mathbb{K}, \quad P(z) = \sum_{|p| \leq k} a_p z^p, \quad k \in \mathbb{N}_0, \quad p = (p_1, \ldots, p_n) \in (\mathbb{N}_0)^n, \\ |p| = p_1 + \cdots + p_n, \quad z^p = z_1^{p_1} z_2^{p_2} \cdots z_n^{p_n}, \quad a_p \in \mathbb{K}. \end{gathered} \tag{1.50}$$

First, from Ex. 1.56(b), we know that the projections $\pi_j : \mathbb{K}^n \longrightarrow \mathbb{K}$, $\pi_j(z) := z_j$, $j \in \{1, \ldots, n\}$, are continuous. An induction and Th. 1.61 then show the monomials $z \mapsto a_p z^p$ to be continuous, and another induction then shows P to be continuous. Applying Th. 1.61 once more finally shows that each rational function P/Q is continuous at each $z \in \mathbb{K}^n$ such that $Q(z) \neq 0$. ■

Example 1.66. For $n \in \mathbb{N}$, recall the notion of an $n \times n$ matrix over $\mathbb{K}$ (see App. A.3), and note that the set $\mathcal{M}(n, \mathbb{K})$ of $n \times n$ matrices over $\mathbb{K}$ is nothing but $\mathbb{K}^{n^2}$ and, thus, can be considered as a normed vector space in the usual way. Also recall the determinant function $\det : \mathcal{M}(n, \mathbb{K}) \longrightarrow \mathbb{K}$ (see App. A.4).

(a) According to (A.82), the determinant $\det$ is a polynomial on $\mathcal{M}(n, \mathbb{K})$ (i.e. on $\mathbb{K}^{n^2}$), i.e. $\det$ is continuous as a consequence of Th. 1.65.

(b) According to Th. A.48(a), $A \in \mathcal{M}(n, \mathbb{K})$ is invertible if, and only if, $\det(A) \neq 0$. Using (a) and Th. 1.54(ii), this implies that $\mathrm{GL}(n, \mathbb{K}) := \det^{-1}(\mathbb{K} \setminus \{0\})$ is an open subset of $\mathcal{M}(n, \mathbb{K})$ (in Linear Algebra, $\mathrm{GL}(n, \mathbb{K})$ is known as the *general linear group* of degree n over $\mathbb{K}$). Moreover, we claim the map

$$\mathrm{inv} : \mathrm{GL}(n, \mathbb{K}) \longrightarrow \mathrm{GL}(n, \mathbb{K}), \quad \mathrm{inv}(A) := A^{-1},$$

is continuous: Indeed, according to Th. A.51(c), all the coordinate maps inv_{kl} (i.e. the entries of the inverse matrix) are rational functions on $\mathcal{M}(n, \mathbb{K})$ (i.e. on $\mathbb{K}^{n^2}$), i.e. they are continuous as a consequence of Th. 1.65, i.e. inv is continuous by Th. 1.58(ii).

Theorem 1.67. *For a $\mathbb{K}$-linear function $A : X \longrightarrow Y$ between normed vector spaces $(X, \|\cdot\|_X)$ and $(Y, \|\cdot\|_Y)$ over $\mathbb{K}$, the following statements are equivalent:*

(i) *A is continuous.*

(ii) *There exists $\xi \in X$ such that A is continuous in ξ.*

(iii) *A is Lipschitz continuous.*

Proof. (iii) implies (i) according to Lem. 1.51, (i) trivially implies (ii), and it merely remains to show that (ii) implies (iii). To that end, let $\xi \in X$ such that A is continuous in ξ. Thus, for each $\epsilon >$, there is $\delta > 0$ such that $\|x-\xi\|_X < \delta$ implies $\|A(x)-A(\xi)\|_Y < \epsilon$. As A is linear, for each $x \in X$ with $\|x\|_X < \delta$, one has $\|A(x)\|_Y = \|A(x+\xi)-A(\xi)\|_Y < \epsilon$, due to $\|x+\xi-\xi\|_X = \|x\|_X < \delta$. Moreover, one has $\|(\delta x)/2\|_X \leq \delta/2 < \delta$ for each $x \in X$ with $\|x\|_X \leq 1$. Letting $L := 2\epsilon/\delta$, this means that $\|A(x)\|_Y = \|A((\delta x)/2)\|_Y/(\delta/2) < 2\epsilon/\delta = L$ for each $x \in X$ with $\|x\|_X \leq 1$. Thus, for each $x, y \in X$ with $x \neq y$, one has

$$\|A(x) - A(y)\|_Y = \|A(x-y)\|_Y = \|x-y\|_X \left\| A\left(\frac{x-y}{\|x-y\|_X}\right)\right\|_Y \leq L\,\|x-y\|_X. \quad (1.51)$$

Since (1.51) is trivially true for $x = y$, this shows that A is Lipschitz continuous. ■

We will now see two examples that show that, in contrast to linear maps between finite-dimensional spaces as considered in Example 1.62 above, linear maps between infinite-dimensional spaces can be discontinuous.

Example 1.68. (a) Once again, consider the space X from Example 1.39(b) consisting of all sequences in $\mathbb{K}$ that are finally constant and equal to zero, endowed with the norm $\|\cdot\|_{\sup}$. The function

$$A:\ X \longrightarrow \mathbb{K}, \quad A\big((z_n)_{n\in\mathbb{N}}\big) := \sum_{n=1}^{\infty} z_n, \quad (1.52)$$

is clearly linear. However, we will see that A is not continuous: The sequence $(z^k)_{k\in\mathbb{N}}$ defined by

$$z_n^k := \begin{cases} 1/k & \text{for } 1 \leq n \leq k, \\ 0 & \text{for } n > k, \end{cases} \quad (1.53)$$

converges to $0 = (0, 0, \dots) \in X$ with respect to $\|\cdot\|_{\sup}$. However, for each $k \in \mathbb{N}$, $A(z^k) = \sum_{n=1}^{k}(1/k) = 1$, i.e. $\lim_{k\to\infty} A(z^k) = 1 \neq 0 = A(0)$, showing that A is not continuous at 0.

(b) Let X be the normed vector space consisting of all bounded and differentiable functions $f : \mathbb{R} \longrightarrow \mathbb{R}$, endowed with the sup-norm. Then the function $d : X \longrightarrow \mathbb{R}$, $d(f) := f'(0)$, is linear, but not continuous (exercise).

—

A notion related to, but different from, continuity is componentwise continuity (see Def. 1.69). Both notions have to be distinguished carefully, as componentwise continuity does *not* imply continuity (see Example 1.71).

Definition 1.69. Let (Y,d) be a metric space and let $\zeta = (\zeta_1,\ldots,\zeta_n) \in \mathbb{K}^n$, $n \in \mathbb{N}$. A function $f : \mathbb{K}^n \longrightarrow Y$ is called continuous in ζ *with respect to the jth component*, $j \in \{1,\ldots,n\}$, if, and only if, the function

$$\phi : \mathbb{K} \longrightarrow Y, \quad \phi(\alpha) := f(\zeta_1,\ldots,\zeta_{j-1},\alpha,\zeta_{j+1},\ldots,\zeta_n), \tag{1.54}$$

is continuous in $\alpha = \zeta_j$.

Lemma 1.70. *Let (Y,d) be a metric space and let $\zeta = (\zeta_1,\ldots,\zeta_n) \in \mathbb{K}^n$, $n \in \mathbb{N}$. If f is continuous in ζ, then f is continuous in ζ with respect to all components.*

Proof. Let $j \in \{1,\ldots,n\}$ and let $(\alpha_k)_{k\in\mathbb{N}}$ be a sequence in $\mathbb{K}$ with $\lim_{k\to\infty}\alpha_k = \zeta_j$. Then $(z^k)_{k\in\mathbb{N}}$ with $z^k := (\zeta_1,\ldots,\zeta_{j-1},\alpha_k,\zeta_{j+1},\ldots,\zeta_n)$ is a sequence in $\mathbb{K}^n$ with $\lim_{k\to\infty} z^k = \zeta$. Thus, the continuity of f yields $\lim_{k\to\infty} f(z^k) = f(\zeta)$. If ϕ is defined as in (1.54), then $\phi(\alpha_k) = f(z^k)$, showing $\lim_{k\to\infty}\phi(\alpha_k) = f(\zeta) = \phi(\zeta_j)$, i.e. ϕ is continuous in ζ_j. We have, hence, shown, for each $j \in \{1,\ldots,n\}$, that f is continuous in ζ with respect to the jth component. ■

Example 1.71. A function can be continuous with respect to all components at a point ζ without being continuous at ζ: Consider the function

$$f : \mathbb{K}^2 \longrightarrow \mathbb{K}, \quad f(z,w) := \begin{cases} 0 & \text{for } zw = 0, \\ 1 & \text{for } zw \neq 0. \end{cases} \tag{1.55}$$

Let $\phi_1,\phi_2 : \mathbb{K} \longrightarrow \mathbb{K}$, $\phi_1(\alpha) := f(\alpha,0)$, $\phi_2(\alpha) := f(0,\alpha)$. Then both ϕ_1 and ϕ_2 are identically 0 and, in particular, continuous at $\alpha = 0$. However, f is not continuous at $(0,0)$, since, for example,

$$(z^k,w^k) := \begin{cases} (1/k,0) & \text{for } k \text{ even}, \\ (1/k,1/k) & \text{for } k \text{ odd} \end{cases} \tag{1.56}$$

yields a sequence that converges to $(0,0)$, but $f(z^k,w^k) = 0$ if k is even and $f(z^k,w^k) = 1$ if k is odd, i.e. the sequence $(f(z^k,w^k))_{k\in\mathbb{N}}$ does not converge.

1.6 Convex Functions and Norms on $\mathbb{K}^n$

Even though convex functions are an important topic in their own right, here the main motivation is to provide a proof for the so-called Minkowski inequality, i.e. for the triangle inequality of the p-norm on $\mathbb{K}^n$, defined by $\|z\|_p := (\sum_{j=1}^n |z_j|^p)^{1/p}$.

The idea is to call a real-valued function f convex if, and only if, each line segment connecting two points on the graph of f lies above this graph, and to call f concave if, and only if, each such line segment lies below the graph of f. Noting that, for $x_1 < x_2$, the line through the two points $(x_1,f(x_1))$ and $(x_2,f(x_2))$ is represented by the equation

$$L(x) = \frac{x_2 - x}{x_2 - x_1} f(x_1) + \frac{x - x_1}{x_2 - x_1} f(x_2), \tag{1.57}$$

this leads to the following definition:

Definition 1.72. Let $I \subseteq \mathbb{R}$ be an interval (I can be open, closed, or half-open, it can be for finite or of infinite length) and $f : I \longrightarrow \mathbb{R}$. Then f is called *convex* if, and only if, for each $x_1, x, x_2 \in I$ such that $x_1 < x < x_2$, one has

$$f(x) \leq \frac{x_2 - x}{x_2 - x_1} f(x_1) + \frac{x - x_1}{x_2 - x_1} f(x_2); \tag{1.58a}$$

f is called *concave* if, and only if, for each $x_1, x, x_2 \in I$ such that $x_1 < x < x_2$, one has

$$f(x) \geq \frac{x_2 - x}{x_2 - x_1} f(x_1) + \frac{x - x_1}{x_2 - x_1} f(x_2). \tag{1.58b}$$

Moreover, f is called strictly convex (resp. strictly concave) if, and only if, (1.58a) (resp. (1.58b)) always holds with strict inequality.

Lemma 1.73. *Let $I \subseteq \mathbb{R}$ be an interval and $f : I \longrightarrow \mathbb{R}$.*

(a) *f is concave if, and only if, $-f$ is convex.*

(b) *f is (strictly) convex if, and only if, for each $a, b \in I$ such that $a \neq b$ and each $\lambda \in]0,1[$, the following estimate holds (with strict inequality):*

$$f\big(\lambda a + (1-\lambda)b\big) \leq \lambda f(a) + (1-\lambda) f(b). \tag{1.59a}$$

(c) *f is (strictly) concave if, and only if, for each $a, b \in I$ such that $a \neq b$ and each $\lambda \in]0,1[$, the following estimate holds (with strict inequality):*

$$f\big(\lambda a + (1-\lambda)b\big) \geq \lambda f(a) + (1-\lambda) f(b). \tag{1.59b}$$

Proof. (a): Merely multiply (1.58b) by (-1) and compare with (1.58a).

(b): Given $x_1, x, x_2 \in I$ with $x_1 < x < x_2$, let $\lambda := (x_2 - x)/(x_2 - x_1)$. Then $0 < \lambda < 1$ as well as

$$\begin{aligned} \lambda x_1 + (1-\lambda) x_2 &= \frac{x_1(x_2 - x)}{x_2 - x_1} + x_2 \left(\frac{x_2 - x_1}{x_2 - x_1} - \frac{x_2 - x}{x_2 - x_1} \right) = \frac{x_1(x_2 - x) + x_2(x - x_1)}{x_2 - x_1} \\ &= \frac{x(x_2 - x_1)}{x_2 - x_1} = x. \end{aligned} \tag{1.60}$$

Letting $a := x_1$ and $b := x_2$, this shows that (1.59a) implies (1.58a). Conversely, given $a, b \in I$ with $a \neq b$, let $x_1 := \min\{a, b\}$, $x_2 := \max\{a, b\}$. If $0 < \lambda < 1$, then, letting $x := \lambda a + (1-\lambda)b$, note that

$$x_1 = \lambda x_1 + (1-\lambda) x_1 < \lambda a + (1-\lambda) b = x < \lambda x_2 + (1-\lambda) x_2 = x_2. \tag{1.61}$$

Then

$$a = x_2 \quad \Rightarrow \quad \frac{x_2 - x}{x_2 - x_1} = \frac{(1-\lambda)(a-b)}{x_2 - x_1} = 1 - \lambda, \qquad \frac{x - x_1}{x_2 - x_1} = \lambda, \tag{1.62a}$$

$$a = x_1 \quad \Rightarrow \quad \frac{x_2 - x}{x_2 - x_1} = \frac{\lambda(b-a)}{x_2 - x_1} = \lambda, \qquad \frac{x - x_1}{x_2 - x_1} = 1 - \lambda. \tag{1.62b}$$

Thus, in each case,

$$\frac{x_2 - x}{x_2 - x_1} f(x_1) + \frac{x - x_1}{x_2 - x_1} f(x_2) = \lambda f(a) + (1 - \lambda) f(b), \tag{1.63}$$

i.e. (1.58a) implies (1.59a).

(c) follows by combining (a) and (b). ■

Example 1.74. Since $|\lambda x + (1 - \lambda)y| \leq \lambda|x| + (1 - \lambda)|y|$ for each $0 < \lambda < 1$ and each $x, y \in \mathbb{R}$, the absolute value function is convex. This example also shows that a convex function does not need to be differentiable.

Lemma 1.75. *Let $I \subseteq \mathbb{R}$ be an interval. Then $f : I \longrightarrow \mathbb{R}$ is convex if, and only if, for each $x_1, x, x_2 \in I$ such that $x_1 < x < x_2$, one has*

$$\frac{f(x) - f(x_1)}{x - x_1} \leq \frac{f(x_2) - f(x)}{x_2 - x}, \tag{1.64a}$$

and also if, and only if, for each $x_1, x, x_2 \in I$ such that $x_1 < x < x_2$, one has

$$\frac{f(x) - f(x_1)}{x - x_1} \leq \frac{f(x_2) - f(x_1)}{x_2 - x_1} \leq \frac{f(x_2) - f(x)}{x_2 - x}. \tag{1.64b}$$

Proof. By definition, convexity of f is equivalent to the validity of (1.58a) for each $x_1, x, x_2 \in I$ such that $x_1 < x < x_2$. Multiplying (1.58a) with the positive number $x_2 - x_1$ shows its equivalence with

$$\big((x_2 - x) + (x - x_1)\big) f(x) \leq (x_2 - x) f(x_1) + (x - x_1) f(x_2), \tag{1.65a}$$

which, in turn, is equivalent to

$$(x_2 - x)\big(f(x) - f(x_1)\big) \leq (x - x_1)\big(f(x_2) - f(x)\big), \tag{1.65b}$$

which, after division by the positive number $(x_2 - x)(x - x_1)$ is equivalent to (1.64a). That (1.64b) implies (1.64a) is trivial. On the other hand, (1.58a) implies

$$f(x) - f(x_1) \leq \left(1 - \frac{x - x_1}{x_2 - x_1}\right) f(x_1) - f(x_1) + \frac{x - x_1}{x_2 - x_1} f(x_2) = \frac{x - x_1}{x_2 - x_1}\big(f(x_2) - f(x_1)\big),$$

i.e. the left-hand inequality of (1.64b). Analogously, (1.58a) also implies

$$f(x_2) - f(x) \geq f(x_2) - \frac{x_2 - x}{x_2 - x_1} f(x_1) - \left(1 - \frac{x_2 - x}{x_2 - x_1}\right) f(x_2) = \frac{x_2 - x}{x_2 - x_1}\big(f(x_2) - f(x_1)\big)$$

i.e. the right-hand inequality of (1.64b), thereby completing the proof of the lemma. ■

For differentiable functions, one can formulate convexity criteria in terms of the derivative:

Proposition 1.76. *Let $a < b$, and suppose that $f : [a,b] \longrightarrow \mathbb{R}$ is continuous on $[a,b]$ and differentiable on $]a,b[$. Then f is convex (resp. concave) on $[a,b]$ if, and only if, the derivative f' is increasing (resp. decreasing) on $]a,b[$.*

Proof. Since $(-f)' = -f'$ and $-f'$ is increasing if, and only if, f' is decreasing, it suffices to consider the convex case. So assume that f is convex. Then for each $x_1, x, x_2 \in]a,b[$ such that $x_1 < x < x_2$, one has the validity of (1.64b). Thus

$$f'(x_1) = \lim_{x \downarrow x_1} \frac{f(x) - f(x_1)}{x - x_1} \leq \frac{f(x_2) - f(x_1)}{x_2 - x_1} \leq \lim_{x \uparrow x_2} \frac{f(x_2) - f(x)}{x_2 - x} = f'(x_2), \tag{1.66}$$

showing that f' is increasing on $]a,b[$. On the other hand, if f' is increasing on $]a,b[$, then for each $x_1, x, x_2 \in [a,b]$ such that $x_1 < x < x_2$, the mean value theorem [Phi15a, Th. 9.17] yields $\xi_1 \in]x_1, x[$ and $\xi_2 \in]x, x_2[$ such that

$$\frac{f(x) - f(x_1)}{x - x_1} = f'(\xi_1) \quad \text{and} \quad \frac{f(x_2) - f(x)}{x_2 - x} = f'(\xi_2). \tag{1.67}$$

As $\xi_1 < \xi_2$ and f' is increasing, (1.67) implies (1.64a) and, thus, the convexity of f. ■

Proposition 1.77. *Let $a < b$, and suppose that $f : [a,b] \longrightarrow \mathbb{R}$ is continuous on $[a,b]$ and twice differentiable on $]a,b[$.*

(a) *f is convex (resp. concave) on $[a,b]$ if, and only if, $f'' \geq 0$ (resp. $f'' \leq 0$) on $]a,b[$.*

(b) *If $f'' > 0$ (resp. $f'' < 0$) on $]a,b[$, then f is strictly convex (resp. strictly concave).*

Proof. Since $-f'' \geq 0$ if, and only if $f'' \leq 0$; and $-f'' > 0$ if, and only if $f'' < 0$, if suffices to consider the convex cases. Moreover, for (a), one merely has to combine Prop. 1.76 with the fact that f' is increasing on $]a,b[$ if, and only if, $f'' \geq 0$ on $]a,b[$. For (b), we proceed by contraposition and assume that f is not strictly convex. Then the argument from the proof of Lem. 1.75 shows that there are $x_1, x, x_2 \in [a,b]$ such that $x_1 < x < x_2$ and

$$\frac{f(x) - f(x_1)}{x - x_1} \geq \frac{f(x_2) - f(x)}{x_2 - x}. \tag{1.68}$$

As in Prop. 1.76, the mean value theorem [Phi15a, Th. 9.17] yields $\xi_1 \in]x_1, x[$ and $\xi_2 \in]x, x_2[$ such that (1.67) holds. Together with (1.68), one obtains $f'(\xi_1) \geq f'(\xi_2)$, i.e. f' is not strictly increasing, i.e. $f'' > 0$ does not hold everywhere on $]a,b[$. ■

Example 1.78. (a) Since for $f : \mathbb{R} \longrightarrow \mathbb{R}$, $f(x) = e^x$, it is $f''(x) = e^x > 0$, the exponential function is strictly convex on $\mathbb{R}$.

(b) Since for $f : \mathbb{R}^+ \longrightarrow \mathbb{R}$, $f(x) = \ln x$, it is $f''(x) = -1/x^2 < 0$, the natural logarithm is strictly concave on $\mathbb{R}^+$.

Theorem 1.79 (Jensen's inequality)**.** *Let $I \subseteq \mathbb{R}$ be an interval and let $f : I \longrightarrow \mathbb{R}$ be convex. If $n \in \mathbb{N}$ and $\lambda_1, \ldots, \lambda_n > 0$ such that $\lambda_1 + \cdots + \lambda_n = 1$, then*

$$\underset{x_1,\ldots,x_n \in I}{\forall} \quad f(\lambda_1 x_1 + \cdots + \lambda_n x_n) \leq \lambda_1 f(x_1) + \cdots + \lambda_n f(x_n). \tag{1.69a}$$

If f is concave, then

$$\underset{x_1,\ldots,x_n \in I}{\forall} \quad f(\lambda_1 x_1 + \cdots + \lambda_n x_n) \geq \lambda_1 f(x_1) + \cdots + \lambda_n f(x_n). \tag{1.69b}$$

If f is strictly convex or strictly concave, then equality in the above inequalities can only hold if $x_1 = \cdots = x_n$.

Proof. If one lets $a := \min\{x_1, \ldots, x_n\}$, $b := \max\{x_1, \ldots, x_n\}$, and $\bar{x} := \lambda_1 x_1 + \cdots + \lambda_n x_n$, then

$$a = \sum_{j=1}^{n} \lambda_j \, a \leq \bar{x} \leq \sum_{j=1}^{n} \lambda_j \, b = b \quad \Rightarrow \quad \bar{x} \in I. \tag{1.70}$$

Since f is (strictly) concave if, and only if, $-f$ is (strictly) convex, it suffices to consider the cases where f is convex and where f is strictly convex. Thus, we assume that f is convex and prove (1.69a) by induction. For $n = 1$, one has $\lambda_1 = 1$ and there is nothing to prove. For $n = 2$, (1.69a) reduces to (1.59a), which holds due to the convexity of f. Finally, let $n > 2$ and assume that (1.69a) already holds for each $1 \leq l \leq n-1$. Set

$$\lambda := \lambda_1 + \cdots + \lambda_{n-1}, \quad x := \frac{\lambda_1}{\lambda} x_1 + \cdots + \frac{\lambda_{n-1}}{\lambda} x_{n-1}. \tag{1.71}$$

Then $x \in I$ follows as in (1.70). One computes

$$\begin{aligned} f(\lambda_1 x_1 + \cdots + \lambda_n x_n) &= f\left(\sum_{j=1}^{n-1} \lambda_j x_j + \lambda_n x_n\right) = f(\lambda x + \lambda_n x_n) \\ &\overset{l=2}{\leq} \lambda f(x) + \lambda_n f(x_n) \overset{l=n-1}{\leq} \lambda \sum_{j=1}^{n-1} \frac{\lambda_j}{\lambda} f(x_j) + \lambda_n f(x_n) \\ &= \lambda_1 f(x_1) + \cdots + \lambda_n f(x_n), \end{aligned} \tag{1.72}$$

thereby completing the induction, and, thus, the proof of (1.69a). If f is strictly convex and (1.69a) holds with equality, then one can also proceed by induction to prove the equality of the x_j. Again, if $n = 1$, then there is nothing to prove. If $n = 2$, and $x_1 \neq x_2$, then strict convexity requires (1.69a) to hold with strict inequality. Thus $x_1 = x_2$. Now let $n > 2$. It is noted that (1.72) still holds. By hypothesis, the first and last term in (1.72) are now equal, implying that all terms in (1.72) must be equal. Using the induction hypothesis for $l = 2$ and the corresponding equality in (1.72), we conclude that $x = x_n$. Using the induction hypothesis for $l = n-1$ and the corresponding equality in (1.72), we conclude that $x_1 = \cdots = x_{n-1}$. Finally, $x = x_n$ and $x_1 = \cdots = x_{n-1}$ are combined using (1.71) to get $x_1 = x_n$, finishing the proof of the theorem. ■

Theorem 1.80 (Inequality Between the Weighted Arithmetic Mean and the Weighted Geometric Mean)**.** *If $n \in \mathbb{N}$, $x_1, \ldots, x_n \geq 0$ and $\lambda_1, \ldots, \lambda_n > 0$ such that $\lambda_1 + \cdots + \lambda_n = 1$, then*

$$x_1^{\lambda_1} \cdots x_n^{\lambda_n} \leq \lambda_1 x_1 + \cdots + \lambda_n x_n, \tag{1.73}$$

where equality occurs if, and only if, $x_1 = \cdots = x_n$. In particular, for $\lambda_1 = \cdots = \lambda_n = \frac{1}{n}$, one recovers the inequality between the arithmetic and the geometric mean without weights, known from [Phi15a, Th. 7.63].

Proof. If at least one of the x_j is 0, then (1.73) becomes the true statement $0 \leq \sum_{j=1}^n \lambda_j x_j$ with strict inequality if, and only if, at least one $x_j > 0$. Thus, it remains to consider the case $x_1, \ldots, x_n > 0$. As we noted in Ex. 1.78(b), the natural logarithm $\ln : \mathbb{R}^+ \longrightarrow \mathbb{R}$ is concave and even strictly concave. Employing Jensen's inequality (1.69b) yields

$$\ln(\lambda_1 x_1 + \cdots + \lambda_n x_n) \geq \lambda_1 \ln x_1 + \cdots + \lambda_n \ln x_n = \ln(x_1^{\lambda_1} \cdots x_n^{\lambda_n}). \tag{1.74}$$

Applying the exponential function to both sides of (1.74), one obtains (1.73). Since (1.74) is equivalent to (1.73), the strict concavity of ln yields that equality in (1.74) implies $x_1 = \cdots = x_n$. ■

Definition 1.81. For $n \in \mathbb{N}$, $p \in [1, \infty[$, the function

$$\|\cdot\|_p : \mathbb{K}^n \longrightarrow \mathbb{R}_0^+, \quad \|x\|_p := \left(\sum_{j=1}^n |x_j|^p \right)^{1/p}, \tag{1.75}$$

is called the *p-norm* on $\mathbb{K}^n$ (that the p-norm is, indeed, a norm is the result formulated as Cor. 1.84 below).

Theorem 1.82 (Hölder inequality)**.** *If $n \in \mathbb{N}$ and $p, q > 1$ such that $\frac{1}{p} + \frac{1}{q} = 1$, then*

$$|a \cdot b| \leq \|a\|_p \|b\|_q \quad \text{for each } a, b \in \mathbb{K}^n. \tag{1.76}$$

Proof. If $a = 0$ or $b = 0$, then there is nothing to prove. So let $a \neq 0$ and $b \neq 0$. For each $j \in \{1, \ldots, n\}$, apply (1.73) with $\lambda_1 = 1/p$, $\lambda_2 = 1/q$, $x_1 = |a_j|^p / \|a\|_p^p$ and $x_2 = |b_j|^q / \|b\|_q^q$, to get

$$\frac{|a_j||b_j|}{\|a\|_p \|b\|_q} \leq \frac{1}{p} \frac{|a_j|^p}{\|a\|_p^p} + \frac{1}{q} \frac{|b_j|^q}{\|b\|_q^q}. \tag{1.77a}$$

Summing (1.77a) over $j \in \{1, \ldots, n\}$ yields 1 on the right-hand side, and, thus,

$$|a \cdot b| = \left| \sum_{j=1}^n a_j \bar{b}_j \right| \leq \sum_{j=1}^n |a_j| \, |b_j| \overset{\text{summed (1.77a)}}{\leq} \|a\|_p \|b\|_q, \tag{1.77b}$$

proving (1.76). ■

Theorem 1.83 (Minkowski inequality)**.** *For each $p \geq 1$, $z, w \in \mathbb{K}^n$, $n \in \mathbb{N}$, one has*

$$\|z + w\|_p \leq \|z\|_p + \|w\|_p. \tag{1.78}$$

Proof. For $p = 1$, (1.78) follows directly from the triangle inequality for the absolute value in $\mathbb{K}$. It remains to consider the case $p > 1$. In that case, define $q := p/(p-1)$, i.e. $1/p + 1/q = 1$. Also define $a \in \mathbb{R}^n$ by letting $a_j := |z_j + w_j|^{p-1} \in \mathbb{R}_0^+$ for each $j \in \{1, \ldots, n\}$, and notice

$$|z_j + w_j|^p = |z_j + w_j|\, a_j \leq |z_j|\, a_j + |w_j|\, a_j. \tag{1.79a}$$

Summing (1.79a) over $j \in \{1, \ldots, n\}$ and applying the Hölder inequality (1.76), one obtains

$$\|z + w\|_p^p \leq (|z_1|, \ldots, |z_n|) \cdot a + (|w_1|, \ldots, |w_n|) \cdot a \leq \|z\|_p \|a\|_q + \|w\|_p \|a\|_q. \tag{1.79b}$$

As $q(p-1) = p$, it is $a_j^q = |z_j + w_j|^p$, and, thus

$$\|a\|_q = \left(\sum_{j=1}^n |z_j + w_j|^p \right)^{\frac{1}{p}\frac{p}{q}} = \|z + w\|_p^{p-1}, \tag{1.79c}$$

where $p/q = p - 1$ was used in the last step. Finally, combining (1.79b) with (1.79c) yields (1.78). ∎

Corollary 1.84. *For each $n \in \mathbb{N}$, $p \in [1, \infty[$, the p-norm on $\mathbb{K}^n$ constitutes, indeed, a norm on $\mathbb{K}^n$.*

Proof. If $z = 0$, then $\|z\|_p = 0$ follows directly from (1.75). If $z \neq 0$, then there is $j \in \{1, \ldots, n\}$ such that $|z_j| > 0$. Then (1.75) provides $\|z\|_p \geq |z_j| > 0$. If $\lambda \in \mathbb{K}$ and $z \in \mathbb{K}^n$, then $\|\lambda z\|_p = (\sum_{j=1}^n |\lambda z_j|^p)^{1/p} = (|\lambda|^p \sum_{j=1}^n |z_j|^p)^{1/p} = |\lambda| \|z\|_p$. The proof is concluded by noticing that the triangle inequality is the same as the Minkowski inequality (1.78). ∎

1.7 Inner Products and Hilbert Space

Definition 1.85. Let X be a vector space over $\mathbb{K}$. A function $\langle \cdot, \cdot \rangle : X \times X \longrightarrow \mathbb{K}$ is called an *inner product* or a *scalar product* on X if, and only if, the following three conditions are satisfied:

(i) $\langle x, x \rangle \in \mathbb{R}^+$ for each $0 \neq x \in X$.

(ii) $\langle \lambda x + \mu y, z \rangle = \lambda \langle x, z \rangle + \mu \langle y, z \rangle$ for each $x, y, z \in X$ and each $\lambda, \mu \in \mathbb{K}$ (i.e. an inner product is $\mathbb{K}$-linear in its first argument).

(iii) $\langle x, y \rangle = \overline{\langle y, x \rangle}$ for each $x, y \in X$ (i.e. an inner product is *conjugate-symmetric*, even *symmetric* for $\mathbb{K} = \mathbb{R}$).

Lemma 1.86. *For each inner product $\langle \cdot, \cdot \rangle$ on a vector space X over $\mathbb{K}$, the following formulas are valid:*

(a) $\langle x, \lambda y + \mu z\rangle = \bar{\lambda}\langle x, y\rangle + \bar{\mu}\langle x, z\rangle$ *for each* $x, y, z \in X$ *and each* $\lambda, \mu \in \mathbb{K}$, *i.e.* $\langle \cdot, \cdot \rangle$ *is* conjugate-linear *(also called* antilinear*) in its second argument, even linear for* $\mathbb{K} = \mathbb{R}$. *Together with Def.* 1.85(ii), *this means that* $\langle \cdot, \cdot \rangle$ *is a* sesquilinear form, *even a* bilinear form *for* $\mathbb{K} = \mathbb{R}$.

(b) $\langle 0, x\rangle = \langle x, 0\rangle = 0$ *for each* $x \in X$.

Proof. (a): One computes, for each $x, y, z \in X$ and each $\lambda, \mu \in \mathbb{K}$,

$$\begin{aligned} \langle x, \lambda y + \mu z\rangle &\overset{\text{Def. 1.85(iii)}}{=} \overline{\langle \lambda y + \mu z, x\rangle} \overset{\text{Def. 1.85(ii)}}{=} \overline{\lambda\langle y, x\rangle + \mu\langle z, x\rangle} \\ &= \bar{\lambda}\,\overline{\langle y, x\rangle} + \bar{\mu}\,\overline{\langle z, x\rangle} \overset{\text{Def. 1.85(iii)}}{=} \bar{\lambda}\langle x, y\rangle + \bar{\mu}\langle x, z\rangle. \end{aligned} \tag{1.80a}$$

(b): One computes, for each $x \in X$,

$$\overline{\langle x, 0\rangle} \overset{\text{Def. 1.85(iii)}}{=} \langle 0, x\rangle = \langle 0x, x\rangle \overset{\text{Def. 1.85(ii)}}{=} 0\langle x, x\rangle = 0, \tag{1.80b}$$

thereby completing the proof of the lemma. ■

Theorem 1.87. *The following* Cauchy-Schwarz inequality (1.81) *holds for each inner product* $\langle \cdot, \cdot \rangle$ *on a vector space* X *over* $\mathbb{K}$*:*

$$|\langle x, y\rangle| \le \|x\|\,\|y\| \quad \textit{for each} \quad x, y \in X, \tag{1.81}$$

where

$$\|x\| := \sqrt{\langle x, x\rangle},\ \|y\| := \sqrt{\langle y, y\rangle}. \tag{1.82}$$

Moreover, equality in (1.81) *holds if, and only if,* x *and* y *are linearly dependent, i.e. if, and only if,* $y = 0$ *or there exists* $\lambda \in \mathbb{K}$ *such that* $x = \lambda y$.

Proof. If $y = 0$, then it is immediate that both sides of (1.81) vanish. If $x = \lambda y$ with $\lambda \in \mathbb{K}$, then $|\langle x, y\rangle| = |\lambda\langle y, y\rangle| = |\lambda|\|y\|^2 = \sqrt{\lambda\bar{\lambda}\langle y, y\rangle}\|y\| = \|x\|\,\|y\|$, showing that (1.81) holds with equality. If x and y are not linearly independent, then $y \neq 0$ and $x - \lambda y \neq 0$ for each $\lambda \in \mathbb{K}$, i.e.

$$\begin{aligned} 0 &< \langle x - \lambda y, x - \lambda y\rangle = \langle x, x - \lambda y\rangle - \lambda\langle y, x - \lambda y\rangle \\ &= \langle x, x\rangle - \bar{\lambda}\langle x, y\rangle - \lambda\langle y, x\rangle + \lambda\bar{\lambda}\langle y, y\rangle = \|x\|^2 - \bar{\lambda}\langle x, y\rangle - \lambda\,\overline{\langle x, y\rangle} + |\lambda|^2\|y\|^2. \end{aligned} \tag{1.83}$$

Since (1.83) is valid for each $\lambda \in \mathbb{K}$, one can set $\lambda := \langle x, y\rangle / \|y\|^2$ (using $y \neq 0$) to get

$$0 < \|x\|^2 - \frac{2\,\langle x, y\rangle\,\overline{\langle x, y\rangle}}{\|y\|^2} + \frac{\langle x, y\rangle\,\overline{\langle x, y\rangle}}{\|y\|^2} = \frac{\|x\|^2\|y\|^2 - \langle x, y\rangle\,\overline{\langle x, y\rangle}}{\|y\|^2}, \tag{1.84}$$

or $\langle x, y\rangle\,\overline{\langle x, y\rangle} < \|x\|^2\|y\|^2$. Finally, taking the square root on both sides shows that (1.81) holds with strict inequality. ■

Proposition 1.88. *If X is a vector space over $\mathbb{K}$ with an inner product $\langle\cdot,\cdot\rangle$, then the map*

$$\|\cdot\| : X \longrightarrow \mathbb{R}_0^+, \quad \|x\| := \sqrt{\langle x, x\rangle}, \tag{1.85}$$

defines a norm on X. One calls this the norm induced *by the inner product.*

Proof. If $x = 0$, then $\langle x, x\rangle = 0$ and $\|x\| = 0$ as well. Conversely, if $x \neq 0$, then $\langle x, x\rangle > 0$ and $\|x\| > 0$ as well, showing that $\|\cdot\|$ is positive definite. For $\lambda \in \mathbb{K}$ and $x \in X$, one has $\|\lambda x\| = \sqrt{\lambda\bar{\lambda}\langle x, x\rangle} = \sqrt{|\lambda|^2\langle x, x\rangle} = |\lambda|\|x\|$, showing that $\|\cdot\|$ is homogeneous of degree 1. Finally, if $x, y \in X$, then

$$\begin{aligned} \|x+y\|^2 &= \langle x+y, x+y\rangle = \|x\|^2 + \langle x, y\rangle + \langle y, x\rangle + \|y\|^2 \\ &\overset{(1.81)}{\leq} \|x\|^2 + 2\|x\|\,\|y\| + \|y\|^2 = \big(\|x\| + \|y\|\big)^2, \end{aligned} \tag{1.86}$$

establishing that $\|\cdot\|$ satisfies the triangle inequality. In conclusion, we have shown that $\|\cdot\|$ constitutes a norm on X. ■

Definition 1.89. Let X be a vector space over $\mathbb{K}$. If $\langle\cdot,\cdot\rangle$ is an inner product on X, then $\big(X, \langle\cdot,\cdot\rangle\big)$ is called an *inner product space* or a *pre-Hilbert space*. An inner product space is called a *Hilbert space* if, and only if, $(X, \|\cdot\|)$ is a Banach space, where $\|\cdot\|$ is the induced norm, i.e. $\|x\| := \sqrt{\langle x, x\rangle}$. Frequently, the inner product on X is understood and X itself is referred to as an inner product space or Hilbert space.

Example 1.90. We now come back to the space $\mathbb{K}^n$, $n \in \mathbb{N}$, with the inner product defined by (1.1c) and the length (norm) defined by (1.1d). Let us verify that (1.1c), indeed, defines an inner product in the sense of Def. 1.85: If $z \neq 0$, then there is $j_0 \in \{1, \ldots, n\}$ such that $z_{j_0} \neq 0$. Thus, $z \cdot z = \sum_{j=1}^n |z_j|^2 \geq |z_{j_0}|^2 > 0$, i.e. Def. 1.85(i) is satisfied. Next, let $z, w, u \in \mathbb{K}^n$ and $\lambda, \mu \in \mathbb{K}$. One computes

$$(\lambda z + \mu w)\cdot u = \sum_{j=1}^n (\lambda z_j + \mu w_j)\bar{u}_j = \sum_{j=1}^n \lambda z_j \bar{u}_j + \sum_{j=1}^n \mu w_j \bar{u}_j = \lambda(z\cdot u) + \mu(w\cdot u), \tag{1.87a}$$

i.e. Def. 1.85(ii) is satisfied. For Def. 1.85(iii), merely note that

$$z \cdot w = \sum_{j=1}^n z_j \bar{w}_j = \overline{\sum_{j=1}^n w_j \bar{z}_j} = \overline{w \cdot z}. \tag{1.87b}$$

Hence, we have shown that (1.1c) defines an inner product according to Def. 1.85. Since the norm defined by (1.1d) is the same as the norm induced by the inner product, this also proves the triangle inequality of Lem. 1.3(c). Due to Th. 1.16(a), the norm of (1.1d) is complete, i.e. $\mathbb{K}^n$ with the norm of (1.1d) is a Banach space and $\mathbb{K}^n$ with the inner product of (1.1c) is a Hilbert space.

Definition 1.91. If $\big(X, \langle\cdot,\cdot\rangle\big)$ is an inner product space, then $x, y \in X$ are called *orthogonal* or *perpendicular* (denoted $x \perp y$) if, and only if, $\langle x, y\rangle = 0$. A *unit vector* is $x \in X$ such that $\|x\| = 1$, where $\|\cdot\|$ is the induced norm. An *orthogonal system* is a family $(x^\alpha)_{\alpha\in I}$, $x^\alpha \in X$, I being some index set, such that $\langle x^\alpha, x^\beta\rangle = 0$ for each $\alpha, \beta \in I$ with $\alpha \neq \beta$. An orthogonal system is called an *orthonormal system* if, and only if, it consists entirely of unit vectors.

Remark 1.92. If $\big(X, \langle\cdot,\cdot\rangle\big)$ is an inner product space, then one has *Pythagoras' theorem*, namely that for each $x, y \in X$ with $x \perp y$:

$$\|x+y\|^2 = \|x\|^2 + \langle x, y\rangle + \langle y, x\rangle + \|y\|^2 = \|x\|^2 + \|y\|^2. \tag{1.88}$$

1.8 Equivalence of Metrics and Equivalence of Norms

Now that we have seen an uncountable number of different norms on $\mathbb{K}^n$ (namely the p-norms of Def. 1.81), it is an important result that they all generate the same open sets (and, thus, the same notions of convergence) on $\mathbb{K}^n$ – all norms on $\mathbb{K}^n$ are equivalent. Before we can state and prove this result in Th. 1.95, we have to introduce the notion of equivalence for metrics and norms. We will also see that, even though all norms on $\mathbb{K}^n$ are equivalent, norms on other normed vector spaces are not necessarily equivalent (see Example 1.98 below).

Definition 1.93. (a) Let d_1 and d_2 be metrics on a set X. Then d_1 and d_2 are said to be *equivalent* if, and only if, both metrics generate precisely the same open sets, i.e. if, and only if, for each $A \subseteq X$, the following holds:

$$A \text{ is } d_1\text{-open} \quad \Leftrightarrow \quad A \text{ is } d_2\text{-open}. \tag{1.89}$$

(b) Let $\|\cdot\|_1$ and $\|\cdot\|_2$ be norms on a vector space X over $\mathbb{K}$. Then $\|\cdot\|_1$ and $\|\cdot\|_2$ are said to be *equivalent* if, and only if, there exist positive constants $\alpha, \beta \in \mathbb{R}^+$ such that

$$\alpha\|x\|_1 \leq \|x\|_2 \leq \beta\|x\|_1 \quad \text{for each } x \in X. \tag{1.90}$$

Proposition 1.94. *Let $\|\cdot\|_1$ and $\|\cdot\|_2$ be norms on a vector space X over $\mathbb{K}$, and let d_1 and d_2 be the respective induced metrics on X. Then $\|\cdot\|_1$ and $\|\cdot\|_2$ are equivalent norms if, and only if, d_1 and d_2 are equivalent metrics.*

Proof. If $X = \{0\}$, then there is nothing to show. Thus, assume that there exists some $x \in X \setminus \{0\}$.

First, assume (1.90) holds, i.e. the norms are equivalent. If $A \subseteq X$ ís open with respect to d_1 and $x \in A$, then there exists $\epsilon > 0$ such that $B_{\epsilon,d_1}(x) \subseteq A$. Thus, for each $y \in B_{\delta,d_2}(x)$ satisfying $\delta := \alpha\epsilon$, one obtains

$$d_1(x,y) \leq \frac{1}{\alpha}\,\|x-y\|_2 < \frac{\delta}{\alpha} = \epsilon,$$

showing $B_{\delta,d_2}(x) \subseteq B_{\epsilon,d_1}(x) \subseteq A$ and that $A \subseteq X$ is d_2-open. Now assume $A \subseteq X$ to be open with respect to d_2. If $x \in A$, then there exists $\epsilon > 0$ such that $B_{\epsilon,d_2}(x) \subseteq A$. Then, for each $y \in B_{\delta,d_1}(x)$ with $\delta := \epsilon/\beta$, it holds that

$$d_2(x,y) \leq \beta\,\|x-y\|_1 < \beta\delta = \epsilon,$$

showing $B_{\delta,d_1}(x) \subseteq B_{\epsilon,d_2}(x) \subseteq A$. Hence, $A \subseteq X$ is d_1-open.

So far, we have proved that the validity of (1.90) implies that A is d_1-open if, and only if, A is d_2-open (i.e. d_1 and d_2 are equivalent).

Conversely, assume that the induced metrics d_1 and d_2 are equivalent. According to Def. 1.93(a), $0 \in X$ has to be a d_1-interior point of both the open d_1-ball $B_{1,d_1}(0)$ and the open d_2-ball $B_{1,d_2}(0)$. Moreover, 0 also has to be a d_2-interior point of both open balls. We claim that the set $M := \big\{\|x\|_2 : \|x\|_1 = 1\big\} \subseteq \mathbb{R}_0^+$ is bounded. Proceeding by contraposition, assume that M is unbounded (from above, as it is always bounded from below by 0). Then there exists a sequence $(x^k)_{k\in\mathbb{N}}$ such that $\|x^k\|_1 = 1$ for each $k \in \mathbb{N}$ and $\lim_{k\to\infty} \|x^k\|_2 = \infty$. Define $\eta^k := \|x^k\|_2$ and $y^k := x^k/\eta^k$ (note $\eta^k \neq 0$, since $\|x^k\|_1 = 1$). Then $\|y^k\|_2 = 1$ for each $k \in \mathbb{N}$. Moreover, $\|y^k\|_1 = 1/\eta^k$, showing $\lim_{k\to\infty} d_1(0, y^k) = \lim_{k\to\infty} \|y^k\|_1 = 0$. Thus, for each $\epsilon > 0$, $B_{\epsilon,d_1}(0)$ contains elements y^k with $\|y^k\|_2 = 1$, i.e. 0 is not a d_1-interior point of $B_{1,d_2}(0)$. Thus, if 0 is a d_1-interior point of $B_{1,d_2}(0)$, then M must be bounded. Letting

$$\beta := \sup\big\{\|x\|_2 : \|x\|_1 = 1\big\} \in \mathbb{R}^+$$

(indeed, $\beta > 0$, as $\|x\|_1 = 1$ implies $x \neq 0$ and $\|x\|_2 > 0$), one has

$$\forall_{x\in X\setminus\{0\}} \quad \|x\|_2 = \left\| \|x\|_1 \frac{x}{\|x\|_1} \right\|_2 \leq \beta \, \|x\|_1.$$

We have therefore found a constant $\beta > 0$ such that the corresponding part of (1.90) is satisfied. One can now proceed completely analogously to show that the hypothesis of 0 being a d_2-interior point of $B_{1,d_1}(0)$ implies that the set $\big\{\|x\|_1 : \|x\|_2 = 1\big\}$ is bounded and

$$\gamma := \sup\big\{\|x\|_1 : \|x\|_2 = 1\big\} \in \mathbb{R}^+$$

satisfies $\|x\|_1 \leq \gamma\|x\|_2$ for each $x \in X$. Finally, letting $\alpha := \gamma^{-1}$ completes the proof of the equivalence of $\|\cdot\|_1$ and $\|\cdot\|_2$. ■

Theorem 1.95. *All norms on $\mathbb{K}^n$, $n \in \mathbb{N}$, are equivalent.*

Proof. It suffices to show that every norm on $\mathbb{K}^n$ is equivalent to the 2-norm on $\mathbb{K}^n$. So let $\|\cdot\|_2$ denote the 2-norm on $\mathbb{K}^n$ and let $\|\cdot\|$ denote an arbitrary norm on $\mathbb{K}^n$. We recall the standard unit vectors e_j from (1.3) as well as that every $z \in \mathbb{K}^n$ can be written as $z = \sum_{j=1}^n z_j e_j$. Moreover, the 2-norm satisfies the Cauchy-Schwarz inequality (1.81), which can be exploited to get

$$\begin{aligned} \|z\| &= \left\| \sum_{j=1}^n z_j e_j \right\| \leq \sum_{j=1}^n |z_j| \|e_j\| = \big(|z_1|, \ldots, |z_n|\big) \cdot \big(\|e_1\|, \ldots, \|e_n\|\big) \\ &\overset{(1.81)}{\leq} \|z\|_2 \, \big\|\big(\|e_1\|, \ldots, \|e_n\|\big)\big\|_2, \end{aligned} \tag{1.91}$$

that means, with $\beta := \sqrt{\sum_{j=1}^n \|e_j\|^2} > 0$,

$$\|z\| \leq \beta \|z\|_2 \quad \text{for each } z \in \mathbb{K}^n. \tag{1.92}$$

We claim that there is also $\alpha > 0$ such that

$$\alpha\|z\|_2 \leq \|z\| \quad \text{for each } z \in \mathbb{K}^n. \tag{1.93}$$

Seeking a contradiction, assume that there is no $\alpha > 0$ satisfying (1.93). Then there is a sequence $(z^k)_{k\in\mathbb{N}}$ in $\mathbb{K}^n$ such that, for each $k \in \mathbb{N}$, $\frac{1}{k}\|z^k\|_2 > \|z^k\|$. Letting $w^k := z^k/\|z^k\|_2$, one gets $\frac{1}{k}\|w^k\|_2 > \|w^k\|$ and $\|w^k\|_2 = 1$ for each $k \in \mathbb{N}$. The Bolzano-Weierstrass Th. 1.16(b) yields a subsequence $(u^k)_{k\in\mathbb{N}}$ of $(w^k)_{k\in\mathbb{N}}$ that converges with respect to $\|\cdot\|_2$ to some $u \in \mathbb{K}^n$. As each norm is continuous according to Lem. 1.40(b), the convergence $u^k \to u$ with respect to $\|\cdot\|_2$ implies $\|u\|_2 = \lim_{k\to\infty}\|u^k\|_2 = 1$, and, in particular, $u \neq 0$. On the other hand, using (1.92), one has $\|u^k - u\| \leq \beta\|u^k - u\|_2 \to 0$, i.e. $(u^k)_{k\in\mathbb{N}}$ also converges to u with respect to $\|\cdot\|$. Then the continuity of $\|\cdot\|$ yields $\|u\| = \lim_{k\to\infty}\|u^k\| \leq \lim_{k\to\infty}\frac{1}{k}\|u^k\|_2 = \lim_{k\to\infty}\frac{1}{k} = 0$, i.e. $u = 0$ in a contradiction to $u \neq 0$. Thus, the assumption that there is no $\alpha > 0$ satisfying (1.93) must have been wrong, i.e. (1.93) must hold for some $\alpha > 0$. The proof is concluded by the observation that (1.92) together with (1.93) is precisely the statement that $\|\cdot\|_2$ and $\|\cdot\|$ are equivalent. ■

Caveat 1.96. Even though it follows from Th. 1.95 and Prop. 1.94 that all metrics on $\mathbb{K}^n$ induced by norms on $\mathbb{K}^n$ are equivalent, there exist nonequivalent metrics on $\mathbb{K}^n$ (examples?).

Proposition 1.97. *For metrics d_1 and d_2 on a set X, the following two statements are equivalent:*

(i) *d_1 and d_2 are equivalent.*

(ii) *Every sequence $(x^k)_{k\in\mathbb{N}}$ in X converges with respect to d_1 if, and only if, it converges with respect to d_2.*

In consequence, the analogous result also holds for two norms on a real vector space.

Proof. "(i) ⇒ (ii)": Suppose d_1 and d_2 are equivalent. Suppose $(x^k)_{k\in\mathbb{N}}$ converges to $x \in X$ with respect to d_1. Let $\epsilon > 0$. Since $B_{\epsilon,d_2}(x)$ is d_2-open and since d_1 and d_2 are equivalent, $B_{\epsilon,d_2}(x)$ is also d_1-open. Thus, there is $\delta > 0$ such that $B_{\delta,d_1}(x) \subseteq B_{\epsilon,d_2}(x)$. By Lem. 1.37, there is $N \in \mathbb{N}$ such that, for each $k > N$, $x^k \in B_{\delta,d_1}(x) \subseteq B_{\epsilon,d_2}(x)$. Thus, again by Lem. 1.37, $(x^k)_{k\in\mathbb{N}}$ converges to $x \in X$ with respect to d_2. An analogous argument shows that, if $(x^k)_{k\in\mathbb{N}}$ converges to $x \in X$ with respect to d_2, then $(x^k)_{k\in\mathbb{N}}$ converges to $x \in X$ with respect to d_1.

"(ii) ⇒ (i)": Suppose $O \subseteq X$ is d_1-open. If $x \in O$ and x, then, due to Lem. 1.31, x is in the d_2-interior of O, x is in the d_2-boundary of O, or x is in the d_2-interior of $X \setminus O$. The last case can not occur, as x is not in $X \setminus O$. We also need to exclude the case that x is in the d_2-boundary of O: Suppose it were. Then, for each $k \in \mathbb{N}$, there is $x^k \in B_{\frac{1}{k},d_2}(x) \cap (X \setminus O)$. Thus, $d_2(x^k, x) \to 0$ for $k \to \infty$. Then, by the hypothesis, also $d_1(x^k, x) \to 0$. As x is a d_1-interior point of O, there is $\epsilon > 0$ such that $B_{\epsilon,d_1}(x) \subseteq O$. Then $d_1(x^k, x) \to 0$ implies that $x^k \in B_{\epsilon,d_1}(x) \subseteq O$ for sufficiently large

k, in contradiction to $x^k \in X \setminus O$. This contradiction excludes the case that x is in the d_2-boundary of O. It only remains that x is a d_2-interior point of O. Since x was arbitrary, O is d_2-open. Interchanging the roles of d_1 and d_2 in the previous argument, one sees that each d_2-open set is also d_1-open, completing the proof that d_1 and d_2 are equivalent. ■

The following Ex. 1.98 shows that, in general, there can be norms on a real vector space X that are not equivalent.

Example 1.98. As in Examples 1.39(b) and 1.68(a) before, let X be vector space over $\mathbb{K}$, consisting of the sequences in $\mathbb{K}$ that are finally constant and equal to zero. Then

$$\big\|(z_n)_{n\in\mathbb{N}}\big\|_1 := \sum_{n=1}^{\infty} |z_n| \quad \text{and} \tag{1.94a}$$

$$\big\|(z_n)_{n\in\mathbb{N}}\big\|_{\sup} := \max\big\{|z_n| : n \in \mathbb{N}\big\} \tag{1.94b}$$

define norms on X ($\|\cdot\|_{\sup}$ is the same norm that was considered in the earlier examples). As was already observed in Example 1.68(a), the sequence $(z^k)_{k\in\mathbb{N}}$ in X defined by

$$z_n^k := \begin{cases} 1/k & \text{for } 1 \le n \le k, \\ 0 & \text{for } n > k, \end{cases} \tag{1.95}$$

converges to $(0, 0, \dots) \in X$ with respect to $\|\cdot\|_{\sup}$; however, the sequence does not converge in X with respect to $\|\cdot\|_1$ (exercise). Then Prop. 1.97 implies that $\|\cdot\|_1$ and $\|\cdot\|_{\sup}$ are not equivalent.

2 Differential Calculus in $\mathbb{R}^n$

2.1 Partial Derivatives and Gradients

The goal of the following is to generalize the notion of derivative from one-dimensional functions to functions $f : G \longrightarrow \mathbb{K}$, where $G \subseteq \mathbb{R}^n$ with $n \in \mathbb{N}$. Later we will also allow functions with values in $\mathbb{K}^m$. For $\xi \in G$, $G \subseteq \mathbb{R}^n$, we will define a function $f : G \longrightarrow \mathbb{K}$ to have a so-called partial derivative (or just partial for short) at ξ with respect to the variable x_j if, and only if, the one-dimensional function that results from keeping all but the jth variable fixed, namely

$$x_j \mapsto \phi(x_j) := f(\xi_1, \dots, \xi_{j-1}, x_j, \xi_{j+1}, \dots, \xi_n), \tag{2.1}$$

is differentiable at $x_j = \xi_j$ in the usual sense for one-dimensional functions. The partial derivative of f at ξ with respect to x_j is then identified with $\phi'(\xi_j)$. This leads to the following definition:

Definition 2.1. Let $G \subseteq \mathbb{R}^n$, $n \in \mathbb{N}$, $f : G \longrightarrow \mathbb{K}$, $\xi \in G$, $j \in \{1, \ldots, n\}$. If there is $\epsilon > 0$ such that $\xi + he_j \in G$ for each $h \in]-\epsilon, \epsilon[$ (this condition is trivially satisfied if ξ is an interior point of G), then f is said to have a *partial derivative* at ξ with respect to the variable x_j (or a jth partial for short) if, and only if, the limit

$$\lim_{h \to 0} \frac{f(\xi + he_j) - f(\xi)}{h} \qquad \Big(\, 0 \neq h \in]-\epsilon, \epsilon[\, \Big) \tag{2.2}$$

exists in $\mathbb{K}$. In that case, the limit is defined to the jth partial of f at ξ and it is denoted with one of the symbols

$$\partial_j f(\xi), \ \partial_{x_j} f(\xi), \ \frac{\partial f(\xi)}{\partial x_j}, \ f_{x_j}(\xi), \ D_j f(\xi). \tag{2.3}$$

If ξ is a boundary point of G and there is $\epsilon > 0$ such that, for each $h \in]0, \epsilon[$, $\xi + he_j \in G$ and $\xi - he_j \notin G$ (resp. $\xi - he_j \in G$ and $\xi + he_j \notin G$), then, instead of the limit in (2.2), one uses the one-sided limit

$$\lim_{h \downarrow 0} \frac{f(\xi + he_j) - f(\xi)}{h} \qquad \Big(\text{ resp. } \lim_{h \uparrow 0} \frac{f(\xi + he_j) - f(\xi)}{h} \Big) \tag{2.4}$$

in the above definition of the jth partial at ξ. If all the partials of f exist in ξ, then the vector

$$\nabla f(\xi) := \big(\partial_1 f(\xi), \ldots, \partial_n f(\xi) \big) \tag{2.5}$$

is called the *gradient* of f at ξ (the symbol ∇ is called *nabla*, the corresponding operator is sometimes called *del*). It is customary to consider the gradient as a row vector. If the jth partial $\partial_j f(\xi)$ exists for each $\xi \in G$, then the function

$$\partial_j f : G \longrightarrow \mathbb{K}, \quad \xi \mapsto \partial_j f(\xi), \tag{2.6}$$

is also called the jth partial of f.

Example 2.2. The following example shows that, in general, the existence of partial derivatives does not imply continuity: Consider the function

$$f : \mathbb{R}^2 \longrightarrow \mathbb{R}, \quad f(x,y) := \begin{cases} \frac{xy}{x^2+y^2} & \text{for } (x,y) \neq (0,0), \\ 0 & \text{for } (x,y) = (0,0). \end{cases} \tag{2.7}$$

Using the quotient rule for $(x,y) \neq (0,0)$ and the fact that $f(x,0) = f(0,y) = 0$ for all $(x,y) \in \mathbb{R}^2$, one obtains

$$\nabla f : \mathbb{R}^2 \longrightarrow \mathbb{R}^2, \quad \nabla f(x,y) = \begin{cases} \left(\frac{y(y^2-x^2)}{(x^2+y^2)^2}, \ \frac{x(x^2-y^2)}{(x^2+y^2)^2} \right) & \text{for } (x,y) \neq (0,0), \\ (0,0) & \text{for } (x,y) = (0,0). \end{cases} \tag{2.8}$$

In particular, both partials $\partial_x f$ and $\partial_y f$ exist everywhere in $\mathbb{R}^2$. However, f is not continuous in $(0,0)$: For $k \in \mathbb{N}$, let $x_k := (1/k)$, $y_k := (1/k)$. Then $\lim_{k \to \infty}(x_k, y_k) = (0,0)$, but

$$f(x_k, y_k) = \frac{\frac{1}{k^2}}{\frac{1}{k^2} + \frac{1}{k^2}} = \frac{1}{2} \tag{2.9}$$

for each $k \in \mathbb{N}$. In particular, $\lim_{k\to\infty} f(x_k, y_k) = \frac{1}{2} \neq 0 = f(0,0)$, showing that f is not continuous in $(0,0)$.

—

The problem in Example 2.2 is the discontinuity of the partials in $(0,0)$. The following Prop. 2.3 shows that, if all partials of f exist and are continuous in some neighborhood of a point ξ, then f is continuous in ξ.

Proposition 2.3. *Let $G \subseteq \mathbb{R}^n$, $n \in \mathbb{N}$, $f : G \longrightarrow \mathbb{K}$, and let ξ be an interior point of G. If there is $\epsilon > 0$ such that $\partial_j f(x)$ exists for each $x \in B_\epsilon(\xi)$ and for each $j \in \{1, \ldots, n\}$ and such that each function $\partial_j f : B_\epsilon(\xi) \longrightarrow \mathbb{K}$ is continuous in ξ, then f is continuous in ξ.*

Proof. According to Th. 1.58, it suffices to show that both $\operatorname{Re} f$ and $\operatorname{Im} f$ are continuous in ξ. Since, moreover, the continuity of $\partial_j f(x)$ on $B_\epsilon(\xi)$ implies the continuity of both $\operatorname{Re} \partial_j f(x) = \partial_j \operatorname{Re} f(x)$ and $\operatorname{Im} \partial_j f(x) = \partial_j \operatorname{Im} f(x)$ on $B_\epsilon(\xi)$, it suffices to carry out the proof for $\mathbb{K} = \mathbb{R}$. Thus, for the remainder of the proof, assume $\mathbb{K} = \mathbb{R}$.

Since all norms on $\mathbb{R}^n$ are equivalent (i.e. they all generate precisely the same open sets), it suffices to carry out the proof for $\mathbb{R}^n$ with the 1-norm $\|\cdot\|_1$. The continuity of the $\partial_j f$ at ξ implies that there is some $\delta \leq \epsilon$ such that $x \in B_\delta(\xi)$ implies $|\partial_j f(\xi) - \partial_j f(x)| < 1$ for each $j \in \{1, \ldots, n\}$. Thus, if $\tilde{L} := \max\{|\partial_1 f(\xi)|, \ldots, |\partial_n f(\xi)|\}$ and $x \in B_\delta(\xi)$, then

$$|\partial_j f(x)| \leq |\partial_j f(\xi)| + |\partial_j f(\xi) - \partial_j f(x)| < \tilde{L} + 1 =: L \quad \text{for each } j \in \{1, \ldots, n\}. \tag{2.10}$$

In other words, the partials are bounded in $B_\delta(\xi)$. Let $h \in \mathbb{R}^n$ with $\|h\|_1 < \delta$. Then

$$\begin{aligned}
&f(\xi + h) - f(\xi) \\
&= f(\xi_1 + h_1, \ldots, \xi_{n-1} + h_{n-1}, \xi_n + h_n) - f(\xi_1 + h_1, \ldots, \xi_{n-1} + h_{n-1}, \xi_n) \\
&\quad + f(\xi_1 + h_1, \ldots, \xi_{n-1} + h_{n-1}, \xi_n) - f(\xi_1 + h_1, \ldots, \xi_{n-1}, \xi_n) \\
&\quad + - \cdots + f(\xi_1 + h_1, \xi_2, \ldots, \xi_n) - f(\xi_1, \xi_2, \ldots, \xi_n) \\
&= f(\xi + h) - f\left(\xi + \sum_{k=1}^{n-1} h_k e_k\right) \\
&\qquad + f\left(\xi + \sum_{k=1}^{n-1} h_k e_k\right) - f\left(\xi + \sum_{k=1}^{n-2} h_k e_k\right) + - \cdots + f(\xi + h_1 e_1) - f(\xi) \\
&= \sum_{j=0}^{n-1} f\left(\xi + \sum_{k=1}^{n-j} h_k e_k\right) - f\left(\xi + \sum_{k=1}^{n-(j+1)} h_k e_k\right) \\
&= \sum_{j=0}^{n-1} \phi_j(h_{n-j}) - \phi_j(0),
\end{aligned} \tag{2.11}$$

where, for each $j \in \{0, \ldots, n-1\}$,

$$\phi_j : [0, h_{n-j}] \longrightarrow \mathbb{R}, \quad \phi_j(t) := f\left(\xi + t e_{n-j} + \sum_{k=1}^{n-(j+1)} h_k e_k\right). \tag{2.12}$$

If $h_{n-j} = 0$, then set $\theta_j := 0$. Otherwise, apply the one-dimensional mean value theorem [Phi15a, Th. 9.17] to the one-dimensional function ϕ_j to get numbers $\theta_j \in]0, h_{n-j}[$ such that

$$\phi_j(h_{n-j}) - \phi_j(0) = h_{n-j}\,\phi_j'(\theta_j) = h_{n-j}\,\partial_{n-j} f\left(\xi + \theta_j e_{n-j} + \sum_{k=1}^{n-(j+1)} h_k e_k\right). \tag{2.13}$$

Combining (2.11) with (2.13) and (2.10) yields

$$\begin{aligned}\big|f(\xi+h) - f(\xi)\big| &\leq \sum_{j=0}^{n-1} \left| h_{n-j}\,\partial_{n-j} f\left(\xi + \theta_j e_{n-j} + \sum_{k=1}^{n-(j+1)} h_k e_k\right)\right| \\ &\leq L \sum_{j=0}^{n-1} |h_{n-j}| = L\|h\|_1. \end{aligned} \tag{2.14}$$

Thus, if $(x^l)_{l\in\mathbb{N}}$ is a sequence in $B_\delta(\xi)$ such that $\lim_{l\to\infty} x^l = \xi$, then, with $h^l := x^l - \xi$, one has $\lim_{l\to\infty} \|h^l\|_1 = 0$ (note that this holds for the 1-norm as well as for all other norms on $\mathbb{R}^n$, since all norms on $\mathbb{R}^n$ are equivalent). Using (2.14) with h^l instead of h, one gets $|f(x^l) - f(\xi)| = |f(\xi + h^l) - f(\xi)| \leq L\|h^l\|_1$, showing $\lim_{l\to\infty} f(x^l) = f(\xi)$, i.e. the continuity of f at ξ. ■

2.2 The Jacobian

If $f : G \longrightarrow \mathbb{K}^m$, where $G \subseteq \mathbb{R}^n$, then we can compute partials for each of the coordinate functions f_j of f (provided the partials exist).

Definition 2.4. Let $G \subseteq \mathbb{R}^n$, $f : G \longrightarrow \mathbb{K}^m$, $(n, m) \in \mathbb{N}^2$, $\xi \in G$. If, for each $l \in \{1, \ldots, m\}$, the coordinate function $f_l = \pi_l \circ f$ (recall that $f = (f_1, \ldots, f_m)$) has all partials $\partial_k f_l$ at ξ, then these $m \cdot n$ partials form an $m \times n$ matrix, namely

$$J_f(\xi) := \frac{\partial(f_1, \ldots, f_m)}{\partial(x_1, \ldots, x_n)}(\xi) := \begin{pmatrix} \partial_1 f_1(\xi) & \ldots & \partial_n f_1(\xi) \\ \vdots & & \vdots \\ \partial_1 f_m(\xi) & \ldots & \partial_n f_m(\xi) \end{pmatrix} = \begin{pmatrix} \nabla f_1(\xi) \\ \vdots \\ \nabla f_m(\xi) \end{pmatrix}, \tag{2.15}$$

called the *Jacobian matrix* of f at ξ. In the case that $m = n$, the Jacobian matrix $J_f(\xi)$ is quadratic and one can compute its determinant $\det J_f(\xi)$. This determinant is then called the *Jacobian determinant* of f at ξ.

Both the Jacobi matrix and the Jacobi determinant are sometimes referred to as the *Jacobian*. One then has to determine from the context which of the two is meant.

Remark 2.5. In many situations, it does not matter if you interpret $z \in \mathbb{K}^n$ as a column vector or a row vector, and the same is true for the gradient. However, in the context of matrix multiplications, it is important to work with a consistent interpretation of such vectors. We will therefore adhere to the following agreement: In the context of matrix

multiplications, we always interpret $x \in \mathbb{R}^n$ and $f(x) \in \mathbb{K}^m$ for $\mathbb{K}^m$-valued functions f as *column vectors*, whereas we always interpret the gradients $\nabla g(x)$ of $\mathbb{K}$-valued functions g as *row vectors*.

Example 2.6. (a) Let A be an $m \times n$ matrix over $\mathbb{K}$,

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

Then the map $x \mapsto Ax$, $A : \mathbb{R}^n \longrightarrow \mathbb{K}^m$, is $\mathbb{R}$-linear for $\mathbb{K} = \mathbb{R}$, and it is the restriction to $\mathbb{R}^n$ of the $\mathbb{C}$-linear map A on $\mathbb{C}^n$ for $\mathbb{K} = \mathbb{C}$ (note that, due to the agreement from Rem. 2.5, Ax can be interpreted as a matrix multiplication in the usual way). Thus, if we denote the coordinate functions $\pi_l \circ A$ by A_l, $l \in \{1, \dots, m\}$, then $A_l(x) = \sum_{k=1}^n a_{lk} x_k$ and $\partial_k A_l(x) = \frac{\partial A_l(x)}{\partial x_k} = a_{lk}$. Thus, $J_A(x) = A$ for each $x \in \mathbb{R}^n$.

(b) Consider $(f, g) : \mathbb{R}^3 \longrightarrow \mathbb{C}^2$, $(f(x, y, z), g(x, y, z)) := (z^2 e^{ixy}, ye^{2yz} + ix)$. Then one computes the following Jacobian:

$$J_{(f,g)}(x, y, z) = \begin{pmatrix} \nabla f(x, y, z) \\ \nabla g(x, y, z) \end{pmatrix} = \begin{pmatrix} iyz^2 e^{ixy} & ixz^2 e^{ixy} & 2ze^{ixy} \\ i & (1 + 2yz)e^{2yz} & 2y^2 e^{2yz} \end{pmatrix}. \tag{2.16}$$

(c) Consider $(f, g) : \mathbb{R}^2 \longrightarrow \mathbb{C}^2$, $(f(x, y), g(x, y)) := (e^{ixy}, ixye^{2y})$. Then one computes the following Jacobian determinant:

$$\begin{aligned} \det J_{(f,g)}(x, y) &= \begin{vmatrix} iye^{ixy} & ixe^{ixy} \\ iye^{2y} & i(x + 2xy)e^{2y} \end{vmatrix} \\ &= -ye^{ixy}(x + 2xy)e^{2y} + ye^{2y} xe^{ixy} = -2xy^2 e^{(2+ix)y}. \end{aligned} \tag{2.17}$$

Remark 2.7. The linearity of forming the derivative of one-dimensional functions directly implies the linearity of forming partial derivatives, gradients, and Jacobians (provided they exist). More precisely, if $G \subseteq \mathbb{R}^n$, $f, g : G \longrightarrow \mathbb{K}^m$, $(n, m) \in \mathbb{N}^2$, $\xi \in G$, and $\lambda \in \mathbb{K}$, then, for each $(l, k) \in \{1, \dots, m\} \times \{1, \dots, n\}$,

$$\partial_k (f + g)_l(\xi) = \partial_k f_l(\xi) + \partial_k g_l(\xi), \qquad \partial_k (\lambda f)_l(\xi) = \lambda \partial_k f_l(\xi), \tag{2.18a}$$

$$\nabla (f + g)_l(\xi) = \nabla f_l(\xi) + \nabla g_l(\xi), \qquad \nabla (\lambda f)_l(\xi) = \lambda \nabla f_l(\xi), \tag{2.18b}$$

$$J_{f+g}(\xi) = J_f(\xi) + J_g(\xi), \qquad J_{\lambda f}(\xi) = \lambda J_f(\xi), \tag{2.18c}$$

where, in each case, the assumed existence of the objects on the right-hand side of the equation implies the existence of the object on the left-hand side.

2.3 Higher Order Partials and the Spaces C^k

Partial derivatives can, in turn, have partial derivatives themselves and so on. For example, a function $f : \mathbb{R}^3 \longrightarrow \mathbb{K}$ might have the following partial derivative of 6th

order: $\partial_1\partial_3\partial_2\partial_1\partial_2\partial_2 f$. We will see that, in general, it is important in which order the different partial derivatives are carried out (see Example 2.9). If all partial derivatives are continuous, then the situation is much better and the result is the same, no matter what order is used for the partial derivatives (continuous partials commute, see Th. 2.12). We start with the definition of higher order partials:

Definition 2.8. Let $G \subseteq \mathbb{R}^n$, $f : G \longrightarrow \mathbb{K}$, $\xi \in G$. Fix $k \in \mathbb{N}$. For each element $p = (p_1, \ldots, p_k) \in \{1, \ldots, n\}^k$, define the following partial derivative of kth order provided that it exists:

$$\partial_p f(\xi) := \frac{\partial^k f(\xi)}{\partial x_{p_1} \ldots \partial x_{p_k}} := \partial_{p_1} \ldots \partial_{p_k} f(\xi). \tag{2.19}$$

One also defines f itself to be its own partial derivative of order 0. Analogous to Def. 2.4, if $f : G \longrightarrow \mathbb{K}^m$, $m \in \mathbb{N}$, then one defines the higher order partials for each coordinate function f_l, $l = 1, \ldots, m$, i.e. one uses f_l instead of f in (2.19).

Example 2.9. The following example shows that, in general, partial derivatives do not commute: Consider the function

$$f : \mathbb{R}^2 \longrightarrow \mathbb{R}, \quad f(x,y) := \begin{cases} \frac{xy^3}{x^2+y^2} & \text{for } (x,y) \neq (0,0), \\ 0 & \text{for } (x,y) = (0,0). \end{cases} \tag{2.20}$$

Analogous to Example 2.2, using the quotient rule for $(x,y) \neq (0,0)$ and the fact that $f(x,0) = f(0,y) = 0$ for all $(x,y) \in \mathbb{R}^2$, one obtains

$$\begin{aligned} \nabla f : \mathbb{R}^2 \longrightarrow \mathbb{R}^2, \quad \nabla f(x,y) &= \big(\partial_1 f(x,y), \partial_2 f(x,y)\big) = \big(\partial_x f(x,y), \partial_y f(x,y)\big) \\ &= \begin{cases} \left(\frac{y^3(y^2-x^2)}{(x^2+y^2)^2}, \ \frac{xy^2(3x^2+y^2)}{(x^2+y^2)^2}\right) & \text{for } (x,y) \neq (0,0), \\ (0,0) & \text{for } (x,y) = (0,0). \end{cases} \end{aligned} \tag{2.21}$$

In particular, we have $\partial_1 f(0,y) = \partial_x f(0,y) = y$ for each $y \in \mathbb{R}$ and $\partial_2 f(x,0) = \partial_y f(x,0) = 0$ for each $x \in \mathbb{R}$. Thus, $\partial_y\partial_x f(0,y) \equiv 1$ and $\partial_x\partial_y f(x,0) \equiv 0$. Evaluating at $(0,0)$ yields $\partial_2\partial_1 f(0,0) = \partial_y\partial_x f(0,0) = 1 \neq 0 = \partial_x\partial_y f(0,0) = \partial_1\partial_2 f(0,0)$.

—

As in Example 2.2, the problem in Example 2.9 lies in the discontinuity of the partials in $(0,0)$. As mentioned above, if all partials are continuous, then they do commute. To prove this result is our next goal. We will accomplish this in several steps. We start with a preparatory lemma that provides a variant of the mean value theorem in two dimensions.

Lemma 2.10. *Let $a < \tilde{a}$, $b < \tilde{b}$, and consider the square $I = [a, \tilde{a}] \times [b, \tilde{b}]$ (which constitutes a closed interval in $\mathbb{R}^2$). Suppose $f : I \longrightarrow \mathbb{R}$, $(x,y) \mapsto f(x,y)$, and set*

$$\Delta_I(f) := f(\tilde{a}, \tilde{b}) + f(a,b) - f(a, \tilde{b}) - f(\tilde{a}, b). \tag{2.22}$$

(a) *If $\partial_x f$ and $\partial_y\partial_x f$ exist everywhere in I, then there is some point $(\xi, \eta) \in I^\circ$ (i.e. with $a < \xi < \tilde{a}$ and $b < \eta < \tilde{b}$) satisfying*

$$\Delta_I(f) = (\tilde{a} - a)(\tilde{b} - b)\partial_y\partial_x f(\xi, \eta).$$

(b) *If $\partial_y f$ and $\partial_x\partial_y f$ exist everywhere in I, then there is some point $(\xi,\eta)\in I^\circ$ satisfying*

$$\Delta_I(f) = (\tilde{a}-a)(\tilde{b}-b)\partial_x\partial_y f(\xi,\eta).$$

Proof. We prove (a); the proof of (b) is completely analogous. Since the function $g : [a,\tilde{a}] \longrightarrow \mathbb{R}$, $g(x) := f(x,\tilde{b}) - f(x,b)$, is differentiable, the one-dimensional mean value theorem [Phi15a, Th. 9.17] yields the existence of some $\xi \in]a,\tilde{a}[$ satisfying

$$\Delta_I(f) = g(\tilde{a}) - g(a) = (\tilde{a}-a)g'(\xi) = (\tilde{a}-a)\big(\partial_x f(\xi,\tilde{b}) - \partial_x f(\xi,b)\big). \tag{2.23a}$$

Since the function $G : [b,\tilde{b}] \longrightarrow \mathbb{R}$, $G(y) := \partial_x f(\xi,y)$, is differentiable, the one-dimensional mean value theorem [Phi15a, Th. 9.17] yields the existence of some $\eta \in]b,\tilde{b}[$ satisfying

$$\partial_x f(\xi,\tilde{b}) - \partial_x f(\xi,b) = G(\tilde{b}) - G(b) = (\tilde{b}-b)G'(\eta) = (\tilde{b}-b)\partial_y\partial_x f(\xi,\eta). \tag{2.23b}$$

Combining (2.23a) and (2.23b) proves (a). ■

Results like the following Th. 2.11 are often named after H.A. Schwarz. His result was actually stronger than our Th. 2.11 in the sense that he showed that partials commute even under weaker hypotheses. However, Th. 2.11 will suffice for our purposes.

Theorem 2.11. *Let G be an open subset of $\mathbb{R}^2$. Suppose that $f : G \longrightarrow \mathbb{K}$, $(x,y)\mapsto f(x,y)$, has partial derivatives $\partial_x f$, $\partial_y f$, $\partial_y\partial_x f$, and $\partial_x\partial_y f$ everywhere in G. If $\partial_y\partial_x f$ and $\partial_x\partial_y f$ are continuous on G, then $\partial_y\partial_x f = \partial_x\partial_y f$ (in particular, $\partial_y\partial_x f = \partial_x\partial_y f$ if all the functions f, $\partial_x f$, $\partial_y f$, $\partial_y\partial_x f$, and $\partial_x\partial_y f$ are continuous).*

Proof. We first note that it suffices to prove the theorem for $\mathbb{K} = \mathbb{R}$, as one can then apply the result to both $\operatorname{Re} f$ and $\operatorname{Im} f$ to obtain the case $\mathbb{K} = \mathbb{C}$. Thus, for the remainder of the proof, we assume f to be $\mathbb{R}$-valued. Consider an arbitrary point $(a,b) \in G$. Since G is open, there is $N \in \mathbb{N}$ such that, for each $n > N$, the closed interval $I_n := [a, a+\frac{1}{n}] \times [b, b+\frac{1}{n}]$ is contained in G, $I_n \subseteq G$. Using Lem. 2.10(a),(b), for each $n > N$, we get $\xi_n, \xi_n' \in]a, a+\frac{1}{n}[$ and $\eta_n, \eta_n' \in]b, b+\frac{1}{n}[$ such that

$$n^2\Delta_{I_n}(f) = \partial_y\partial_x f(\xi_n,\eta_n) = \partial_x\partial_y f(\xi_n',\eta_n'). \tag{2.24}$$

It follows from $\lim_{n\to\infty}(a+\frac{1}{n}) = a$ and $\lim_{n\to\infty}(b+\frac{1}{n}) = b$ that $\lim_{n\to\infty}(\xi_n,\eta_n) = (a,b)$ and that $\lim_{n\to\infty}(\xi_n',\eta_n') = (a,b)$. Thus, the continuity of $\partial_y\partial_x f$ and $\partial_x\partial_y f$ in (a,b) together with Th. 1.55(b) yields

$$\partial_y\partial_x f(a,b) = \lim_{n\to\infty} \partial_y\partial_x f(\xi_n,\eta_n) \overset{(2.24)}{=} \lim_{n\to\infty} \partial_x\partial_y f(\xi_n',\eta_n') = \partial_x\partial_y f(a,b), \tag{2.25}$$

finishing the proof that $\partial_y\partial_x f$ and $\partial_x\partial_y f$ commute in G. ■

Using the combinatorial result that one can achieve an arbitrary permutation by a finite sequence of permutations of precisely two juxtaposed elements (cf. Th. A.42(b)) one can easily extend Th. 2.11 to partial derivatives of order $k > 2$.

Theorem 2.12. *Let G be an open subset of $\mathbb{R}^n$, $n \in \mathbb{N}$, and let $k \in \mathbb{N}$. Suppose that for $f : G \longrightarrow \mathbb{K}$ all partial derivatives of order less than or equal to k exist and are continuous on G. Than the value of each partial derivative of f of order k is independent of the order in which the individual partial derivatives are carried out. In other words, if $p = (p_1, \ldots, p_k) \in \{1, \ldots, n\}^k$ and $q = (q_1, \ldots, q_k) \in \{1, \ldots, n\}^k$ such that there exists a permutation (i.e. a bijective map) $\pi : \{1, \ldots, k\} \longrightarrow \{1, \ldots, k\}$ satisfying $q = (p_{\pi(1)}, \ldots, p_{\pi(k)})$, then $\partial_p f(\xi) = \partial_q f(\xi)$ for each $\xi \in G$. If $f : G \longrightarrow \mathbb{K}^m$, $m \in \mathbb{N}$, then the same holds with respect to each coordinate function f_j of f, $j \in \{1, \ldots, m\}$.*

Proof. For $k = 1$, there is nothing to prove. So let $k > 1$. For $l \in 1, \ldots, k-1$, let $\tau_l : \{1, \ldots, k\} \longrightarrow \{1, \ldots, k\}$ be the transposition that interchanges l and $l+1$ and leaves all other elements fixed (i.e. $\tau_l(l) = l+1$, $\tau_l(l+1) = l$, $\tau(\alpha) = \alpha$ for each $\alpha \in \{1, \ldots, k\} \setminus \{l, l+1\}$) and let $T := \{\tau_1, \ldots, \tau_{k-1}\}$. Then Th. 2.11 directly implies that the theorem holds for $\pi = \tau$ for each $\tau \in T$. For a general permutation $\pi : \{1, \ldots, k\} \longrightarrow \{1, \ldots, k\}$, the abovementioned combinatorial result provides a finite sequence $(\tau^1, \ldots, \tau^N)$, $N \in \mathbb{N}$, of elements of T such that $\pi = \tau^N \circ \cdots \circ \tau^1$. Thus, as we already know that the theorem holds for $N = 1$, the case $N > 1$ follows by induction. ■

Now that we have seen that functions with continuous partials are particularly benign, we introduce some special notation dedicated to such functions:

Definition 2.13. Let $G \subseteq \mathbb{R}^n$, $f : G \longrightarrow \mathbb{K}$, $k \in \mathbb{N}_0$. If all partials of f up to order k exist everywhere in G, and if f and all its partials up to order k are continuous on G, then f is said to be *of class* C^k (one also says that f has continuous partials up to order k). The set of all $\mathbb{K}$-valued functions of class C^k is denoted by $C^k(G, \mathbb{K})$ (in particular, $C^0(G, \mathbb{K}) = C(G, \mathbb{K})$). If f has continuous partials of all orders, than f is said to be of class C^∞, i.e. $C^\infty(G, \mathbb{K}) := \bigcap_{k=0}^\infty C^k(G, \mathbb{K})$. For $\mathbb{R}$-valued functions, we introduce the shorter notation $C^k(G) := C^k(G, \mathbb{R})$ for each $k \in \mathbb{N}_0 \cup \{\infty\}$. Finally, for $f : G \longrightarrow \mathbb{K}^m$, we say that f is of class C^k if, and only if, each coordinate function f_j, $j \in \{1, \ldots, m\}$, is of class C^k. The set of all such functions is denoted by $C^k(G, \mathbb{K}^m)$.

Notation 2.14. For two vectors $u = (u_1, u_2, u_3) \in \mathbb{K}^3$, $v = (v_1, v_2, v_3) \in \mathbb{K}^3$, the cross product is an element of $\mathbb{K}^3$ defined as follows:

$$u \times v := \Big(u_2 v_3 - u_3 v_2,\ u_3 v_1 - u_1 v_3,\ u_1 v_2 - u_2 v_1\Big). \tag{2.26}$$

Definition 2.15. Let $G \subseteq \mathbb{R}^n$, $n \in \mathbb{N}$, $\xi \in G$.

(a) If $f : G \longrightarrow \mathbb{K}^n$ and the partials $\partial_j f_j(\xi)$ exist for each $j \in \{1, \ldots, n\}$, then the *divergence* of f in ξ is defined as

$$\operatorname{div} f(\xi) := \sum_{j=1}^n \partial_j f_j(\xi) = \frac{\partial f_1(\xi)}{\partial x_1} + \cdots + \frac{\partial f_n(\xi)}{\partial x_n}. \tag{2.27}$$

If $\operatorname{div} f(\xi)$ exists for all $\xi \in G$, then $\operatorname{div} f : G \longrightarrow \mathbb{K}$. Sometimes, one defines the del operator $\nabla = (\partial_1, \ldots, \partial_n)$ and then writes $\operatorname{div} f = \nabla \cdot f$, using the analogue between

(2.27) and the definition of the Euclidean scalar product. Also note that $\operatorname{div} f(\xi)$ is precisely the trace of the corresponding Jacobi matrix, $\operatorname{div} f(\xi) = \operatorname{tr} J_f(\xi)$.

(b) If $f : G \longrightarrow \mathbb{K}$ has second-order partials at ξ, then one defines the *Laplacian* (also known as the *Laplace operator*) of f in ξ by

$$\Delta f(\xi) := \operatorname{div} \nabla f(\xi) = \sum_{j=1}^{n} \partial_j \partial_j f(\xi) = \partial_1^2 f(\xi) + \cdots + \partial_n^2 f(\xi). \tag{2.28}$$

If $\Delta f(\xi)$ exists for all $\xi \in G$, then $\Delta f : G \longrightarrow \mathbb{K}$.

(c) If $n = 3$ and $f : G \longrightarrow \mathbb{K}^3$ has first-order partials at ξ, then one defines the *curl* of f in ξ by

$$\begin{aligned} \operatorname{curl} f(\xi) &:= \big(\partial_2 f_3(\xi) - \partial_3 f_2(\xi),\, \partial_3 f_1(\xi) - \partial_1 f_3(\xi),\, \partial_1 f_2(\xi) - \partial_2 f_1(\xi)\big) \\ &= \left(\frac{\partial f_3(\xi)}{\partial x_2} - \frac{\partial f_2(\xi)}{\partial x_3}, \frac{\partial f_1(\xi)}{\partial x_3} - \frac{\partial f_3(\xi)}{\partial x_1}, \frac{\partial f_2(\xi)}{\partial x_1} - \frac{\partial f_1(\xi)}{\partial x_2} \right). \end{aligned} \tag{2.29}$$

If $\operatorname{curl} f(\xi)$ exists for all $\xi \in G$, then $\operatorname{curl} f : G \longrightarrow \mathbb{K}^3$. Again, one sometimes defines the del operator $\nabla = (\partial_1, \partial_2, \partial_3)$ and then writes $\operatorname{curl} f = \nabla \times f$, using the analogue between (2.29) and the definition of the cross product or two vectors in $\mathbb{K}^3$.

Proposition 2.16. *Let $G \subseteq \mathbb{R}^3$, let $f : G \longrightarrow \mathbb{K}$ be a scalar-valued function and let $v : G \longrightarrow \mathbb{K}^3$ be a vector-valued function.*

(a) *If $\xi \in G$ is such that f and v have all partials of first order at ξ, then*

$$\operatorname{curl}(fv)(\xi) = f(\xi) \operatorname{curl} v(\xi) + \nabla f(\xi) \times v(\xi).$$

(b) *If G is open and $f \in C^2(G, \mathbb{K})$, then* $\operatorname{curl} \nabla f$ *vanishes identically on G, i.e.*

$$\operatorname{curl} \nabla f \equiv 0.$$

(c) *If G is open and $v \in C^2(G, \mathbb{K}^3)$, then* $\operatorname{div} \operatorname{curl} v$ *vanishes identically on G, i.e.*

$$\operatorname{div} \operatorname{curl} v \equiv 0.$$

Proof. Exercise. ■

2.4 Interlude: Graphical Representation in Two Dimensions

In this section, we will briefly address the problem of drawing graphs of functions $f : D_f \longrightarrow \mathbb{R}$ with $D_f \subseteq \mathbb{R}^2$. If the function f is sufficiently benign (for example, if $f \in C^1(\mathbb{R}^2)$), then the graph of f, namely the set $\{(x, y, z) \in \mathbb{R}^3 : (x, y) \in D_f,\ z = f(x, y)\} \subseteq \mathbb{R}^3$ will represent a two-dimensional surface in the three-dimensional space $\mathbb{R}^3$. The two most important methods for depicting the graph of f as a picture in a two-dimensional plane (such as a sheet of paper or a board) are:

(a) The use of perspective.

(b) The use of level sets, in particular, level curves (also known as contour lines).

The Use of Perspective

Nowadays, this is most effectively accomplished by the use of computer graphics software. Widely used programs include commercial software such as *MATLAB* and *Mathematica* as well as the noncommercial software *Gnuplot*.

The Use of Level Sets

By a *level set* or an *isolevel*, we mean a set of the form $f^{-1}\{C\} = \{(x, y) \in D_f : f(x, y) = C\}$ with $C \in \mathbb{R}$. If $f^{-1}\{C\}$ constitutes a curve in $\mathbb{R}^2$, then we speak of a *level curve* or a *contour line*. Representation of functions depending on two variables by contour lines is well-known from everyday live. For example, contour lines are used to depict the height above sea level on hiking maps; on meteorological maps, isobars and isotherms are used to depict levels of equal pressure and equal temperature, respectively. Determining level sets and contour lines can be difficult, and the appropriate method depends on the function under consideration. In some cases, it is possible to determine the contour line corresponding to the level $C \in f(D_f)$ by solving the equation $C = f(x, y)$ for y (the difficulty is that an explicit solution of this equation can not always be found). The following Example 2.17 provides some cases, where $C = f(x, y)$ can be solved explicitly:

Example 2.17. (a) For $f : \mathbb{R}^2 \longrightarrow \mathbb{R}$, $f(x, y) := x^2 + y^2$, and $C \in \mathbb{R}_0^+$, one has

$$|y| = \sqrt{C - x^2} \quad \text{for } -\sqrt{C} \leq x \leq \sqrt{C}.$$

(b) For $f : \mathbb{R}^2 \longrightarrow \mathbb{R}$, $f(x, y) := xy$, and $C \in \mathbb{R}$, one has

$$y = \frac{C}{x} \quad \text{for } x \neq 0.$$

For $C = 0$, one actually gets $x = 0$ or $y = 0$, which provides one additional contour line.

—

In some cases, it helps to write $C = f(x, y)$ in polar coordinates (here, we will not pursue this further, as we did not discuss polar coordinates).

In general, the question if $C = f(x, y)$ can be solved for y (or x) is related to the implicit function Th. C.7 of the Appendix. Here, let us briefly discuss another method for the determination of contour lines which makes use of ordinary differential equations (ODE): Let $a < 0 < b$ and consider the path

$$\phi :]a, b[\longrightarrow \mathbb{R}^2, \quad \phi(t) := (x(t), y(t)) \tag{2.30}$$

as well as the function $h :]a,b[\longrightarrow \mathbb{R}$, $h = f \circ \phi$. We need the derivative of h. However, to fully understand the following formula (2.31), you will have to wait until we have discussed the chain rule (2.49) in Sec. 2.6 below. If ϕ and f are differentiable (see Def. 2.19 in Sec. 2.5 below), then the chain rule of Th. 2.28 below yields that h is differentiable with

$$h'(t) = \partial_1 f\big(x(t), y(t)\big)\, x'(t) + \partial_2 f\big(x(t), y(t)\big)\, y'(t). \tag{2.31}$$

If $h' \equiv 0$, then h is constant according to [Phi15a, Cor. 9.18(b)], i.e. $h(t) = f\big(x(t), y(t)\big) = C \in \mathbb{R}$ for each $t \in]a,b[$, i.e. f is constant with value C along the curve ϕ. Thus, if h is constant, then ϕ represents a contour line of f. A sufficient condition for h to be constant is the existence of some function $\lambda :]a,b[\times D_f \longrightarrow \mathbb{R}$ such that

$$x'(t) = \lambda\big(t, x(t), y(t)\big)\partial_2 f\big(x(t), y(t)\big) \quad \text{and} \quad y'(t) = -\lambda\big(t, x(t), y(t)\big)\partial_1 f\big(x(t), y(t)\big) \tag{2.32}$$

as one immediately verifies by plugging (2.32) into (2.31). For given λ, (2.32) constitutes a system of two ODE for the functions $t \mapsto x(t)$ and $t \mapsto y(t)$. One has the freedom to choose λ such that the system of ODE becomes as simple as possible (note that the choice $\lambda \equiv 0$ is not useful as, in this case, ϕ represents a point rather than a curve). To determine the contour line through a given point $(x_0, y_0) \in D_f$, one has to solve an *initial value problem* that consists of the system of ODE (2.32) completed with the *initial condition*

$$x(0) = x_0, \quad y(0) = y_0. \tag{2.33}$$

The following example shows a case, where one can exploit this method to determine contour lines:

Example 2.18. Consider

$$f : \mathbb{R}^2 \setminus \{(0,0)\} \longrightarrow \mathbb{R}, \quad f(x,y) := \frac{xy}{x^2 + y^2}. \tag{2.34}$$

From Example 2.2, we already know that $\nabla f(x,y) = \Big(\frac{y(y^2-x^2)}{(x^2+y^2)^2},\ \frac{x(x^2-y^2)}{(x^2+y^2)^2}\Big)$. Choosing $\lambda(t,x,y) := \frac{(x^2+y^2)^2}{x^2-y^2}$ for $x^2 \neq y^2$, we get from (2.32):

$$x'(t) = x(t) \quad \text{and} \quad y'(t) = y(t). \tag{2.35}$$

This system of ODE together with the initial conditions (2.33) is solved by

$$x(t) = x_0\, e^t \quad \text{and} \quad y(t) = y_0\, e^t. \tag{2.36}$$

This clearly represents the ray which originates from $(0,0)$ and passes through the point (x_0, y_0). Thus, these rays are the contour lines of f (i.e. f is constant along each ray). Note that we did not get any information on the behavior of f along the diagonals, where $y = \pm x$. However, in that case, $\nabla f(x,y) = 0$, such that (2.31) yields $h' \equiv 0$ also on the diagonals, which, thus, turn out to be contour lines as well.

2.5 The Total Derivative and the Notion of Differentiability

Roughly, a function $f : G \longrightarrow \mathbb{R}^m$, $G \subseteq \mathbb{R}^n$, will be called differentiable if, locally, it can be approximated by an affine function, i.e., if, for each $\xi \in G$, there exists an $\mathbb{R}$-linear function $L(\xi)$ such that $f(\xi + h) \approx f(\xi) + L(h)$ for sufficiently small $h \in \mathbb{R}^n$.

Analogous to the treatment in the one-dimensional situation in [Phi15a, Sec. 9], we will also consider $\mathbb{C}^m$-valued functions and call them differentiable if, and only if, both $\mathbb{R}^m$-valued functions $\operatorname{Re} f$ and $\operatorname{Im} f$ are differentiable. Also as in [Phi15a, Sec. 9], in this class, we will not study *complex* differentiability, which would mean locally approximating functions $f : G \longrightarrow \mathbb{C}^m$, $G \subseteq \mathbb{C}^n$, by $\mathbb{C}$-linear (more precisely, by $\mathbb{C}$-affine) functions. While the theory of complex differentiability has many similarities with the theory of real differentiability, there are also many significant differences, and it would take us too far afield to pursue this route, called the field of (multidimensional) complex analysis, in this class.

Definition 2.19. Let G be an open subset of $\mathbb{R}^n$, $n \in \mathbb{N}$, $f : G \longrightarrow \mathbb{R}^m$, $m \in \mathbb{N}$, $\xi \in G$. Then f is called *differentiable* in ξ if, and only if, there exists a linear map $L : \mathbb{R}^n \longrightarrow \mathbb{R}^m$ such that

$$\lim_{h \to 0} \frac{f(\xi + h) - f(\xi) - L(h)}{\|h\|_2} = 0. \tag{2.37a}$$

Note that, in general, L will depend on ξ. If f is differentiable in ξ, then L is called the *total derivative* or the *total differential* of f in ξ. In that case, one writes $Df(\xi)$ instead of L.

We call $f : G \longrightarrow \mathbb{C}^m$ differentiable in ξ if, and only if, both $\operatorname{Re} f$ and $\operatorname{Im} f$ are differentiable in ξ in the above sense. If f is differentiable in ξ, define $Df(\xi) := D \operatorname{Re} f(\xi) + iD \operatorname{Im} f(\xi)$ to be the *total derivative* or the *total differential* of f in ξ. It is then an easy exercise to show

$$\lim_{h \to 0} \frac{f(\xi + h) - f(\xi) - Df(\xi)(h)}{\|h\|_2} = 0. \tag{2.37b}$$

Finally, f is called *differentiable* if, and only if, f is differentiable in every $\xi \in G$.

Remark 2.20. (a) As the set $G \subseteq \mathbb{R}^n$ in Def. 2.19 is open, it is guaranteed that $\xi + h \in G$ for $\|h\|_2$ sufficiently small: There exists $\epsilon > 0$ such that $\|h\|_2 < \epsilon$ implies $\xi + h \in G$.

(b) As all norms on $\mathbb{R}^n$ are equivalent, instead of the Euclidean norm $\| \cdot \|_2$, one can use any other norm on $\mathbb{R}^n$ in (2.37a) without changing the definition.

Lemma 2.21. *Let G be an open subset of $\mathbb{R}^n$, $n \in \mathbb{N}$, $\xi \in G$. Then $f : G \longrightarrow \mathbb{R}^m$, $m \in \mathbb{N}$, is differentiable in ξ if, and only if, there exists a linear map $L : \mathbb{R}^n \longrightarrow \mathbb{R}^m$ and another (not necessarily linear) map $r : \mathbb{R}^n \longrightarrow \mathbb{R}^m$ such that*

$$f(\xi + h) - f(\xi) = L(h) + r(h) \tag{2.38a}$$

for each $h \in \mathbb{R}^n$ with sufficiently small $\|h\|_2$, and

$$\lim_{h\to 0} \frac{r(h)}{\|h\|_2} = 0. \tag{2.38b}$$

Proof. Suppose L, r are as above and satisfy (2.38). Then, for each $0 \neq h \in \mathbb{R}^n$ with sufficiently small $\|h\|_2$, it holds that

$$\frac{f(\xi+h) - f(\xi) - L(h)}{\|h\|_2} = \frac{r(h)}{\|h\|_2}. \tag{2.39}$$

Thus, (2.38b) implies (2.37a), showing that f is differentiable. Conversely, if f is differentiable in ξ, then there exists a linear map $L : \mathbb{R}^n \longrightarrow \mathbb{R}^m$ satisfying (2.37a). Choose $\epsilon > 0$ such that $B_{\epsilon,\|\cdot\|_2}(\xi) \subseteq G$ and define

$$r : \mathbb{R}^n \longrightarrow \mathbb{R}^m, \quad r(h) := \begin{cases} f(\xi+h) - f(\xi) - L(h) & \text{for } h \in B_{\epsilon,\|\cdot\|_2}(\xi), \\ 0 & \text{otherwise.} \end{cases} \tag{2.40}$$

Then (2.38a) is immediate. Since (2.39) also holds, (2.37a) implies (2.38b). ■

Theorem 2.22. *Let G be an open subset of $\mathbb{R}^n$, $n \in \mathbb{N}$, $\xi \in G$. If $f : G \longrightarrow \mathbb{K}$ is differentiable in ξ, then f is continuous in ξ, all partials at ξ, i.e. $\partial_j f(\xi)$, $j \in \{1,\ldots,n\}$, exist, and $Df(\xi) = \nabla f(\xi)$ (that means, for each $h = (h_1,\ldots,h_n) \in \mathbb{R}^n$, one has $Df(\xi)(h) = \nabla f(\xi)h = \sum_{j=1}^n \partial_j f(\xi) h_j$). In particular, $Df(\xi)$ is unique and, hence, well-defined.*

Proof. Assume f is differentiable in ξ. We first consider the case $\mathbb{K} = \mathbb{R}$. Let the linear map $L : \mathbb{R}^n \longrightarrow \mathbb{R}$ and $r : \mathbb{R}^n \longrightarrow \mathbb{R}$ be as in Lem. 2.21. We already know from Example 1.62 that each linear map from $\mathbb{R}^n$ into $\mathbb{R}$ is continuous. In particular, L must be continuous. Now let $(x^k)_{k\in\mathbb{N}}$ be a sequence in G that converges to ξ, i.e. $\lim_{k\to\infty} \|x^k - \xi\|_2 = 0$. Then $(h^k)_{k\in\mathbb{N}}$ with $h^k := x^k - \xi$ constitutes a sequence in $\mathbb{R}^n$ such that $\lim_{k\to\infty} \|h^k\|_2 = 0$. Note that (2.38b) implies that $0 \leq |r(h)| < \|h\|_2$ for $\|h\|_2$ sufficiently small. Thus, $\lim_{k\to\infty} \|h^k\|_2 = 0$ implies $\lim_{k\to\infty} |r(h^k)| = 0$. As the continuity of L also yields $\lim_{k\to\infty} |L(h^k)| = 0$, (2.38a) provides

$$\begin{aligned} \lim_{k\to\infty} \big|f(x^k) - f(\xi)\big| &= \lim_{k\to\infty} \big|f(\xi+h^k) - f(\xi)\big| \\ &= \lim_{k\to\infty} \big|L(h^k)\big| + \lim_{k\to\infty} \big|r(h^k)\big| = 0, \end{aligned} \tag{2.41}$$

establishing the continuity of f in ξ. To see that the partials exist and that L is given by the gradient, set $l_j := L(e_j)$ for each $j \in \{1,\ldots,n\}$. If $h = te_j$ with $t \in \mathbb{R}$ sufficiently close to 0, than (2.38a) yields

$$f(\xi + te_j) - f(\xi) = t\, l_j + r(te_j). \tag{2.42}$$

For $t \neq 0$, we can divide by t. Letting $t \to 0$, we see from (2.38b) that the right-hand side converges to l_j. But this means that the left-hand side must converge as

well, and comparing with (2.2), we see that its limit is precisely $\partial_j f(\xi)$, thereby proving $l_j = \partial_j f(\xi)$ as claimed.

We now consider the case $\mathbb{K} = \mathbb{C}$. From the case $\mathbb{K} = \mathbb{R}$, we know $\operatorname{Re} f$ and $\operatorname{Im} f$ are both continuous at ξ, such that, by Th. 1.58, f must be continuous at ξ as well. Moreover, from the case $\mathbb{K} = \mathbb{R}$, we know $\partial_j \operatorname{Re} f(\xi)$ and $\partial_j \operatorname{Im} f(\xi)$ exist for each $j \in \{1, \ldots, n\}$. Thus, $\partial_j f(\xi) = \partial_j \operatorname{Re} f(\xi) + i\, \partial_j \operatorname{Im} f(\xi)$ exist as well by [Phi15a, Rem. 9.2]. ■

By applying Th. 2.22 to coordinate functions, we can immediately extend it to $\mathbb{K}^m$-valued functions:

Corollary 2.23. *Let G be an open subset of $\mathbb{R}^n$, $n \in \mathbb{N}$, $\xi \in G$. If $f : G \longrightarrow \mathbb{K}^m$ is differentiable in ξ, then f is continuous in ξ, all partials at ξ, i.e. $\partial_k f_l(\xi)$, $k \in \{1, \ldots, n\}$, $l \in \{1, \ldots, m\}$, exist, and $Df(\xi) = J_f(\xi)$: For each $h = (h_1, \ldots, h_n) \in \mathbb{R}^n$, one has*

$$Df(\xi)(h) = J_f(\xi) \begin{pmatrix} h_1 \\ \vdots \\ h_n \end{pmatrix} = \begin{pmatrix} \nabla f_1(\xi)(h) \\ \vdots \\ \nabla f_m(\xi)(h) \end{pmatrix} = \begin{pmatrix} \sum_{k=1}^n \partial_k f_1(\xi) h_k \\ \vdots \\ \sum_{k=1}^n \partial_k f_m(\xi) h_k \end{pmatrix}.$$

In particular, $Df(\xi)$ is unique and, hence, well-defined. ■

Example 2.24. (a) If $G \subseteq \mathbb{R}^n$, $n \in \mathbb{N}$, is open and $f : G \longrightarrow \mathbb{K}^m$ is constant (i.e. there is $c \in \mathbb{K}^m$ such that $f(x) = c$ for each $x \in G$), than f is differentiable with $Df \equiv 0$: It suffices to notice that, for a constant f and $L \equiv 0$, the numerator in (2.37a) vanishes identically.

(b) If $A : \mathbb{R}^n \longrightarrow \mathbb{R}^m$ is $\mathbb{R}$-linear, then A is differentiable with $Df(\xi) = A$ for each $\xi \in \mathbb{R}^n$: If $\xi, h \in \mathbb{R}^n$, then $A(\xi + h) - A(\xi) - A(h) = 0$, showing that, as in (a), the numerator in (2.37a) (with $f = L = A$) vanishes identically. If $A : \mathbb{R}^n \longrightarrow \mathbb{C}^m$ is the restriction of a $\mathbb{C}$-linear map, then both $\operatorname{Re} A$ and $\operatorname{Im} A$ are $\mathbb{R}$-linear, and the above implies A is still differentiable with $Df(\xi) = A$ for each $\xi \in \mathbb{R}^n$.

(c) Let us revisit the well-known case of a differentiable one-dimensional function (cf. [Phi15a, Def. 9.1]), and compare this notion of differentiability with the more general one of Def. 2.19. Thus, let G be an open subset of $\mathbb{R}$, $\xi \in G$, and $f : G \longrightarrow \mathbb{K}$. We claim that f is differentiable at ξ in the sense of [Phi15a, (9.1)] if, and only if, f is differentiable at ξ in the sense of Def. 2.19 with

$$Df(\xi) : \mathbb{R} \longrightarrow \mathbb{K}, \quad Df(\xi)(h) := f'(\xi)h. \tag{2.43}$$

As, in both situations, a $\mathbb{C}$-valued f is differentiable at ξ if, and only if, both $\operatorname{Re} f$ and $\operatorname{Im} f$ are differentiable at ξ, it suffices to consider $\mathbb{K} = \mathbb{R}$. Thus, let f be $\mathbb{R}$-valued. If f is differentiable at ξ as a one-dimensional function and we use the $Df(\xi)$ according to (2.43) for the linear map L of Def. 2.19, then we get, for each $0 \neq h \in \mathbb{R}$ sufficiently close to 0,

$$\begin{aligned} \frac{f(\xi + h) - f(\xi) - L(h)}{\|h\|_2} &= \frac{f(\xi + h) - f(\xi) - f'(\xi)h}{|h|} \\ &= \begin{cases} \frac{f(\xi+h)-f(\xi)}{h} - f'(\xi) & \text{for } h > 0, \\ f'(\xi) - \frac{f(\xi+h)-f(\xi)}{h} & \text{for } h < 0. \end{cases} \end{aligned} \tag{2.44a}$$

Furthermore, $f'(\xi) = \lim_{h\to 0} \frac{f(\xi+h)-f(\xi)}{h}$ by its definition, i.e.

$$\lim_{h\to 0} \left| \frac{f(\xi+h)-f(\xi)}{h} - f'(\xi) \right| = 0. \tag{2.44b}$$

Combining (2.44a) and (2.44b), one obtains

$$\lim_{h\to 0} \frac{f(\xi+h)-f(\xi)-L(h)}{\|h\|_2} = 0, \tag{2.44c}$$

showing that f is differentiable in ξ in the sense of Def. 2.19. Conversely, if f is differentiable in ξ in the sense of Def. 2.19, then, according to Th. 2.22, $\partial_1 f(\xi)$ exists and $Df(\xi)(h) = \partial_1 f(\xi)h$. Thus, the one-dimensional differentiability of f at ξ as well as (2.43) follow by noticing that the definitions of $\partial_1 f(\xi)$ and of $f'(\xi)$ are identical.

Proposition 2.25. *Forming the total derivative is a linear operation: Let G be an open subset of $\mathbb{R}^n$, $n \in \mathbb{N}$, $\xi \in G$.*

(a) *If $f, g : G \longrightarrow \mathbb{K}^m$, $m \in \mathbb{N}$, are both differentiable at ξ, then $f+g$ is differentiable at ξ and $D(f+g)(\xi) = Df(\xi) + Dg(\xi)$.*

(b) *If $f : G \longrightarrow \mathbb{K}^m$, $m \in \mathbb{N}$, is differentiable at ξ and $\lambda \in \mathbb{K}$, then λf is differentiable at ξ and $D(\lambda f)(\xi) = \lambda Df(\xi)$.*

Proof. (a): We first consider $\mathbb{K} = \mathbb{R}$ and note that, for each $h \in \mathbb{R}^n$ with $0 \neq \|h\|_2$ sufficiently small,

$$\begin{aligned} &\frac{(f+g)(\xi+h)-(f+g)(\xi)-Df(\xi)(h)-Dg(\xi)(h)}{\|h\|_2} \\ &= \frac{f(\xi+h)-f(\xi)-Df(\xi)(h)}{\|h\|_2} + \frac{g(\xi+h)-g(\xi)-Dg(\xi)(h)}{\|h\|_2}. \end{aligned} \tag{2.45a}$$

Thus, if the limit $\lim_{h\to 0}$ exists and equals 0 for both summands on the right-hand side of (2.45a), then the same must be true for the left-hand side of (2.45a). The case $\mathbb{K} = \mathbb{C}$ now follows by applying the case $\mathbb{K} = \mathbb{R}$ to $\mathrm{Re}(f+g) = \mathrm{Re}\, f + \mathrm{Re}\, g$ and to $\mathrm{Im}(f+g) = \mathrm{Im}\, f + \mathrm{Im}\, g$.

(b): Again, we consider the case $\mathbb{K} = \mathbb{R}$ first. For $\lambda \in \mathbb{R}$, one computes

$$\lim_{h\to 0} \frac{(\lambda f)(\xi+h)-(\lambda f)(\xi)-\lambda Df(\xi)(h)}{\|h\|_2} = \lambda \lim_{h\to 0} \frac{f(\xi+h)-f(\xi)-Df(\xi)(h)}{\|h\|_2} = 0, \tag{2.45b}$$

thereby establishing the case. For $\mathbb{K} = \mathbb{C}$, one now applies the case $\mathbb{K} = \mathbb{R}$ and (a) to $\mathrm{Re}(\lambda f) = \mathrm{Re}\,\lambda\, \mathrm{Re}\, f - \mathrm{Im}\,\lambda\, \mathrm{Im}\, f$ and to $\mathrm{Im}(\lambda f) = \mathrm{Re}\,\lambda\, \mathrm{Im}\, f + \mathrm{Im}\,\lambda\, \mathrm{Re}\, f$. ■

Even though we have seen in Example 2.2 that the existence of all partial derivatives does not even imply continuity, let alone differentiability, the next theorem and its corollary will show that if all partial derivatives exist and are *continuous*, then that does, indeed, imply differentiability.

Theorem 2.26. *Let G be an open subset of $\mathbb{R}^n$, $n \in \mathbb{N}$, $\xi \in G$, and $f : G \longrightarrow \mathbb{K}$. If all partials $\partial_j f$, $j \in \{1, \ldots, n\}$ exist everywhere in G and are continuous in ξ, then f is differentiable in ξ, and, in particular, f is continuous in ξ.*

Proof. As usual, the case $\mathbb{K} = \mathbb{C}$ follows by applying the case $\mathbb{K} = \mathbb{R}$ to $\operatorname{Re} f$ and $\operatorname{Im} f$. We, therefore proceed to treat the case $\mathbb{K} = \mathbb{R}$. We first consider the special case where $\partial_j f(\xi) = 0$ for each $j \in \{1, \ldots, n\}$. In that case, we need to show

$$\lim_{h \to 0} \frac{f(\xi + h) - f(\xi)}{\|h\|_2} = 0. \tag{2.46}$$

Since the norms $\| \cdot \|_1$ and $\| \cdot \|_2$ are equivalent on $\mathbb{R}^n$, there is $C \in \mathbb{R}^+$ such that $\|h\|_1 \leq C\|h\|_2$ for each $h \in \mathbb{R}^n$. Since G is open and since the $\partial_j f$ are continuous in ξ, given $\epsilon > 0$, there is $\delta > 0$ such that, for each $h \in \mathbb{R}^n$ with $\|h\|_2 < \delta$, one has $\xi + h \in G$ and $\big|\partial_j f(\xi + h)\big| < \epsilon/C$ for every $j \in \{1, \ldots, n\}$. Fix $h \in \mathbb{R}^n$ with $\|h\|_2 < \delta$. For $j \in \{1, \ldots, n\}$, if $h_{n-j} = 0$, then set $\theta_j := 0$. If $h_{n-j} \neq 0$, then, as in the proof of Prop. 2.3, we make use of the fact that the one-dimensional mean value theorem [Phi15a, Th. 9.17] provides a number $\theta_j \in]0, h_{n-j}[$ such that

$$f(\xi + h) - f(\xi) = \sum_{j=0}^{n-1} h_{n-j}\, \partial_{n-j} f \left(\xi + \theta_j e_{n-j} + \sum_{k=1}^{n-(j+1)} h_k e_k \right) \tag{2.47}$$

((2.47) follows by combining (2.11) and (2.13)). Noting that $\|h\|_2 < \delta$ implies $\big\|\theta_j e_{n-j} + \sum_{k=1}^{n-(j+1)} h_k e_k\big\|_2 < \delta$, we obtain from (2.47) that, for $0 \neq h$ with $\|h\|_2 < \delta$,

$$\frac{\big|f(\xi + h) - f(\xi)\big|}{\|h\|_2} < \frac{1}{\|h\|_2} \sum_{j=0}^{n-1} |h_{n-j}|\, \frac{\epsilon}{C} \leq \epsilon, \tag{2.48}$$

thereby proving (2.46) and establishing the case. It remains to consider a general $f : G \longrightarrow \mathbb{R}$, without the restriction of a vanishing gradient. For such a general f, consider the modified function $g : G \longrightarrow \mathbb{R}$, $g(x) := f(x) - \nabla f(\xi)(x) = f(x) - \sum_{j=1}^{n} \partial_j f(\xi) x_j$. For g, we then get $\partial_j g(x) = \partial_j f(x) - \partial_j f(\xi)$ for each $x \in G$. In particular, the $\partial_j g$ exist in G, are continuous at $x = \xi$, and vanish at $x = \xi$. Thus, the first part of the proof applies to g, showing that g is differentiable at ξ. Since $f = g + \nabla f(\xi)$ and both g and the linear map $\nabla f(\xi)$ are differentiable at ξ, so is f by Prop. 2.25(a). ■

Corollary 2.27. *Let G be an open subset of $\mathbb{R}^n$, $n \in \mathbb{N}$, $\xi \in G$, and $f : G \longrightarrow \mathbb{K}^m$, $m \in \mathbb{N}$. If all partials $\partial_k f_l$, $k \in \{1, \ldots, n\}$, $l \in \{1, \ldots, m\}$, exist everywhere in G and are continuous in ξ, then f is differentiable in ξ, and, in particular, f is continuous in ξ.*

Proof. Applying Th. 2.26 to the coordinate functions f_l, $l \in \{1, \ldots, m\}$, yields that each f_l is differentiable at ξ. However, since a $\mathbb{K}^m$-valued function converges if, and only if, each of its coordinate functions converges, f must also be differentiable at ξ. ■

2.6 The Chain Rule

As for one-dimensional differentiable functions, one can also prove a chain rule for vector-valued differentiable functions:

Theorem 2.28. *Let $m, n, p \in \mathbb{N}$. Let $G_f \subseteq \mathbb{R}^n$ be open, $f : G_f \longrightarrow \mathbb{R}^m$, let $G_g \subseteq \mathbb{R}^m$ be open, $g : G_g \longrightarrow \mathbb{K}^p$, $f(G_f) \subseteq G_g$. If f is differentiable at $\xi \in G_f$ and g is differentiable at $f(\xi) \in G_g$, then $g \circ f : G_f \longrightarrow \mathbb{K}^p$ is differentiable at ξ and, for the $\mathbb{R}$-linear maps $D(g\circ f)(\xi) : \mathbb{R}^n \longrightarrow \mathbb{K}^p$, $Df(\xi) : \mathbb{R}^n \longrightarrow \mathbb{R}^m$, and $Dg\big(f(\xi)\big) : \mathbb{R}^m \longrightarrow \mathbb{K}^p$, the following* chain rule *holds:*

$$D(g \circ f)(\xi) = Dg\big(f(\xi)\big) \circ Df(\xi). \tag{2.49}$$

In particular, if both f and g are differentiable, then $g \circ f$ is differentiable.

Proof. The proof is noticeably harder than in the one-dimensional case and can be found in Sec. C.1 of the appendix. ■

Example 2.29. In the setting of the chain rule of Th. 2.28, we consider the special case $n = p = 1$. Thus, we have an open subset G_f of $\mathbb{R}$ and $f : G_f \longrightarrow \mathbb{R}^m$. The map g maps G_g into $\mathbb{K}$ and for $h := g \circ f : G_f \longrightarrow \mathbb{K}$, we have $h(t) = g\big(f_1(t), \ldots, f_m(t)\big)$. In this case, one computes the one-dimensional function h by making a detour through the m-dimensional space $\mathbb{R}^m$. If f is differentiable at $\xi \in G_f$ and g is differentiable at $f(\xi) \in G_g$, the chain rule (2.49) now reads

$$Dh(\xi) = D(g \circ f)(\xi) = Dg\big(f(\xi)\big) \circ Df(\xi) = \nabla g\big(f(\xi)\big) J_f(\xi) = \sum_{j=1}^{m} \partial_j g\big(f(\xi)\big) \partial_1 f_j(\xi). \tag{2.50}$$

Recall from Example 2.24(c) that, for one-dimensional functions such as h, the function $Dh(\xi) : \mathbb{R} \longrightarrow \mathbb{K}$ corresponds to the number $h'(\xi) \in \mathbb{K}$ via (2.43). Also recall that, for one-dimensional functions such as f_j, the partial derivative $\partial_1 f_j$ coincides with the one-dimensional derivative f_j'. Thus, (2.50) implies

$$h'(\xi) = \sum_{j=1}^{m} \partial_j g\big(f(\xi)\big) f_j'(\xi). \tag{2.51}$$

Definition 2.30. Let $G \subseteq \mathbb{R}^m$, $m \in \mathbb{N}$. A *differentiable path* is a differentiable function $\phi :]a, b[\longrightarrow G$, $a, b \in \mathbb{R}$, $a < b$. The set G is called *connected by differentiable paths* if, and only if, for each $x, y \in G$, there some differentiable path $\phi :]a, b[\longrightarrow G$ such that $\phi(s) = x$ and $\phi(t)$ for suitable $s, t \in]a, b[$.

Proposition 2.31. *Let $G \subseteq \mathbb{R}^m$ be open, $m \in \mathbb{N}$. If G is connected by differentiable paths and $f : G \longrightarrow \mathbb{K}$ is differentiable with $\nabla f \equiv 0$, then f is constant.*

Proof. Let $x, y \in G$, and let $\phi :]a, b[\longrightarrow G$ be a differentiable path connecting x and y, i.e. $\phi(s) = x$ and $\phi(t)$ for suitable $s, t \in]a, b[$. Define the auxiliary function $h :]a, b[\longrightarrow \mathbb{K}$,

$h = f \circ \phi$. By the chain rule of Th. 2.28, h is differentiable and, using (2.51) and $\partial_j f \equiv 0$ for each $j \in \{1, \ldots, m\}$,

$$h'(\xi) = \sum_{j=1}^{m} \partial_j f\big(\phi(\xi)\big)\phi_j'(\xi) = 0 \quad \text{for each } \xi \in]a, b[. \tag{2.52}$$

As a one-dimensional function on an open interval with vanishing derivative, h must be constant (as both $\operatorname{Re} h$ and $\operatorname{Im} h$ must be constant by [Phi15a, Cor. 9.18(b)]), implying $f(x) = f(\phi(s)) = h(s) = h(t) = f(\phi(t)) = f(y)$, showing that f is constant as well. ■

2.7 The Mean Value Theorem

Another application of the chain rule in several variables is the mean value theorem in several variables:

Theorem 2.32. *Let $G \subseteq \mathbb{R}^n$ be open, $n \in \mathbb{N}$, $f : G \longrightarrow \mathbb{R}$. If f is differentiable on G and $x, y \in G$ such that the entire line segment connecting x and y is also contained in G, i.e. $S_{x,y} := \{x + t(y - x) : 0 < t < 1\} \subseteq G$, then there is $\xi \in S_{x,y}$ satisfying*

$$f(y) - f(x) = Df(\xi)(y - x) = \nabla f(\xi)(y - x) = \sum_{j=1}^{n} \partial_j f(\xi)(y_j - x_j). \tag{2.53}$$

Proof. We merely need to combine the one-dimensional mean value theorem [Phi15a, Th. 9.17] with the chain rule of Th. 2.28. A small problem arises from the fact that, in Th. 2.28, we required G_f to be open. We therefore note that the openness of G allows us to find some $\epsilon > 0$ such that the small extension $S_{x,y,\epsilon} := \{x + t(y - x) : -\epsilon < t < 1 + \epsilon\}$ is still contained in G: $S_{x,y,\epsilon} \subseteq G$. Consider the auxiliary functions

$$\phi :]-\epsilon, 1+\epsilon[\longrightarrow \mathbb{R}^n, \qquad \phi(t) := x + t(y - x) \tag{2.54a}$$

$$h :]-\epsilon, 1+\epsilon[\longrightarrow \mathbb{R}, \qquad h(t) := (f \circ \phi)(t) = f\big(x + t(y - x)\big). \tag{2.54b}$$

As the sum of a constant function and a linear function, ϕ is differentiable, and $D\phi(t) : \mathbb{R} \longrightarrow \mathbb{R}^n$, $D\phi(t) = y - x$ (that means, for each $\alpha \in \mathbb{R}$, one has $D\phi(t)(\alpha) = \alpha(y - x)$). Thus, according to Th. 2.28, h is differentiable, and, using (2.50),

$$Dh(t) = Df\big(\phi(t)\big) \circ D\phi(t) = \nabla f\big(\phi(t)\big)(y - x). \tag{2.55}$$

The one-dimensional mean value theorem [Phi15a, Th. 9.17] provides $\theta \in]0, 1[$ such that

$$f(y) - f(x) = h(1) - h(0) = h'(\theta). \tag{2.56}$$

As in Example 2.29, we recall from (2.43) that the real number $h'(\theta)$ represents the linear map $Dh(\theta)$ such that we can combine (2.55) and (2.56) to obtain

$$f(y) - f(x) = h'(\theta) = \nabla f\big(\phi(\theta)\big)(y - x) = \nabla f(\xi)(y - x) \tag{2.57}$$

with $\xi := \phi(\theta) = x + \theta(y - x) \in S_{x,y}$, concluding the proof of (2.53). ■

Caveat 2.33. Unlike many other results of this class, Th. 2.32 does *not* extend to $\mathbb{C}$-valued functions – actually, even the one-dimensional mean value theorem does not extend to $\mathbb{C}$-valued functions. It is an exercise to find an explicit counterexample of a differentiable function $f : \mathbb{R} \longrightarrow \mathbb{C}$ and $x, y \in \mathbb{R}$, $x < y$, such that there does not exist $\xi \in]x, y[$ satisfying $f(y) - f(x) = f'(\xi)(y - x)$.

—

As an application of Th. 2.32, let us prove that differentiable maps with bounded partials are Lipschitz continuous on convex sets.

Definition 2.34. A set $G \subseteq \mathbb{R}^n$, $n \in \mathbb{N}$, is called *convex* if, and only if, for each $x, y \in G$, one has $S_{x,y} := \{x + t(y - x) : 0 < t < 1\} \subseteq G$.

Theorem 2.35. *Let $G \subseteq \mathbb{R}^n$ be open, $n \in \mathbb{N}$, and let $f : G \longrightarrow \mathbb{K}$ be differentiable. Suppose there exist $M_1, M_2 \in \mathbb{R}_0^+$ such that $|\partial_j \operatorname{Re} f(\xi)| \leq M_1$ and $|\partial_j \operatorname{Im} f(\xi)| \leq M_2$ for each $j \in \{1, \ldots, n\}$ and each $\xi \in G$. If G is convex, then f is Lipschitz continuous with Lipschitz constant $L := M_1 + M_2$ with respect to the 1-norm on $\mathbb{R}^n$ and with Lipschitz constant cL, $c > 0$, with respect to an arbitrary norm on $\mathbb{R}^n$.*

Proof. Since f is differentiable and G is convex, given $x, y \in G$, we can apply Th. 2.32 to obtain $\xi_1, \xi_2 \in G$ such that

$$\begin{aligned} |f(y) - f(x)| &\overset{\text{[Phi15a, Th. 5.11(d)]}}{\leq} |\operatorname{Re} f(y) - \operatorname{Re} f(x)| + |\operatorname{Im} f(y) - \operatorname{Im} f(x)| \\ &\overset{(2.53)}{\leq} \sum_{j=1}^{n} \big(|\partial_j \operatorname{Re} f(\xi_1)| + |\partial_j \operatorname{Im} f(\xi_2)|\big)\, |y_j - x_j| \\ &\overset{\text{[Phi15a, Th. 5.11(d)]}}{\leq} (M_1 + M_2)\|y - x\|_1, \end{aligned} \tag{2.58}$$

showing that, with respect to the 1-norm, f is Lipschitz continuous with Lipschitz constant $M_1 + M_2$. Since all norms on $\mathbb{R}^n$ are equivalent, we also get that f is Lipschitz continuous with Lipschitz constant cL, $c > 0$, with respect to all other norms on $\mathbb{R}^n$. ■

For $\mathbb{R}^m$-valued variants of Th. 2.35, see Th. C.1 and Th. C.3 of the Appendix.

2.8 Directional Derivatives

Given a real-valued function f, the partial derivatives $\partial_j f$ (if they exist) describe the local change of f in the direction of the standard unit vector e_j. We would now like to generalize the notion of partial derivative in such a way that it allows us to study the change of f in an arbitrary direction $e \in \mathbb{R}^n$. This leads to the following notion of directional derivatives.

Definition 2.36. Let $G \subseteq \mathbb{R}^n$, $n \in \mathbb{N}$, $f : G \longrightarrow \mathbb{K}$, $\xi \in G$, $e \in \mathbb{R}^n$. If there is $\epsilon > 0$ such that $\xi + he \in G$ for each $h \in]0, \epsilon[$ (this condition is trivially satisfied if ξ is an

interior point of G), then f is said to have a *directional derivative* at ξ in the direction e if, and only if, the limit

$$\lim_{h\downarrow 0} \frac{f(\xi + he) - f(\xi)}{h} \tag{2.59}$$

exists in $\mathbb{K}$. In that case, this limit is identified with the corresponding directional derivative and denoted by $\frac{\partial f}{\partial e}(\xi)$ or by $\delta f(\xi, e)$. If the directional derivative of f in the direction e exists for each $\xi \in G$, then the function

$$\frac{\partial f}{\partial e} : G \longrightarrow \mathbb{K}, \quad \xi \mapsto \frac{\partial f}{\partial e}(\xi), \tag{2.60}$$

is also called the directional derivative of f in the direction e.

Remark 2.37. Consider the setting of Def. 2.36 and suppose $e = e_j$ for some $j \in \{1, \ldots, n\}$. If ξ is an interior point of G, then the directional derivative $\frac{\partial f}{\partial e}(\xi)$ coincides with the partial derivative $\partial_j f(\xi)$ of Def. 2.1 if, and only if, both $\frac{\partial f}{\partial e}(\xi)$ and $\frac{\partial f}{\partial(-e)}(\xi)$ exist and $\frac{\partial f}{\partial e}(\xi) = -\frac{\partial f}{\partial(-e)}(\xi)$: If $\partial_j f(\xi)$ exists, then

$$\begin{aligned}
\partial_j f(\xi) &= \lim_{h\to 0} \frac{f(\xi + he_j) - f(\xi)}{h} = \lim_{h\downarrow 0} \frac{f(\xi + he_j) - f(\xi)}{h} = \frac{\partial f}{\partial e}(\xi) \\
&= \lim_{h\uparrow 0} \frac{f(\xi + he_j) - f(\xi)}{h} = \lim_{h\downarrow 0} \frac{f(\xi - he_j) - f(\xi)}{-h} \\
&= -\lim_{h\downarrow 0} \frac{f(\xi + h(-e_j)) - f(\xi)}{h} = -\frac{\partial f}{\partial(-e)}(\xi).
\end{aligned} \tag{2.61}$$

On the other hand, if both $\frac{\partial f}{\partial e}(\xi)$ and $\frac{\partial f}{\partial(-e)}(\xi)$ exist and $\frac{\partial f}{\partial e}(\xi) = -\frac{\partial f}{\partial(-e)}(\xi)$, then the corresponding equalities in (2.61) show that both one-sided partials exist at ξ and that their values agree, showing that $\partial_j f(\xi) = \frac{\partial f}{\partial e}(\xi)$ exists.

—

We can now generalize Th. 2.22:

Theorem 2.38. *Let G be an open subset of $\mathbb{R}^n$, $n \in \mathbb{N}$, $\xi \in G$. If $f : G \longrightarrow \mathbb{K}$ is differentiable in ξ, then, for each $e = (\epsilon_1, \ldots, \epsilon_n) \in \mathbb{R}^n$, the directional derivative $\frac{\partial f}{\partial e}(\xi)$ exists and*

$$\frac{\partial f}{\partial e}(\xi) = \nabla f(\xi) \cdot e = \sum_{j=1}^{n} \epsilon_j \partial_j f(\xi). \tag{2.62}$$

Moreover, if we consider $\mathbb{K} = \mathbb{R}$ and only allow normalized $e \in \mathbb{R}^n$ with $\|e\|_2 = 1$, then the directional derivatives can take only values between $\alpha := \| \nabla f(\xi)\|_2$ and $-\alpha$, where the largest value (i.e. α) is attained in the direction $e_{\max} := \nabla f(\xi)/\alpha$ and the smallest value (i.e. $-\alpha$) is attained in the direction $e_{\min} := -e_{\max}$. For $n = 1$, $e = \pm 1$ are the only possible directions, yielding precisely the values α and $-\alpha$. For $n \geq 2$, all values in $[-\alpha, \alpha]$ are attained.

Proof. Since G is open, there is $\epsilon > 0$ such that $\xi + he \in G$ for each $h \in]-\epsilon, \epsilon[$. Similarly to the proof of Th. 2.32, consider auxiliary functions

$$\phi :]-\epsilon, \epsilon[\longrightarrow \mathbb{R}^n, \qquad \phi(h) := \xi + he, \tag{2.63a}$$

$$g :]-\epsilon, \epsilon[\longrightarrow \mathbb{K}, \qquad g(h) := (f \circ \phi)(h) = f\big(\xi + he\big). \tag{2.63b}$$

Theorem 2.28 yields the differentiability of g, and, as $D\phi \equiv e$ (i.e., for each $h \in]-\epsilon, \epsilon[$ and each $\alpha \in \mathbb{R}$, it is $D\phi(h)(\alpha) = \alpha e$), by (2.50), we have

$$\underset{h\in]-\epsilon,\epsilon[}{\forall} \quad g'(h) = D(f \circ \phi)(h) = Df\big(\phi(h)\big) \circ D\phi(h) = \nabla f(\xi + he) \cdot e \tag{2.64a}$$

and

$$\frac{\partial f}{\partial e}(\xi) = g'(0) = \nabla f(\xi) \cdot e, \tag{2.64b}$$

proving (2.62). Applying the Cauchy-Schwarz inequality (1.81) to (2.62) yields

$$\left|\frac{\partial f}{\partial e}(\xi)\right| = \big|\nabla f(\xi) \cdot e\big| \leq \big\|\nabla f(\xi)\big\|_2 \big\|e\big\|_2 = \alpha \|e\|_2. \tag{2.65}$$

Thus, for $\mathbb{K} = \mathbb{R}$ and $e \in \mathbb{R}^n$ with $\|e\|_2 = 1$, we have $-\alpha \leq \frac{\partial f}{\partial e}(\xi) \leq \alpha$. It remains to show that, for $\mathbb{K} = \mathbb{R}$ and $n \geq 2$, the map

$$D : S_1(0) \longrightarrow [-\alpha, \alpha], \quad D(e) := \nabla f(\xi) \cdot e = \sum_{j=1}^{n} \epsilon_j \partial_j f(\xi),$$

is surjective. The details are bit tedious and are carried out in App. C.3. ■

The following example shows that the existence of all directional derivatives does not imply continuity, let alone differentiability.

Example 2.39. Consider the function

$$f : \mathbb{R}^2 \longrightarrow \mathbb{K}, \quad f(x, y) := \begin{cases} 1 & \text{for } 0 < y < x^2, \\ 0 & \text{otherwise.} \end{cases} \tag{2.66}$$

The function is not continuous in $(0, 0)$: Let $x_n := 1/n$ and $y_n := 1/n^3$. Then $\lim_{n\to\infty}(x_n, y_n) = (0, 0)$. However, since $y_n = 1/n^3 < 1/n^2 = x_n^2$ for $n > 1$, one has

$$\lim_{n\to\infty} f(x_n, y_n) = 1 \neq 0 = f(0, 0). \tag{2.67}$$

We now claim that, for each $e = (\epsilon_x, \epsilon_y) \in \mathbb{R}^2$, the directional derivative $\frac{\partial f}{\partial e}(0, 0)$ exists and $\frac{\partial f}{\partial e}(0, 0) = 0$. For $\epsilon_y \leq 0$ this is immediate since, for each $h \in \mathbb{R}^+$, $f((0, 0) + h(\epsilon_x, \epsilon_y)) = f(h\epsilon_x, h\epsilon_y) = 0$. Now assume $\epsilon_y > 0$. If $\epsilon_x = 0$, then $f(h\epsilon_x, h\epsilon_y) = f(0, h\epsilon_y) = 0$ for each $h \in \mathbb{R}^+$, showing $\frac{\partial f}{\partial e}(0, 0) = 0$. It remains the case, where $\epsilon_y > 0$ and $\epsilon_x \neq 0$. In that case, one obtains $h^2\epsilon_x^2 < h\epsilon_y$ for each $0 < h < \frac{\epsilon_y}{\epsilon_x^2}$. Thus, for such h, $f(h\epsilon_x, h\epsilon_y) = 0$, once again proving $\frac{\partial f}{\partial e}(0, 0) = 0$.

3 Extreme Values and Stationary Points

3.1 Definitions of Extreme Values

The following Def. 3.1 is a generalization of [Phi15a, Def. 7.50].

Definition 3.1. Let (X, d) be a metric space, $M \subseteq X$, and $f : M \longrightarrow \mathbb{R}$.

(a) Given $x \in M$, f has a *(strict) global min* at x if, and only if, $f(x) \leq f(y)$ $(f(x) < f(y))$ for each $y \in M \setminus \{x\}$. Analogously, f has a *(strict) global max* at x if, and only if, $f(x) \geq f(y)$ $(f(x) > f(y))$ for each $y \in M \setminus \{x\}$. Moreover, f has a *(strict) global extreme value* at x if, and only if, f has a (strict) global min or a (strict) global max at x.

(b) Given $x \in M$, f has a *(strict) local min* at x if, and only if, there exists $\epsilon > 0$ such that $f(x) \leq f(y)$ $(f(x) < f(y))$ for each $y \in \{y \in M : d(x, y) < \epsilon\} \setminus \{x\}$. Analogously, f has a *(strict) local max* at x if, and only if, there exists $\epsilon > 0$ such that $f(x) \geq f(y)$ $(f(x) > f(y))$ for each $y \in \{y \in M : d(x, y) < \epsilon\} \setminus \{x\}$. Moreover, f has a *(strict) local extreme value* at x if, and only if, f has a (strict) local min or a (strict) local max at x.

Remark 3.2. In the context of Def. 3.1, it is immediate from the respective definitions that f has a (strict) global min at $x \in M$ if, and only if, $-f$ has a (strict) global max at x. Moreover, the same holds if "global" is replaced by "local". It is equally obvious that every (strict) global min/max is a (strict) local min/max.

3.2 Extreme Values of Continuous Functions on Compact Sets

Definition 3.3. A subset C of a metric space X is called *compact* if, and only if, every sequence in C has a subsequence that converges to some limit $c \in C$.

Proposition 3.4. *Let (X, d) be a metric space and $C \subseteq X$.*

(a) *If C is compact, then C is closed and bounded.*

(b) *If C is compact and $A \subseteq C$ is closed, then A is compact.*

Proof. (a): The proof is analogous to the second part of the proof of [Phi15a, Th. 7.48]: Suppose C is compact. Let $(x^k)_{k\in\mathbb{N}}$ be a sequence in C that converges in X, i.e. $\lim_{k\to\infty} x^k = x \in X$. Since C is compact, $(x^k)_{k\in\mathbb{N}}$ must have a subsequence that converges to some $c \in C$. However, according to Prop. 1.38(c), it must be $x = c \in C$. Due to the equivalence between statements (iv) and (i) of Cor. 1.44, C must be closed. If C is not bounded, then, for each $x \in X$, there is a sequence $(x^k)_{k\in\mathbb{N}}$ in C such that $\lim_{k\to\infty} d(x, x^k) = \infty$. If $y \in X$, then $d(x, x^k) \leq d(x, y) + d(y, x^k)$, i.e. $d(y, x^k) \geq d(x, x^k) - d(x, y)$, showing that $\lim_{k\to\infty} d(y, x^k) = \infty$ as well. Thus, y can not be a limit of any subsequence of $(x^k)_{k\in\mathbb{N}}$. As y was arbitrary, C can not be compact.

(b): If $(x^k)_{k\in\mathbb{N}}$ is a sequence in A, then $(x^k)_{k\in\mathbb{N}}$ is a sequence in C. Since C is compact, it must have a subsequence that converges to some $c \in C$. However, as A is closed, c must be in A, showing that $(x^k)_{k\in\mathbb{N}}$ has a subsequence that converges to some $c \in A$, i.e. A is compact. ■

Corollary 3.5. *A subset C of $\mathbb{K}^n$, $n \in \mathbb{N}$, is compact if, and only if, C is closed and bounded.*

Proof. Every compact set is closed and bounded by Prop. 3.4(a). If C is closed and bounded, and $(x^k)_{k\in\mathbb{N}}$ is a sequence in C, then the boundedness and the Bolzano-Weierstrass Th. 1.16(b) yield a subsequence that converges to some $x \in \mathbb{K}^n$. However, since C is closed, $x \in C$, showing that C is compact. ■

The following examples show that, in general, sets can be closed and bounded without being compact.

Example 3.6. (a) If (X, d) is a noncomplete metric space, than it contains a Cauchy sequence that does not converge. It is not hard to see that such a sequence can not have a convergent subsequence, either. This shows that no noncomplete metric space can be compact. Moreover, the closure of every bounded subset of X that contains such a nonconvergent Cauchy sequence is an example of a closed and bounded set that is noncompact. Concrete examples are given by $\mathbb{Q} \cap [a, b]$ for each $a, b \in \mathbb{R}$ with $a < b$ (these sets are $\mathbb{Q}$-closed, but not $\mathbb{R}$-closed!) and $]a, b[$ for each $a, b \in \mathbb{R}$ with $a < b$, in each case endowed with the usual metric $d(x, y) := |x - y|$.

(b) There can also be closed and bounded sets in complete spaces that are not compact. Consider the space X of all bounded sequences $(x_n)_{n\in\mathbb{N}}$ in $\mathbb{K}$, endowed with the sup-norm $\|(x_n)_{n\in\mathbb{N}}\|_{\sup} := \sup\{|x_n| : n \in \mathbb{N}\}$. It is not too difficult to see that X with the sup-norm is a Banach space: Let $(x^k)_{k\in\mathbb{N}}$ with $x^k = (x^k_n)_{n\in\mathbb{N}}$ be a Cauchy sequence in X. Then, for each $n \in \mathbb{N}$, $(x^k_n)_{k\in\mathbb{N}}$ is a Cauchy sequence in $\mathbb{K}$, and, thus, it has a limit $y_n \in \mathbb{K}$. Let $y := (y_n)_{n\in\mathbb{N}}$. Then

$$\|x^k - y\|_{\sup} = \sup\{|x^k_n - y_n| : n \in \mathbb{N}\}.$$

Let $\epsilon > 0$. As $(x^k)_{k\in\mathbb{N}}$ is a Cauchy sequence with respect to the sup-norm, there is $N \in \mathbb{N}$ such that $\|x^k - x^l\|_{\sup} < \epsilon$ for all $k, l > N$. Fix some $l > N$ and some $n \in \mathbb{N}$. Then $\epsilon \geq \lim_{k\to\infty} |x^k_n - x^l_n| = \lim_{k\to\infty} |y_n - x^l_n|$. Since this is valid for each $n \in \mathbb{N}$, we get $\|x^l - y\|_{\sup} \leq \epsilon$ for each $l > N$, showing $\lim_{l\to\infty} x^l = y$, i.e. X is complete and a Banach space.

Now consider the sequence $(e^k)_{k\in\mathbb{N}}$ with

$$e^k_n := \begin{cases} 1 & \text{for } k = n, \\ 0 & \text{otherwise.} \end{cases}$$

Then $(e^k)_{k\in\mathbb{N}}$ constitutes a sequence in X with $\|e^k\|_{\sup} = 1$ for each $k \in \mathbb{N}$. In particular, $(e^k)_{k\in\mathbb{N}}$ is a sequence inside the closed unit ball $\overline{B}_1(0)$, and, hence, bounded.

However, if $k, l \in \mathbb{N}$ with $k \neq l$, then $\|e^k - e^l\|_{\sup} = 1$. Thus, neither $(e^k)_{k\in\mathbb{N}}$ nor any subsequence can be a Cauchy sequence. In particular, no subsequence can converge, showing that the closed and bounded unit ball $\overline{B}_1(0)$ is not compact.

Note: There is an important result, provided as Th. B.29 of the Appendix, that shows a normed vector space is finite-dimensional if, and only if, the closed unit ball $\overline{B}_1(0)$ is compact.

Theorem 3.7. *If (X, d_X) and (Y, d_Y) are metric spaces, $C \subseteq X$ is compact, and $f : C \longrightarrow Y$ is continuous, then $f(C)$ is compact.*

Proof. The present theorem is a generalization of [Phi15a, Th. 7.52]; however the proof can still be conducted precisely as for [Phi15a, Th. 7.52]: If $(y^k)_{k\in\mathbb{N}}$ is a sequence in $f(C)$, then, for each $k \in \mathbb{N}$, there is some $x^k \in C$ such that $f(x^k) = y^k$. As C is compact, there is a subsequence $(a^k)_{k\in\mathbb{N}}$ of $(x^k)_{k\in\mathbb{N}}$ with $\lim_{k\to\infty} a^k = a$ for some $a \in C$. Then $(f(a^k))_{k\in\mathbb{N}}$ is a subsequence of $(y^k)_{k\in\mathbb{N}}$ and the continuity of f yields $\lim_{k\to\infty} f(a^k) = f(a) \in f(C)$, showing that $(y^k)_{k\in\mathbb{N}}$ has a convergent subsequence with limit in $f(C)$. We have therefore established that $f(C)$ is compact. ■

The following Th. 3.8 is a generalization of [Phi15a, Th. 7.54].

Theorem 3.8. *If (X, d) is a metric space, $C \subseteq X$ is compact, and $f : C \longrightarrow \mathbb{R}$ is continuous, then f assumes its max and its min, i.e. there are $x_m \in C$ and $x_M \in C$ such that f has a global min at x_m and a global max at x_M.*

Proof. The proof is still conducted precisely as in the special case [Phi15a, Th. 7.54]: Since C is compact and f is continuous, $f(C) \subseteq \mathbb{R}$ is compact according to Th. 3.7. Then, by [Phi15a, Lem. 7.53], $f(C)$ contains a smallest element m and a largest element M. This, in turn, implies that there are $x_m, x_M \in C$ such that $f(x_m) = m$ and $f(x_M) = M$. ■

A drawback of Th. 3.8 (as well as of [Phi15a, Th. 7.54]) is that its proof is not constructive. That means that, even though it guarantees the function has a max and a min, it does not give any clues as how to find them. For differentiable functions, we will see more constructive results in the following sections.

Theorem 3.9. *If (X, d_X) and (Y, d_Y) are metric spaces, $C \subseteq X$ is compact, and $f : C \longrightarrow Y$ is continuous, then f is uniformly continuous.*

Proof. If f is not uniformly continuous, then there must be some $\epsilon > 0$ such that, for each $k \in \mathbb{N}$, there exist $x^k, y^k \in C$ satisfying $d_X(x^k, y^k) < 1/k$ and $d_Y(f(x^k), f(y^k)) \geq \epsilon$. Since C is compact, there is $a \in C$ and a subsequence $(a^k)_{k\in\mathbb{N}}$ of $(x^k)_{k\in\mathbb{N}}$ such that $a = \lim_{k\to\infty} a^k$. Then there is a corresponding subsequence $(b^k)_{k\in\mathbb{N}}$ of $(y^k)_{k\in\mathbb{N}}$ such that $d_X(a^k, b^k) < 1/k$ and $d_Y(f(a^k), f(b^k)) \geq \epsilon$ for all $k \in \mathbb{N}$. Using the compactness of C again, there $b \in C$ and a subsequence $(v^k)_{k\in\mathbb{N}}$ of $(b^k)_{k\in\mathbb{N}}$ such that $b = \lim_{k\to\infty} v^k$. Now there is a corresponding subsequence $(u^k)_{k\in\mathbb{N}}$ of $(a^k)_{k\in\mathbb{N}}$ such that $d_X(u^k, v^k) < 1/k$ and $d_Y(f(u^k), f(v^k)) \geq \epsilon$ for all $k \in \mathbb{N}$. Note that we still have $a = \lim_{k\to\infty} v^k$.

Given $\alpha > 0$, there is $N \in \mathbb{N}$ such that, for each $k > N$, one has $d_X(a, u^k) < \alpha/3$, $d_X(b, v^k) < \alpha/3$, and $d_X(u^k, v^k) < 1/k < \alpha/3$. Thus, $d_X(a, b) < d_X(a, u^k) + d_X(u^k, v^k) + d_X(b, v^k) < \alpha$, implying $d(a, b) = 0$ and $a = b$. Finally, the continuity of f implies $f(a) = \lim_{k\to\infty} f(u^k) = \lim_{k\to\infty} f(v^k)$ in contradiction to $d_Y(f(u^k), f(v^k)) \geq \epsilon$. ■

Theorem 3.10. *If (X, d_X) and (Y, d_Y) are metric spaces, $C \subseteq X$ is compact, and $f : C \longrightarrow Y$ is continuous and one-to-one, then $f^{-1} : f(C) \longrightarrow C$ is continuous.*

Proof. Let $(y^k)_{k\in\mathbb{N}}$ be a sequence $f(C)$ such that $\lim_{k\to\infty} y^k = y \in f(C)$. Then there is a sequence $(x^k)_{k\in\mathbb{N}}$ in C such that $f(x^k) = y^k$ for each $k \in \mathbb{N}$. Let $x := f^{-1}(y)$. It remains to prove that $\lim_{k\to\infty} x^k = x$. As C is compact, there is $a \in C$ and a subsequence $(a^k)_{k\in\mathbb{N}}$ of $(x^k)_{k\in\mathbb{N}}$ such that $a = \lim_{k\to\infty} a^k$. The continuity of f yields $f(a) = \lim_{k\to\infty} f(a^k) = \lim_{k\to\infty} y^k = y = f(x)$ since $(f(a^k))_{k\in\mathbb{N}}$ is a subsequence of $(y^k)_{k\in\mathbb{N}}$. It now follows that $a = x$ since f is one-to-one. The same argument shows that *every* convergent subsequence of $(x^k)_{k\in\mathbb{N}}$ has to converge to x. If $(x^k)_{k\in\mathbb{N}}$ did not converge to x, then there had to be some $\epsilon > 0$ such that infinitely man x^k are not in $B_\epsilon(x)$. However, the compactness of C would provide a convergent subsequence whose limit could not be x, in contradiction to x having to be the limit of all convergent subsequences of $(x^k)_{k\in\mathbb{N}}$. ■

3.3 Taylor's Theorem

We begin with Taylor's theorem for one-dimensional functions. For proving its form with so-called Lagrange form of the remainder term, we will use the following Th. 3.11, which is a consequence of the one-dimensional mean value theorem in its generalized version [Phi15a, Th. 9.22].

Theorem 3.11. *Let $a, b \in \mathbb{R}$, $a \neq b$. Suppose $f, g \in C^{m+1}[a, b]$ for $m \in \mathbb{N}_0$ (i.e. f, g are continuous on $[a, b]$ and all derivatives of f and g up to order $m + 1$ exist in $]a, b[$ and extend continuously to $[a, b]$). Moreover, assume that $g^{(k)}(t) \neq 0$ for each $t \in]a, b[$ and each $k \in \{1, \ldots, m + 1\}$. In addition, assume that $f(a) = g(a) = f^{(k)}(a) = g^{(k)}(a) = 0$ for each $k \in \{1, \ldots, m\}$. Then there is $\theta \in]a, b[$ such that*

$$\frac{f(b)}{g(b)} = \frac{f^{(m+1)}(\theta)}{g^{(m+1)}(\theta)}. \tag{3.1}$$

Proof. From [Phi15a, Th. 9.22], we know

$$\frac{f(b) - f(a)}{g(b) - g(a)} \overset{f(a)=g(a)=0}{=} \frac{f(b)}{g(b)} = \frac{f'(\theta_1)}{g'(\theta_1)} \tag{3.2}$$

for some $\theta_1 \in]a, b[$. An induction then establishes (3.1). ■

Theorem 3.12 (Taylor's Theorem)**.** *Let $I \subseteq \mathbb{R}$ be an open interval and $a, x \in I$, $x \neq a$. If $m \in \mathbb{N}_0$ and $f \in C^{m+1}(I, \mathbb{K})$, then*

$$f(x) = T_m(x, a) + R_m(x, a), \tag{3.3}$$

where

$$T_m(x,a) := \sum_{k=0}^{m} \frac{f^{(k)}(a)}{k!}(x-a)^k = f(a)+f'(a)(x-a)+\frac{f''(a)}{2!}(x-a)^2+\cdots+\frac{f^{(m)}(a)}{m!}(x-a)^m \tag{3.4}$$

is the m*th* Taylor polynomial *and*

$$R_m(x,a) := \int_a^x \frac{(x-t)^m}{m!}\, f^{(m+1)}(t)\,\mathrm{d}t \tag{3.5}$$

is the integral form *of the* remainder term*, where, for* $\mathbb{K}=\mathbb{C}$*, we make use of* [Phi15a, Def. G.2] *that, for* $\mathbb{C}$*-valued* f *and provided both* $\operatorname{Re} f$ *and* $\operatorname{Im} f$ *are Riemann integrable, defines* $\int_I f := \int_I \operatorname{Re} f + i\int_I \operatorname{Im} f$*. For* $\mathbb{K}=\mathbb{R}$*, one can also write the remainder term in* Lagrange form*:*

$$R_m(x,a) = \frac{f^{(m+1)}(\theta)}{(m+1)!}(x-a)^{m+1} \quad \textit{with some suitable } \theta \in]x,a[. \tag{3.6}$$

Proof. The integral form (3.5) of the remainder term we prove by using induction on m: For $m=0$, the assertion is

$$f(x) = f(a) + \int_a^x f'(t)\,\mathrm{d}t\,, \tag{3.7}$$

which holds according to the fundamental theorem of calculus in the form [Phi15a, Th. 10.19(b)] (also see [Phi15a, Th. G.6(b)]). For the induction step, we assume (3.3) holds for fixed $m\in\mathbb{N}_0$ with $R_m(x,a)$ in integral form (3.5) and consider $f\in C^{m+2}(I,\mathbb{K})$. For fixed $x\in I$, we define the function

$$g:\, I \longrightarrow \mathbb{K}, \quad g(t) := \frac{(x-t)^{m+1}}{(m+1)!}\, f^{(m+1)}(t). \tag{3.8}$$

Using the product rule, its derivative is

$$g':\, I \longrightarrow \mathbb{K}, \quad g'(t) = \frac{(x-t)^{m+1}}{(m+1)!}\, f^{(m+2)}(t) - \frac{(x-t)^m}{m!}\, f^{(m+1)}(t). \tag{3.9}$$

Applying the fundamental theorem to g then yields

$$-g(a) = g(x) - g(a) = \int_a^x g'(t)\,\mathrm{d}t \overset{(3.9)}{=} R_{m+1}(x,a) - R_m(x,a), \tag{3.10}$$

with $R_m(x,a)$ and $R_{m+1}(x,a)$ defined according to (3.5). Thus,

$$\begin{aligned} T_{m+1}(x,a) + R_{m+1}(x,a) &\overset{(3.10)}{=} T_m(x,a) + \frac{f^{(m+1)}(a)}{(m+1)!}\,(x-a)^{m+1} + R_m(x,a) - g(a) \\ &= T_m(x,a) + R_m(x,a) \overset{\text{ind. hyp.}}{=} f(x), \end{aligned} \tag{3.11}$$

thereby completing the induction and the proof of (3.5).

It remains to prove the Lagrange form (3.6) of the remainder term for $\mathbb{K} = \mathbb{R}$. It is possible, to deduce the Lagrange form from the integral form. However, here we present the following proof based on Th. 3.11, that does not make use of any integration theory. We define auxiliary functions

$$F : I \longrightarrow \mathbb{R}, \qquad F(t) := f(t) - T_m(t, a), \tag{3.12a}$$

$$G : I \longrightarrow \mathbb{R}, \qquad G(t) := (t-a)^{m+1}. \tag{3.12b}$$

It clearly follows from (3.4) that $T_m^{(k)}(a, a) = f^{(k)}(a)$ for each $k \in \{0, \ldots, m\}$. Thus, $F^{(k)}(a) = 0$ and $G^{(k)}(a) = 0$ for each $k \in \{0, \ldots, m\}$, that means F and G satisfy the hypotheses of Th. 3.11 on $[a, x]$. In the present context, (3.1) takes the form

$$\frac{f(x) - T_m(x, a)}{(x-a)^{m+1}} = \frac{F(x)}{G(x)} = \frac{F^{(m+1)}(\theta)}{G^{(m+1)}(\theta)} = \frac{f^{(m+1)}(\theta) - 0}{(m+1)!} \tag{3.13}$$

for some $\theta \in]x, a[$. As (3.13) is equivalent to (3.3) with $R_m(x, a)$ according to (3.6), we have proved the Lagrange form of the remainder term. ■

Remark 3.13. The importance of Taylor's Th. 3.12 does not lie in the decomposition $f = T_m + R_m$, which can be accomplished simply by defining $R_m := f - T_m$. The importance lies rather in the specific formulas for the remainder term.

—

We will now extend Taylor's theorem to higher dimensions by means of the chain rule. First, we need to introduce some notation.

Notation 3.14. In the context of Taylor's theorem, we need to consider directional derivatives of higher order. In this context, one often uses a slightly different notation than the one we used earlier. Let $n \in \mathbb{N}$ and $h = (h_1, \ldots, h_n) \in \mathbb{R}^n$. If $G \subseteq \mathbb{R}^n$ is open and $f : G \longrightarrow \mathbb{K}$ is differentiable at some $\xi \in G$, then, according to (2.62), we can compute the directional derivative

$$(h\,\nabla)(f)(\xi) := \frac{\partial f}{\partial h}(\xi) = \sum_{j=1}^{n} h_j \partial_j f(\xi) = h_1 \partial_1 f(\xi) + \cdots + h_n \partial_n f(\xi). \tag{3.14}$$

The object $h\,\nabla$ is also called a *differential operator*. If f has all partials of second order at ξ, then we can apply $h\,\nabla$ again to the function in (3.14), obtaining

$$(h\,\nabla)^2(f)(\xi) := (h\,\nabla)(h\,\nabla)(f)(\xi) = \sum_{j=1}^{n} (h\,\nabla)(h_j \partial_j f)(\xi) = \sum_{j,k=1}^{n} h_k h_j \partial_k \partial_j f(\xi). \tag{3.15}$$

Thus, if f has all partials of order k at ξ, $k \in \mathbb{N}$, then an induction yields

$$(h\,\nabla)^k(f)(\xi) = \sum_{j_1, \ldots, j_k = 1}^{n} h_{j_k} \cdots h_{j_1} \partial_{j_k} \cdots \partial_{j_1} f(\xi). \tag{3.16}$$

Finally, it is also useful to define

$$(h\,\nabla)^0(f)(\xi) := f(\xi). \tag{3.17}$$

Theorem 3.15 (Taylor's Theorem)**.** *Let $G \subseteq \mathbb{R}^n$ be open, $n \in \mathbb{N}$, and $f \in C^{m+1}(G, \mathbb{K})$ for some $m \in \mathbb{N}_0$ (i.e. $f : G \longrightarrow \mathbb{K}$ and f has continuous partials up to order $m+1$). Let $\xi \in G$ and $h \in \mathbb{R}^n$ such that the line segment $S_{\xi,\xi+h}$ between ξ and $\xi + h$ is a subset of G. Then the following formula, also known as Taylor's formula, holds:*

$$\begin{aligned} f(\xi + h) &= \sum_{k=0}^{m} \frac{(h\,\nabla)^k(f)(\xi)}{k!} + R_m(\xi) \\ &= f(\xi) + \frac{(h\,\nabla)(f)(\xi)}{1!} + \frac{(h\,\nabla)^2(f)(\xi)}{2!} + \cdots + \frac{(h\,\nabla)^m(f)(\xi)}{m!} + R_m(\xi), \end{aligned} \tag{3.18}$$

where, similar to the one-dimensional case,

$$R_m(\xi) := \int_0^1 \frac{(1-t)^m}{m!}\,(h\,\nabla)^{m+1}(f)(\xi + th)\,\mathrm{d}t \tag{3.19}$$

is the integral form *of the* remainder term*. Also similar to the one-dimensional case, if $\mathbb{K} = \mathbb{R}$, then there is $\theta \in]0,1[$ such that*

$$R_m(\xi) = \frac{(h\,\nabla)^{m+1}(f)(\xi + \theta h)}{(m+1)!}, \tag{3.20}$$

called the Lagrange form *of the remainder term.*

Proof. Since $S_{\xi,\xi+h} \subseteq G$ and G is open, there is $\epsilon > 0$ such that we can consider the auxiliary function

$$\phi :]-\epsilon, 1+\epsilon[\longrightarrow \mathbb{K}, \quad \phi(t) := f(\xi + th). \tag{3.21}$$

This definition immediately implies $\phi(0) = f(\xi)$ and $\phi(1) = f(\xi + h)$. We can apply the chain rule to get

$$\phi'(t) = \nabla f(\xi + th) \cdot h = (h\,\nabla)(f)(\xi + th), \tag{3.22}$$

using the notation from (3.14). Since $f \in C^{m+1}(G, \mathbb{K})$, we can use an induction to get

$$\phi^{(k)}(t) = (h\,\nabla)^k(f)(\xi + th). \tag{3.23}$$

Applying the one-dimensional form of Taylor's theorem (i.e. Th. 3.12) with the remainder term in integral form to ϕ with $x = 1$ and $a = 0$ together with (3.23) yields

$$\begin{aligned} f(\xi + h) &= \phi(1) \\ &= \phi(0) + \phi'(0)(1-0) + \frac{\phi''(0)}{2!}(1-0)^2 + \cdots + \frac{\phi^{(m)}(0)}{m!}(1-0)^m \\ &\quad + \int_0^1 \frac{(1-t)^m}{m!}\,\phi^{(m+1)}(t)\,\mathrm{d}t \\ &= f(\xi) + \frac{(h\,\nabla)(f)(\xi)}{1!} + \frac{(h\,\nabla)^2(f)(\xi)}{2!} + \cdots + \frac{(h\,\nabla)^m(f)(\xi)}{m!} \\ &\quad + \int_0^1 \frac{(1-t)^m}{m!}\,(h\,\nabla)^{m+1}(f)(\xi + th)\,\mathrm{d}t\,, \end{aligned} \tag{3.24}$$

which is precisely (3.18) with $R_m(\xi)$ in the form (3.19).

To prove the Lagrange form of the remainder term, we restate (3.24), this time applying Th. 3.12 to ϕ with the remainder term in Lagrange form, yielding

$$\begin{aligned} f(\xi+h) &= \sum_{k=0}^{m} \frac{(h\,\nabla)^k(f)(\xi)}{k!} + \frac{\phi^{(m+1)}(\theta)}{(m+1)!}(1-0)^{m+1} \\ &= \sum_{k=0}^{m} \frac{(h\,\nabla)^k(f)(\xi)}{k!} + \frac{(h\,\nabla)^{m+1}(f)(\xi+\theta h)}{(m+1)!} \end{aligned}$$

for some suitable $\theta \in]0,1[$, thereby completing the proof. ■

Example 3.16. Let us write Taylor's formula (3.18) explicitly for the function

$$f : \mathbb{R}^2 \longrightarrow \mathbb{R}, \quad f(x,y) := \sin(xy) \tag{3.25}$$

for $m = 1$ and for $\xi = (0,0)$. Here, we have for the gradient

$$\nabla f(x,y) = \Big(y\cos(xy),\, x\cos(xy)\Big) \tag{3.26}$$

and for the Hessian matrix of second order partials

$$\begin{aligned} H_f(x,y) &= \begin{pmatrix} \partial_x\partial_x f(x,y) & \partial_x\partial_y f(x,y) \\ \partial_y\partial_x f(x,y) & \partial_y\partial_y f(x,y) \end{pmatrix} \\ &= \begin{pmatrix} -y^2\sin(xy) & \cos(xy) - xy\sin(xy) \\ \cos(xy) - xy\sin(xy) & -x^2\sin(xy) \end{pmatrix}. \end{aligned} \tag{3.27}$$

For $h = (h_1, h_2) \in \mathbb{R}^2$, we obtain

$$\begin{aligned} &f(h) = \\ &\frac{-h_1^2\theta^2h_2^2\sin(\theta^2h_1h_2) + 2h_1h_2\cos(\theta^2h_1h_2) - 2h_1^2h_2^2\theta^2\sin(\theta^2h_1h_2) - h_2^2\theta^2h_1^2\sin(\theta^2h_1h_2)}{2!} \\ &= -2h_1^2h_2^2\theta^2\sin(\theta^2h_1h_2) + h_1h_2\cos(\theta^2h_1h_2) \end{aligned} \tag{3.28}$$

for some suitable $0 < \theta < 1$.

3.4 Quadratic Forms

Before we get to the quadratic forms, we briefly need to consider the Euclidean norm of matrices.

Notation 3.17. Let $A = (a_{kl})_{(k,l)\in\{1,\dots,m\}\times\{1,\dots,n\}}$ a real $m \times n$ matrix, $m, n \in \mathbb{N}$. We introduce the quantity

$$\|A\|_{\text{HS}} := \sqrt{\sum_{k=1}^{m}\sum_{l=1}^{n} a_{kl}^2}, \tag{3.29}$$

called the *Hilbert-Schmidt* norm or the *Frobenius* norm of A. Thus, $\|A\|_{\mathrm{HS}}$ is the Euclidean norm of A if we consider A as an element of $\mathbb{R}^{mn}$. *Caveat*: For $m, n > 1$, the Hilbert-Schmidt norm is *not!* the operator norm of A with respect to the Euclidean norms on $\mathbb{R}^m$ and $\mathbb{R}^n$ – it is actually not an operator norm at all (see [Phi15b, Ex. B.9]). We could actually use the mentioned operator norm in the following and everything would work just the same (since (3.30) also holds for the operator norm) – the reason we prefer the Hilbert-Schmidt norm here, is that it is much easier to compute and, thus, less abstract.

Lemma 3.18. *Let $A = (a_{kl})_{(k,l)\in\{1,\dots,m\}\times\{1,\dots,n\}}$ a real $m \times n$ matrix, $m, n \in \mathbb{N}$. Then, for each $x \in \mathbb{R}^n$, it holds that*

$$\|Ax\|_2 \leq \|A\|_{\mathrm{HS}} \, \|x\|_2. \tag{3.30}$$

Proof. This follows easily from the Cauchy-Schwarz inequality. For each $k \in \{1, \dots, m\}$, let $a_k := (a_{k1}, \dots, a_{kn})$ denote the kth row vector of the matrix A. Then one computes

$$\|Ax\|_2 = \sqrt{\sum_{k=1}^{m} \left(\sum_{l=1}^{n} a_{kl} x_l \right)^2} \leq \sqrt{\sum_{k=1}^{m} \|a_k\|_2^2 \, \|x\|_2^2} = \|A\|_{\mathrm{HS}} \, \|x\|_2, \tag{3.31}$$

thereby establishing the case. ∎

Definition 3.19. Let $n \in \mathbb{N}$. A *quadratic form* is a map

$$Q_A : \mathbb{R}^n \longrightarrow \mathbb{R}, \quad Q_A(x) := x^{\mathrm{t}} A x = \sum_{k,l=1}^{n} a_{kl} x_k x_l, \tag{3.32}$$

where x^{t} denotes the transpose of x, and $A = (a_{kl})_{k,l=1}^n$ is a symmetric real $n \times n$-matrix, i.e. a quadratic real matrix with $a_{kl} = a_{lk}$.

Remark 3.20. Each quadratic form is a polynomial and, thus, continuous by Th. 1.65. Moreover, if $\lambda \in \mathbb{R}$ and A and B are symmetric real $n \times n$-matrices, then λA and $A + B$ are also symmetric real $n \times n$-matrices, and $Q_{\lambda A} = \lambda Q_A$ as well as $Q_{A+B} = Q_A + Q_B$, showing, in particular, that the symmetric real $n \times n$-matrices form a real vector space and that the quadratic forms also form a real vector space.

Example 3.21. If $G \subseteq \mathbb{R}^n$ is open and $f : G \longrightarrow \mathbb{R}$ is C^2, then, for each $\xi \in G$, the Hessian matrix

$$H_f(\xi) = \Big(\partial_k \partial_l f(\xi) \Big)_{k,l=1}^{n} \tag{3.33}$$

is symmetric, i.e. $Q_{H_f(\xi)} : \mathbb{R}^n \longrightarrow \mathbb{R}$ is a quadratic form.

Lemma 3.22. *Let $A = (a_{kl})_{k,l=1}^n$ is a symmetric real $n \times n$-matrix, $n \in \mathbb{N}$, and let Q_A be the corresponding quadratic form.*

(a) *Q_A is* homogeneous of degree 2, *i.e.*

$$Q_A(\lambda x) = \lambda^2 \, Q_A(x) \quad \textit{for each } x \in \mathbb{R}^n \textit{ and each } \lambda \in \mathbb{R}.$$

(b) *For each $\alpha \in \mathbb{R}$, the following statements are equivalent:*

(i) $Q_A(x) \geq \alpha \|x\|_2^2$ for all $x \in \mathbb{R}^n$.

(ii) $Q_A(x) \geq \alpha$ for all $x \in \mathbb{R}^n$ with $\|x\|_2 = 1$.

(c) *For each $x \in \mathbb{R}^n$:*

$$|Q_A(x)| \leq \|A\|_{\mathrm{HS}} \, \|x\|_2^2.$$

Proof. (a) is an immediate consequence of (3.32).

(b): That (i) implies (ii) is trivial, since (ii) is a special case of (i). It remains to show that (ii) implies (i). For $x = 0$, one has $0 = Q_A(x) = \alpha \|x\|_2^2$, so let $x \neq 0$ and assume (ii). Then one obtains

$$Q_A(x) = Q_A\left(\|x\|_2 \, \frac{x}{\|x\|_2}\right) \overset{\text{(a)}}{=} \|x\|_2^2 \, Q_A\left(\frac{x}{\|x\|_2}\right) \geq \alpha \|x\|_2^2, \tag{3.34}$$

proving (i).

(c): Let $x \in \mathbb{R}^n$. Since $Q_A(x) = x \cdot (Ax)$, the Cauchy-Schwarz inequality yields $|Q_A(x)| \leq \|Ax\|_2 \|x\|_2$, and (3.30) then implies (c). ∎

Definition 3.23. Let $A = (a_{kl})_{k,l=1}^n$ is a symmetric real $n \times n$-matrix, $n \in \mathbb{N}$, and let Q_A be the corresponding quadratic form.

(a) A and Q_A are called *positive definite* if, and only if, $Q_A(x) > 0$ for every $0 \neq x \in \mathbb{R}^n$.

(b) A and Q_A are called *positive semidefinite* if, and only if, $Q_A(x) \geq 0$ for every $x \in \mathbb{R}^n$.

(c) A and Q_A are called *negative definite* if, and only if, $Q_A(x) < 0$ for every $0 \neq x \in \mathbb{R}^n$.

(d) A and Q_A are called *negative semidefinite* if, and only if, $Q_A(x) \leq 0$ for every $x \in \mathbb{R}^n$.

(e) A and Q_A are called *indefinite* if, and only if, they are neither positive semidefinite nor negative semidefinite, i.e. if, and only if, there exist $a, b \in \mathbb{R}^n$ with $Q_A(a) > 0$ and $Q_A(b) < 0$.

Example 3.24. Let $n = 2$ and consider the real symmetric matrix $A = \begin{pmatrix} a & b \\ b & c \end{pmatrix}$. One then obtains

$$Q_A : \mathbb{R}^2 \longrightarrow \mathbb{R}, \quad Q_A(x, y) = ax^2 + 2bxy + cy^2. \tag{3.35}$$

One can now use the value of $\det A = ac - b^2$, which is also called the *discriminant* of Q_A, to determine the definiteness of A. This is due to the following identity, that holds for each $(x, y) \in \mathbb{R}^2$:

$$aQ_A(x, y) = a(ax^2 + 2bxy + cy^2) = (ax + by)^2 + (\det A)y^2. \tag{3.36}$$

One obtains the following cases:

$\det A > 0$: This implies $a \neq 0$. Then (3.36) provides:

$$\begin{aligned} a > 0 \quad &\Leftrightarrow \quad Q_A \text{ positive definite,} \\ a < 0 \quad &\Leftrightarrow \quad Q_A \text{ negative definite.} \end{aligned}$$

$\det A < 0$: In this case, we claim:

$$Q_A \text{ is indefinite.}$$

To verify this claim, first consider $a > 0$. Then $Q_A(1,0) = a > 0$ and, according to (3.36), $Q_A(-b/a, 1) = (\det A)/a < 0$, showing that Q_A is indefinite. Now let $a < 0$. Then $Q_A(1,0) = a < 0$ and, according to (3.36), $Q_A(-b/a, 1) = (\det A)/a > 0$, again showing that Q_A is indefinite. Finally, let $a = 0$. Then $\det A < 0$ implies $b \neq 0$. If $c > 0$, then $Q_A(0,1) = c > 0$ and $Q_A(1/(2b), -1/(2c)) = -1/(2c) + 1/(4c) = -1/(2c) < 0$, i.e. Q_A is indefinite. If $c < 0$, then $Q_A(0,1) = c < 0$ and $Q_A(1/(2b), -1/(2c)) = -1/(2c) + 1/(4c) = -1/(2c) > 0$, i.e. Q_A is again indefinite. If $c = 0$, then $Q_A(1/(2b), 1) = 1$ and $Q_A(1/(2b), -1) = -1$ and Q_A is indefinite also in this last case.

$\det A = 0$: Here, we claim:

$$\begin{aligned} a > 0 \text{ or } (a = 0 \text{ and } c \geq 0) \quad &\Leftrightarrow \quad Q_A \text{ positive semidefinite,} \\ a < 0 \text{ or } (a = 0 \text{ and } c \leq 0) \quad &\Leftrightarrow \quad Q_A \text{ negative semidefinite.} \end{aligned}$$

Once again, for the proof, we need to distinguish the different possible cases. If $a > 0$, then $Q_A(x,y) = (ax + by)^2/a \geq 0$, i.e. Q_A is positive semidefinite. If $a < 0$, then $Q_A(x,y) = (ax + by)^2/a \leq 0$, i.e. Q_A is negative semidefinite. Now let $a = 0$. Then $\det A = 0$ implies $b = 0$. Thus, $Q_A(x,y) = cy^2$, i.e. Q_A is positive semidefinite for $c \geq 0$ and negative semidefinite for $c \leq 0$.

Proposition 3.25. *Let $A = (a_{kl})_{k,l=1}^n$ is a symmetric real $n \times n$-matrix, $n \in \mathbb{N}$, and let Q_A be the corresponding quadratic form.*

(a) *A and Q_A are positive definite if, and only if, there exists $\alpha > 0$ such that*

$$Q_A(x) \geq \alpha > 0 \quad \text{for each } x \in \mathbb{R}^n \text{ with } \|x\|_2 = 1. \tag{3.37a}$$

Analogously, A and Q_A are negative definite if, and only if, there exists $\alpha < 0$ such that

$$Q_A(x) \leq \alpha < 0 \quad \text{for each } x \in \mathbb{R}^n \text{ with } \|x\|_2 = 1. \tag{3.37b}$$

(b) *If A and Q_A are positive definite (respectively negative definite, or indefinite), then there exists $\epsilon > 0$ such that each symmetric real $n \times n$ matrix B with $\|A - B\|_{\mathrm{HS}} < \epsilon$ is also positive definite (respectively negative definite, or indefinite).*

(c) *If A and Q_A are indefinite, then there exists $\epsilon > 0$ and $a, b \in \mathbb{R}^n$ with $\|a\|_2 = \|b\|_2 = 1$ such that, for each symmetric real $n \times n$ matrix B with $\|A - B\|_{\mathrm{HS}} < \epsilon$ and each $0 \neq \lambda \in \mathbb{R}$, it holds that $Q_B(\lambda a) > 0$ and $Q_B(\lambda b) < 0$.*

Proof. (a): We consider the positive definite case; the negative definite case is proved completely analogously. First note that (3.37a) implies that A and Q_A are positive definite according to Lem. 3.22(b). Conversely, assume that A and Q_A are positive definite. The 1-sphere $S_1(0) = \{x \in \mathbb{R}^n : \|x\|_2 = 1\}$ is a closed and bounded subset of $\mathbb{R}^n$ and, hence, compact. Since Q_A is continuous, it must assume its min on $S_1(0)$ according to Th. 3.8, i.e. there is $\alpha \in \mathbb{R}$ and $x_\alpha \in S_1(0)$ such that $Q_A(x_\alpha) = \alpha$ and $Q_A(x) \geq \alpha$ for each $x \in \mathbb{R}^n$ with $\|x\|_2 = 1$. Since Q_A is positive definite, $\alpha > 0$, proving (3.37a).

(b) and (c): We begin by employing (3.37a) to show (b) for A and Q_A being positive definite (employing (3.37b), the case of A and Q_A being negative definite can be treated completely analogously). If A and Q_A are positive definite, then there is $\alpha > 0$ such that (3.37a) holds. Choose $\epsilon := \alpha/2$. If B is a symmetric real $n \times n$ matrix with $\|A - B\|_{\text{HS}} < \epsilon$, then, using Lem. 3.22(c), for each $x \in \mathbb{R}^n$ with $\|x\|_2 = 1$:

$$|Q_A(x) - Q_B(x)| = |Q_{A-B}(x)| \leq \|A - B\|_{\text{HS}} < \epsilon = \frac{\alpha}{2}. \tag{3.38}$$

Since $Q_A(x) \geq \alpha > 0$, this implies $Q_B(x) \geq \alpha/2 > 0$ for each $x \in \mathbb{R}^n$ with $\|x\|_2 = 1$. Due to (a), this proves that B is positive definite. Now consider the case that A and Q_A are indefinite. Then there are $0 \neq a, b \in \mathbb{R}^n$ such that $Q_A(a) > 0$ and $Q_A(b) < 0$. By normalizing and using Lem. 3.22(a), one can even additionally assume $\|a\|_2 = \|b\|_2 = 1$. Set $\alpha := \min\{Q_A(a), |Q_A(b)|\}$. Then $\alpha > 0$. If $\epsilon := \alpha/2$ and B is a symmetric real $n \times n$ matrix with $\|A - B\|_{\text{HS}} < \epsilon$, then, as above, (3.38) holds for each $x \in \mathbb{R}^n$ with $\|x\|_2 = 1$. In particular, $Q_B(a) \geq \alpha/2 > 0$ and $Q_B(b) \leq -\alpha/2 < 0$, showing that Q_B is indefinite, concluding the proof of (b). To complete the proof of (c) as well, it merely remains to remark that, for each $0 \neq \lambda \in \mathbb{R}$, one has $Q_B(\lambda a) \geq \lambda^2\alpha/2 > 0$ and $Q_B(\lambda b) \leq -\lambda^2\alpha/2 < 0$. ■

3.5 Extreme Values and Stationary Points of Differentiable Functions

Definition 3.26. Let $G \subseteq \mathbb{R}^n$, $n \in \mathbb{N}$, $f : G \longrightarrow \mathbb{K}$, and let ξ be an interior point of G. If all first partials of f exist in ξ, then ξ is called a *stationary* or *critical point* of f if, and only if,

$$\nabla f(\xi) = 0. \tag{3.39}$$

—

The following Th. 3.27 generalizes [Phi15a, Th. 9.14] to functions defined on subsets of $\mathbb{R}^n$:

Theorem 3.27. *Let $G \subseteq \mathbb{R}^n$, $n \in \mathbb{N}$, $f : G \longrightarrow \mathbb{R}$, and let ξ be an interior point of G. If all first partials of f exist in ξ and f has a local min or max at ξ, then ξ is a stationary point of f, i.e. $\nabla f(\xi) = 0$.*

Proof. Since ξ is an interior point of G and since f has a local min or max at ξ, there is $\epsilon > 0$ such that $B_\epsilon(\xi) \subseteq G$ and such that $f(\xi) \leq f(x)$ for each $x \in B_\epsilon(\xi)$ or such that $f(\xi) \geq f(x)$ for each $x \in B_\epsilon(\xi)$. Let $j \in \{1, \ldots, n\}$. Then there is $\delta > 0$ such that $(\xi_1, \ldots, \xi_{j-1}, t, \xi_{j+1}, \ldots, \xi_n) \in B_\epsilon(\xi)$ for each $t \in]\xi_j - \delta, \xi_j + \delta[$. Thus, the one-dimensional function $g :]\xi_j - \delta, \xi_j + \delta[\longrightarrow \mathbb{R}$, $g(t) := f(\xi_1, \ldots, \xi_{j-1}, t, \xi_{j+1}, \ldots, \xi_n)$, has a local min or max at ξ_j, and, since $\partial_j f(\xi)$ exists, g is differentiable in ξ_j, implying $0 = g'(\xi_j) = \partial_j f(\xi)$ according to [Phi15a, Th. 9.14]. Since $j \in \{1, \ldots, n\}$ was arbitrary, $\nabla f(\xi) = 0$. ■

One already knows from simple one-dimensional examples such as $f : \mathbb{R} \longrightarrow \mathbb{R}$, $f(x) := x^3$ and $\xi = 0$ that $\nabla f(\xi) = 0$ is not a sufficient condition for f to have a local extreme value at ξ. However, the following Th. 3.28 does provide such sufficient conditions.

Theorem 3.28. *Let $G \subseteq \mathbb{R}^n$ be open, $n \in \mathbb{N}$, $f : G \longrightarrow \mathbb{R}$, $f \in C^2(G)$, and let $\xi \in G$ be a stationary point of f. Then, in the following cases, one can use the Hessian matrix $H_f(\xi)$ to determine if f has a local extreme value at ξ:*

$$H_f(\xi) \text{ positive definite} \quad \Rightarrow \quad f \text{ has a strict local min at } \xi, \tag{3.40a}$$

$$H_f(\xi) \text{ negative definite} \quad \Rightarrow \quad f \text{ has a strict local max at } \xi, \tag{3.40b}$$

$$H_f(\xi) \text{ indefinite} \quad \Rightarrow \quad f \text{ does not have a local extreme value at } \xi. \tag{3.40c}$$

Proof. Since G is open, there is $\epsilon > 0$ such that $\xi + h \in G$ for each $h \in \mathbb{R}^n$ with $\|h\|_2 < \epsilon$. For each such h, by an application of Taylor's Th. 3.15 with $m = 1$, we obtain the existence of $\theta \in]0, 1[$ satisfying

$$\begin{aligned} f(\xi + h) &= f(\xi) + h \cdot \nabla f(\xi) + \frac{1}{2} \sum_{k,l=1}^{n} \partial_k \partial_l f(\xi + \theta h) h_k h_l \\ &= f(\xi) + \frac{h^{\mathrm{t}} H_f(\xi + \theta h) h}{2} = f(\xi) + \frac{Q_{H_f(\xi+\theta h)}(h)}{2}. \end{aligned} \tag{3.41}$$

Rewriting (3.41), one gets

$$f(\xi + h) - f(\xi) = \frac{Q_{H_f(\xi+\theta h)}(h)}{2}. \tag{3.42}$$

Note that the assumed continuity of the functions $\partial_k \partial_l f : G \longrightarrow \mathbb{R}$ ($k, l \in \{1, \ldots, n\}$) implies the continuity of $H_f : G \longrightarrow \mathbb{R}^{n^2}$, $x \mapsto H_f(x)$ (the $\partial_k \partial_l f$ are the coordinate functions of H_f). Thus, if $H_f(\xi)$ is positive definite, then, by Prop. 3.25(b), there is $\delta > 0$ such that $\|h\|_2 < \epsilon$ and $\|H_f(\xi) - H_f(\xi + \theta h)\|_{\mathrm{HS}} < \delta$ imply that $H_f(\xi + \theta h)$ is also positive definite. Moreover, the continuity of H_f means that there exists $0 < \alpha < \epsilon$ such that $\|h\|_2 < \alpha$ implies $\|H_f(\xi) - H_f(\xi + \theta h)\|_{\mathrm{HS}} < \delta$ for each $\theta \in]0, 1[$. For such h, the right-hand side of (3.42) must be positive, showing that f has a strict local min at ξ ($f(\xi) < f(x)$ for each $x \in B_{\alpha, \|\cdot\|_2}(\xi) \setminus \{\xi\}$). For $H_f(\xi)$ being negative definite, an analogous argument shows that f has a strict max at ξ. Similarly, if $H_f(\xi)$ is indefinite, then, by Prop. 3.25(c), there is $\delta > 0$ and $a, b \in \mathbb{R}^n$ with $\|a\|_2 = \|b\|_2 = 1$ such that, $\|h\|_2 < \epsilon$ and $\|H_f(\xi) - H_f(\xi + \theta h)\|_{\mathrm{HS}} < \delta$ imply that $Q_{H_f(\xi+\theta h)}(\lambda a) > 0$ and $Q_{H_f(\xi+\theta h)}(\lambda b) < 0$ for each $0 \neq \lambda \in \mathbb{R}$. The continuity of H_f provides some

$0 < \alpha < \epsilon$ such that $\|h\|_2 < \alpha$ implies $\|H_f(\xi) - H_f(\xi + \theta h)\|_{\text{HS}} < \delta$ for each $\theta \in]0,1[$. For each $0 < \lambda < \alpha$, we get $\|\lambda a\|_2 < \alpha$ and $\|\lambda b\|_2 < \alpha$, such that (3.42) implies $f(\xi + \lambda b) < f(\xi) < f(\xi + \lambda a)$, i.e. f has neither a local min nor a local max at ξ. ■

Example 3.29. Consider the case $n = 2$, i.e. the case of a C^2 function $f : G \longrightarrow \mathbb{R}$, G being an open subset of $\mathbb{R}^2$. Let $(x_0, y_0) \in G$ be a stationary point of f. Then, according to Example 3.24, the definiteness of the Hessian matrix $H_f(x_0, y_0)$ is determined by the sign of

$$\det H_f(x_0, y_0) = \partial_x \partial_x f(x_0, y_0) \partial_y \partial_y f(x_0, y_0) - \big(\partial_x \partial_y f(x_0, y_0)\big)^2 \tag{3.43}$$

(which, by definition, is the same as the discriminant of the corresponding quadratic form $Q_{H_f(x_0,y_0)}$). If $\det H_f(x_0, y_0) > 0$, then Th. 3.28 tells us that f has a strict local extreme value at (x_0, y_0): If $\partial_x \partial_x f(x_0, y_0) > 0$, then, by Example 3.24, $H_f(x_0, y_0)$ is positive definite and f has as strict local min at (x_0, y_0); if $\partial_x \partial_x f(x_0, y_0) < 0$, then, by Example 3.24, $H_f(x_0, y_0)$ is negative definite and f has as strict local max at (x_0, y_0). If $\det H_f(x_0, y_0) < 0$, then $H_f(x_0, y_0)$ is indefinite according to Example 3.24, and Th. 3.28 yields that f has neither a local max nor a local min at (x_0, y_0). Such a stationary point, where $\det H_f(x_0, y_0) < 0$, is called a *saddle point* – in a neighborhood of such a point, the graph of f is shaped like a saddle. In the remaining case, namely $\det H_f(x_0, y_0) = 0$, one knows from Example 3.24 that $H_f(x_0, y_0)$ is positive semidefinite or negative semidefinite. In this case, Th. 3.28 does not provide any information, i.e., without further investigation, one can not say if f does or does not have an extreme value at (x_0, y_0).

Let us look at two concrete cases:

(a) Consider $f : \mathbb{R}^2 \longrightarrow \mathbb{R}$, $f(x, y) := x^2 + y^2$. Then $\nabla f(x, y) = (2x, 2y)$ and $H_f(x, y) = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$. Thus, $(0, 0)$ is the only stationary point of f. Since $\det H_f(0, 0) = 4 > 0$, f has a strict local min at $(0, 0)$ and this is the only point, where f has a local extreme value. Moreover, since $f(x, y) > 0$ for $(x, y) \neq (0, 0)$, f also has a strict global min at $(0, 0)$.

(b) Consider $f : \mathbb{R}^2 \longrightarrow \mathbb{R}$, $f(x, y) := x^2 - y^2$. Then $\nabla f(x, y) = (2x, -2y)$ and $H_f(x, y) = \begin{pmatrix} 2 & 0 \\ 0 & -2 \end{pmatrix}$. Thus, $(0, 0)$ is the only stationary point of f. Here, one has $\det H_f(0, 0) = -4 < 0$, i.e. f does not have a local min or max at $(0, 0)$ (or anywhere else). Thus, $(0, 0)$ is an example of a saddle point.

—

Let us summarize the general strategy for determining extreme values of differentiable functions f defined on a set G: One starts by seeking all stationary points of f, that means the points ξ, where $\nabla f(\xi) = 0$. Every min or max of f that lies in the interior of G must be included in the set of stationary points. To investigate if a stationary point is, indeed, a max or a min, one will compute the Hessian matrix H_f at this point, and one will determine the definiteness properties of H_f. Then one can use Th. 3.28 to

decide if the stationary point is a max, a min, or neither, except for cases, where H_f is only (positive or negative) semidefinite, in which case Th. 3.28 does not help and one has to resort to other means (which can be difficult). As is already know from functions defined on $G \subseteq \mathbb{R}$, one also has to investigate the behavior of f at the boundary of G if one wants to find out if one of the local extrema is actually a global extremum. Moreover, if f is defined on ∂G, then ∂G might contain further local extrema of f.

4 The Riemann Integral on Intervals in $\mathbb{R}^n$

4.1 Definition and Simple Properties

In generalization of [Phi15a, Sec. 10], we will define integrals for suitable functions $f : A \longrightarrow \mathbb{R}$, where A is a subset of $\mathbb{R}^n$. As in [Phi15a, Sec. 10], we will restrict ourselves to considering the Riemann integral for $\mathbb{R}$-valued functions, pointing out that, by applying the theory to the $\mathbb{R}$-valued functions $\operatorname{Re} f$ and $\operatorname{Im} f$, many results can be extended to $\mathbb{C}$-valued functions f. The details are carried out in Appendix D. When stating some important $\mathbb{R}$-valued result that has a $\mathbb{C}$-valued analogue, we will usually provide the corresponding reference to the Appendix.

In generalization of [Phi15a, Sec. 10], given a nonnegative function $f : A \longrightarrow \mathbb{R}_0^+$, where A is a subset of $\mathbb{R}^n$, we aim to compute the $(n+1)$-dimensional volume $\int_A f$ of the set "under the graph" of f, i.e. of the set

$$\big\{(x_1, \ldots, x_n, x_{n+1}) \in \mathbb{R}^{n+1} : (x_1, \ldots, x_n) \in A \text{ and } 0 \leq x_{n+1} \leq f(x_1, \ldots, x_n)\big\}. \tag{4.1}$$

This $(n+1)$-dimensional volume $\int_A f$ (if it exists) will be called the *integral* of f over A. Moreover, for functions $f : A \longrightarrow \mathbb{R}$ that are not necessarily nonnegative, we would like to count volumes of sets of the form (4.1) (which are below the graph of f and above the set $A \cong \big\{(x_1, \ldots, x_n, 0) \in \mathbb{R}^{n+1} : (x_1, \ldots, x_n) \in A\big\} \subseteq \mathbb{R}^{n+1}$) with a positive sign, whereas we would like to count volumes of sets above the graph of f and below the set A with a negative sign. In other words, making use of the positive and negative parts $f^+ = \max(f, 0)$ and $f^- = \max(-f, 0)$ of $f = f^+ - f^-$, we would like our integral to satisfy

$$\int_A f = \int_A f^+ - \int_A f^-. \tag{4.2}$$

Difficulties arise from the fact that both the function f and the set A can be extremely complicated. To avoid dealing with complicated sets A, we restrict ourselves to the situation of integrals over closed and bounded intervals in $\mathbb{R}^n$, i.e. to integrals over sets of the form $A = [x, y]$ as defined in (1.5b). Moreover, we will also restrict ourselves to bounded functions $f \in B(A, \mathbb{R})$ (cf. Ex. 1.22(c)).

As in [Phi15a, Sec. 10.1], the basic idea for the definition of the Riemann integral $\int_A f$ is rather simple: Decompose the set A into small pieces $A_1, \ldots, A_N$ and approximate $\int_A f$ by the finite sum $\sum_{j=1}^N f(a_j)|A_j|$, where $a_j \in A_j$ and $|A_j|$ denotes the volume of the set A_j. Define $\int_A f$ as the limit of such sums as the size of the A_j tends to zero (if

the limit exists). However, to carry out this idea precisely and rigorously is somewhat cumbersome, for example due to the required notation.

As stated before, we will assume that A is a closed and bounded interval, and we will choose the A_j to be closed finite intervals as well. To emphasize that we are dealing with intervals, in the following, we will prefer to use the symbol I instead of A. As each n-dimensional interval I is a product of one-dimensional intervals, we will obtain our decompositions of I from decompositions of one-dimensional intervals (the *sides* of I).

Definition 4.1. In generalization of [Phi15a, Def. 10.2], if $a, b \in \mathbb{R}^n$, $n \in \mathbb{N}$, $a \leq b$, and $I := [a, b] = [a_1, b_1] \times \cdots \times [a_n, b_n]$, then we call

$$|I| := \prod_{j=1}^{n}(b_j - a_j) = \prod_{j=1}^{n}|a_j - b_j| = \prod_{j=1}^{n}|I_j| \quad (I_j := [a_j, b_j]), \tag{4.3}$$

the (n-dimensional) *size*, *volume*, or *measure* of I.

Definition 4.2. Recall the notion of partition for a 1-dimensional interval $[a, b] \subseteq \mathbb{R}$ from [Phi15a, Def. 10.3]. We now generalize this notion to n-dimensional intervals, $n \in \mathbb{N}$: Given an interval $I := [a, b] \subseteq \mathbb{R}^n$, $a, b \in \mathbb{R}^n$, $a < b$, i.e. $I = [a, b] = [a_1, b_1] \times \cdots \times [a_n, b_n]$, a *partition* Δ of I is given by (1-dimensional) partitions $\Delta_k = (x_{k,0}, \ldots, x_{k,N_k})$ of $[a_k, b_k]$, $k \in \{1, \ldots, n\}$, $N_k \in \mathbb{N}$. Given such a partition Δ of I, for each $(k_1, \ldots, k_n) \in P(\Delta) := \prod_{k=1}^{n}\{1, \ldots, N_k\}$, define

$$I_{(j_1,\ldots,j_n)} := \prod_{k=1}^{n}[x_{k,j_k-1}, x_{k,j_k}] = [x_{1,j_1-1}, x_{1,j_1}] \times \cdots \times [x_{n,j_n-1}, x_{n,j_n}]. \tag{4.4}$$

The number

$$|\Delta| := \max\big\{|\Delta_k| : k \in \{1, \ldots, n\}\big\}, \tag{4.5}$$

is called the *mesh size* of Δ.

Moreover, if each Δ_k is tagged by $(t_{k,1}, \ldots, t_{k,N_k}) \in \mathbb{R}^{N_k}$ such that $t_{k,j} \in [x_{k,j-1}, x_{k,j}]$ for each $j \in \{1, \ldots, N_j\}$, then Δ is tagged by $(t_p)_{p \in P(\Delta)}$, where

$$t_{(j_1,\ldots,j_n)} := (t_{1,j_1}, \ldots, t_{n,j_n}) \in I_{(j_1,\ldots,j_n)} \quad \text{for each} \quad (j_1, \ldots, j_n) \in P(\Delta). \tag{4.6}$$

Remark 4.3. If Δ is a partition of $I = [a, b] \subseteq \mathbb{R}^n$, $n \in \mathbb{N}$, $a, b \in \mathbb{R}^n$, $a < b$, as in Def. 4.2 above, then

$$I = \bigcup_{p \in P(\Delta)} I_p \tag{4.7}$$

and

$$|I_p \cap I_q| = 0 \quad \text{for each } p, q \in P(\Delta) \text{ such that } p \neq q, \tag{4.8}$$

since, for $p \neq q$, $I_p \cap I_q$ is either empty or it is an interval such that one side consists of precisely one point. Moreover, as a consequence of (4.7) and (4.8):

$$|I| = \sum_{p \in P(\Delta)} |I_p|. \tag{4.9}$$

Definition 4.4. Consider an interval $I := [a, b] \subseteq \mathbb{R}^n$, $n \in \mathbb{N}$, $a, b \in \mathbb{R}^n$, $a < b$, with a partition Δ of I as in Def. 4.2. In generalization of [Phi15a, Def. 10.4], given a function $f : I \longrightarrow \mathbb{R}$ that is bounded, define

$$m_p := m_p(f) := \inf\{f(x) : x \in I_p\}, \quad M_p := M_p(f) := \sup\{f(x) : x \in I_p\}, \tag{4.10}$$

and

$$r(\Delta, f) := \sum_{p \in P(\Delta)} m_p |I_p|, \tag{4.11a}$$

$$R(\Delta, f) := \sum_{p \in P(\Delta)} M_p |I_p|, \tag{4.11b}$$

where $r(\Delta, f)$ is called the *lower Riemann sum* and $R(\Delta, f)$ is called the *upper Riemann sum* associated with Δ and f. If Δ is tagged by $\tau := (t_p)_{p \in P(\Delta)}$, then we also define the *intermediate Riemann sum*

$$\rho(\Delta, f) := \sum_{p \in P(\Delta)} f(t_p) |I_p|. \tag{4.11c}$$

Definition 4.5. Let $I = [a, b] \subseteq \mathbb{R}^n$ be an interval, $a, b \in \mathbb{R}^n$, $n \in \mathbb{N}$, $a < b$, and suppose $f : I \longrightarrow \mathbb{R}$ is bounded.

(a) Define

$$J_*(f, I) := \sup\big\{r(\Delta, f) : \Delta \text{ is a partition of } I\big\}, \tag{4.12a}$$

$$J^*(f, I) := \inf\big\{R(\Delta, f) : \Delta \text{ is a partition of } I\big\}. \tag{4.12b}$$

We call $J_*(f, I)$ the *lower Riemann integral* of f over I and $J^*(f, I)$ the *upper Riemann integral* of f over I.

(b) The function f is called *Riemann integrable* over I if, and only if, $J_*(f, I) = J^*(f, I)$. If f is Riemann integrable over I, then

$$\int_I f(x)\,\mathrm{d}x := \int_I f := J_*(f, I) = J^*(f, I) \tag{4.13}$$

is called the *Riemann integral* of f over I. The set of all functions $f : I \longrightarrow \mathbb{R}$ that are Riemann integrable over I is denoted by $\mathcal{R}(I)$.

Remark 4.6. If I and f are as before, then (4.10) implies

$$m_p(f) = -M_p(-f) \qquad \text{and} \qquad m_p(-f) = -M_p(f), \tag{4.14a}$$

(4.11) implies

$$r(\Delta, f) = -R(\Delta, -f) \qquad \text{and} \qquad r(\Delta, -f) = -R(\Delta, f), \tag{4.14b}$$

and (4.12) implies

$$J_*(f, I) = -J^*(-f, I) \qquad \text{and} \qquad J_*(-f, I) = -J^*(f, I). \tag{4.14c}$$

Example 4.7. Let $I = [a, b] \subseteq \mathbb{R}^n$ be an interval, $a, b \in \mathbb{R}^n$, $n \in \mathbb{N}$, $a < b$.

(a) Analogous to [Phi15a, Ex. 10.7(a)], if $f : I \longrightarrow \mathbb{R}$ is constant, i.e. $f \equiv c$ with $c \in \mathbb{R}$, then $f \in \mathcal{R}(I)$ and

$$\int_I f = c\,|I| : \tag{4.15}$$

We have, for each partition Δ of I,

$$r(\Delta, f) = \sum_{p \in P(\Delta)} m_p |I_p| = c \sum_{p \in P(\Delta)} |I_p| \overset{(4.9)}{=} c\,|I| = \sum_{p \in P(\Delta)} M_p |I_p| = R(\Delta, f), \tag{4.16}$$

proving $J_*(f, I) = c\,|I| = J^*(f, I)$.

(b) An example of a function that is not Riemann integrable is given by the n-dimensional version of the *Dirichlet function* of [Phi15a, Ex. 10.7(b)], i.e. by

$$f : I \longrightarrow \mathbb{R}, \quad f(x) := \begin{cases} 0 & \text{for } x \in I \setminus \mathbb{Q}^n, \\ 1 & \text{for } x \in I \cap \mathbb{Q}^n. \end{cases} \tag{4.17}$$

Since $r(\Delta, f) = 0$ and $R(\Delta, f) = \sum_{p \in P(\Delta)} |I_p| = |I|$ for every partition Δ of I, one obtains $J_*(f, I) = 0 \neq |I| = J^*(f, I)$, showing that $f \notin \mathcal{R}(I)$.

Definition 4.8. Recall the notions of refinement and superposition of partitions of a 1-dimensional interval $[a, b] \subseteq \mathbb{R}$ from [Phi15a, Def. 10.8]. We now generalize both notions to partitions of n-dimensional intervals, $n \in \mathbb{N}$:

(a) If Δ is a partition of $[a, b] \subseteq \mathbb{R}^n$, $n \in \mathbb{N}$, $a, b \in \mathbb{R}^n$, $a < b$, as in Def. 4.2, then another partition Δ' of $[a, b]$ given by partitions Δ'_k of $[a_k, b_k]$, $k \in \{1, \ldots, n\}$, respectively, is called a *refinement* of Δ if, and only if, each Δ'_k is a (1-dimensional) refinement of Δ_k in the sense of [Phi15a, Def. 10.8(a)].

(b) If Δ and Δ' are partitions of $[a, b] \subseteq \mathbb{R}^n$, $n \in \mathbb{N}$, $a, b \in \mathbb{R}^n$, $a < b$, then the *superposition* $\Delta + \Delta'$ is given by the (1-dimensional) superpositions $\Delta_k + \Delta'_k$ of the $[a_k, b_k]$ in the sense of [Phi15a, Def. 10.8(b)]. Note that the superposition of Δ and Δ' is always a common refinement of Δ and Δ'.

Lemma 4.9. *Let $n \in \mathbb{N}$, $a, b \in \mathbb{R}^n$, $a < b$, $I := [a, b]$, and suppose $f : I \longrightarrow \mathbb{R}$ is bounded with $M := \|f\|_{\sup} \in \mathbb{R}_0^+$. Let Δ' be a partition of I and define*

$$\alpha := \sum_{k=1}^{n} \#\big(\nu(\Delta'_k) \setminus \{a_k, b_k\}\big) \tag{4.18}$$

i.e. α is the number of interior nodes that occur in the Δ'_k. Then, for each partition Δ of I, the following holds:

$$r(\Delta, f) \leq r(\Delta + \Delta', f) \leq r(\Delta, f) + 2\,\alpha\, M \phi(I) |\Delta|, \tag{4.19a}$$

$$R(\Delta, f) \geq R(\Delta + \Delta', f) \geq R(\Delta, f) - 2\,\alpha\, M \phi(I) |\Delta|, \tag{4.19b}$$

where

$$\phi(I) := \max\left\{\frac{|I|}{b_k - a_k} : k \in \{1, \dots, n\}\right\} \tag{4.20}$$

is the maximal volume of the $(n-1)$*-dimensional faces of* I*.*

Proof. We carry out the proof of (4.19a) – the proof of (4.19b) can be conducted completely analogous. Consider the case $\alpha = 1$. Then there is a unique $k_0 \in \{1, \dots, n\}$ such that there exists $\xi \in \nu(\Delta'_{k_0}) \setminus \{a_{k_0}, b_{k_0}\}$. If $\xi \in \nu(\Delta_{k_0})$, then $\Delta + \Delta' = \Delta$, and (4.19a) is trivially true. If $\xi \notin \nu(\Delta_{k_0})$, then $x_{k_0,l-1} < \xi < x_{k_0,l}$ for a suitable $l \in \{1, \dots, N_{k_0}\}$. Recalling the notation from Def. 4.2, we let

$$P_l(\Delta) := \big\{(j_1, \dots, j_n) \in P(\Delta) : j_{k_0} = l\big\}, \tag{4.21}$$

and define, for each $(j_1, \dots, j_n) \in P_l(\Delta)$,

$$I'_{j_1,\dots,j_n} := \prod_{k=1}^{k_0-1} [x_{k,j_k-1}, x_{k,j_k}] \times [x_{k_0,l-1}, \xi] \times \prod_{k=k_0+1}^{n} [x_{k,j_k-1}, x_{k,j_k}], \tag{4.22a}$$

$$I''_{j_1,\dots,j_n} := \prod_{k=1}^{k_0-1} [x_{k,j_k-1}, x_{k,j_k}] \times [\xi, x_{k_0,l}] \times \prod_{k=k_0+1}^{n} [x_{k,j_k-1}, x_{k,j_k}], \tag{4.22b}$$

and

$$m'_p := \inf\{f(x) : x \in I'_p\}, \quad m''_p := \inf\{f(x) : x \in I''_p\} \quad \text{for each } p \in P_l(\Delta). \tag{4.23}$$

Then we obtain

$$\begin{aligned} r(\Delta + \Delta', f) - r(\Delta, f) &= \sum_{p \in P_l(\Delta)} \big(m'_p\,|I'_p| + m''_p\,|I''_p| - m_p\,|I_p|\big) \\ &= \sum_{p \in P_l(\Delta)} \big((m'_p - m_p)\,|I'_p| + (m''_p - m_p)\,|I''_p|\big). \end{aligned} \tag{4.24}$$

Together with the observation

$$0 \le m'_p - m_p \le 2M, \quad 0 \le m''_p - m_p \le 2M, \tag{4.25}$$

(4.24) implies

$$\begin{aligned} 0 \le r(\Delta + \Delta', f) - r(\Delta, f) &\le 2M \sum_{p \in P_l(\Delta)} |I_p| = 2M\,(x_{k_0,l} - x_{k_0,l-1})\,\frac{|I|}{b_{k_0} - a_{k_0}} \\ &\le 2M\,|\Delta_{k_0}|\,\phi(I) \le 2M\,|\Delta|\,\phi(I). \end{aligned} \tag{4.26}$$

The general form of (4.19a) now follows by an induction on α. ∎

Theorem 4.10. *Let* $n \in \mathbb{N}$*,* $a, b \in \mathbb{R}^n$*,* $a < b$*,* $I := [a, b]$*, and let* $f : I \longrightarrow \mathbb{R}$ *be bounded.*

(a) *Suppose Δ and Δ' are partitions of I such that Δ' is a refinement of Δ. Then*

$$r(\Delta, f) \le r(\Delta', f), \quad R(\Delta, f) \ge R(\Delta', f). \tag{4.27}$$

(b) *For arbitrary partitions Δ and Δ', the following holds:*

$$r(\Delta, f) \le R(\Delta', f). \tag{4.28}$$

(c) $J_*(f, I) \le J^*(f, I)$.

(d) *For each sequence of partitions $(\Delta^k)_{k\in\mathbb{N}}$ of I such that $\lim_{k\to\infty} |\Delta^k| = 0$, one has*

$$\lim_{k\to\infty} r(\Delta^k, f) = J_*(f, I), \quad \lim_{k\to\infty} R(\Delta^k, f) = J^*(f, I). \tag{4.29}$$

In particular, if $f \in \mathcal{R}(I)$, then

$$\lim_{k\to\infty} r(\Delta^k, f) = \lim_{k\to\infty} R(\Delta^k, f) = \int_I f, \tag{4.30a}$$

and if $f \in \mathcal{R}(I)$ and the Δ^k are tagged, then also

$$\lim_{k\to\infty} \rho(\Delta^k, f) = \int_I f. \tag{4.30b}$$

(e) *If there exists $\alpha \in \mathbb{R}$ such that*

$$\alpha = \lim_{k\to\infty} \rho(\Delta^k, f) \tag{4.31}$$

for each sequence of tagged partitions $(\Delta^k)_{k\in\mathbb{N}}$ of I such that $\lim_{k\to\infty} |\Delta^k| = 0$, then $f \in \mathcal{R}(I)$ and $\alpha = \int_I f$.

Proof. (a): If Δ' is a refinement of Δ, then $\Delta' = \Delta + \Delta'$. Thus, (4.27) is immediate from (4.19).

(b): This also follows from (4.19):

$$r(\Delta, f) \overset{(4.19a)}{\le} r(\Delta + \Delta', f) \overset{(4.11)}{\le} R(\Delta + \Delta', f) \overset{(4.19b)}{\le} R(\Delta', f). \tag{4.32}$$

(c): One just combines (4.12) with (b).

(d): Let $(\Delta^k)_{k\in\mathbb{N}}$ be a sequence of partitions of I such that $\lim_{k\to\infty} |\Delta^k| = 0$, and let Δ' be an arbitrary partition of I with numbers α, M, and $\phi(I)$ defined as in Lem. 4.9. Then, according to (4.19a):

$$r(\Delta^k, f) \le r(\Delta^k + \Delta', f) \le r(\Delta^k, f) + 2\,\alpha\, M\phi(I)|\Delta^k| \quad \text{for each } k \in \mathbb{N}. \tag{4.33}$$

From (b), we conclude the sequence $\big(r(\Delta^k, f)\big)_{k\in\mathbb{N}}$ is bounded. Recall from [Phi15a, Th. 7.27] that each bounded sequence $(t_k)_{k\in\mathbb{N}}$ in $\mathbb{R}$ has a smallest cluster point $t_* \in \mathbb{R}$,

and a largest cluster point $t^* \in \mathbb{R}$. Moreover, by [Phi15a, Prop. 7.26], there exists at least one subsequence converging to t_* and at least one subsequence converging to t^*, and, in particular, the sequence converges if, and only if, $t_* = t^* = \lim_{k\to\infty} t_k$. We can apply this to the present situation: Suppose $(r(\Delta^{k_l}, f))_{l\in\mathbb{N}}$ is a converging subsequence of $(r(\Delta^k, f))_{k\in\mathbb{N}}$ with

$$\beta := \lim_{l\to\infty} r(\Delta^{k_l}, f). \tag{4.34}$$

First note $\beta \leq J_*(f, I)$ due to the definition of $J_*(f, I)$. Moreover, (4.33) implies $\lim_{l\to\infty} r(\Delta^{k_l} + \Delta', f) = \beta$. Since $r(\Delta', f) \leq r(\Delta^{k_l} + \Delta', f)$ and Δ' is arbitrary, we obtain $J_*(f, I) \leq \beta$, i.e. $J_*(f, I) = \beta$. Thus, we have shown that every subsequence of $(r(\Delta^k, f))_{k\in\mathbb{N}}$ converges to β, showing

$$J_*(f, I) = \beta = \lim_{k\to\infty} r(\Delta^k, f) \tag{4.35}$$

as claimed. In the same manner, one conducts the proof of $J^*(f, I) = \lim_{k\to\infty} R(\Delta^k, f)$. Then (4.30a) is immediate from the definition of Riemann integrability, and (4.30b) follows from (4.30a), since (4.11) implies

$$r(\Delta, f) \leq \rho(\Delta, f) \leq R(\Delta, f) \tag{4.36}$$

for each tagged partition Δ of I.

(e): Due to the definition of inf and sup,

$$\underset{\emptyset\neq A\subseteq I}{\forall} \quad \underset{\epsilon>0}{\forall} \quad \underset{t_*\in A}{\exists} \quad f(t_*) < \inf\{f(x) : x \in A\} + \epsilon, \tag{4.37a}$$

$$\underset{\emptyset\neq A\subseteq I}{\forall} \quad \underset{\epsilon>0}{\forall} \quad \underset{t^*\in A}{\exists} \quad f(t^*) > \sup\{f(x) : x \in A\} - \epsilon. \tag{4.37b}$$

In consequence, for each partition Δ of I and each $\epsilon > 0$, there are tags $\tau_* := (t_{p,*})_{p\in P(\Delta)}$ and $\tau^* := (t^*_p)_{p\in P(\Delta)}$ such that

$$\rho(\Delta, \tau_*, f) < r(\Delta, f) + \epsilon \quad \wedge \quad \rho(\Delta, \tau^*, f) > R(\Delta, f) - \epsilon. \tag{4.38}$$

Now let $(\Delta^k)_{k\in\mathbb{N}}$ be a sequence of partitions of I such that $\lim_{k\to\infty} |\Delta^k| = 0$. According to the above, for each Δ^k, there are tags $\tau^k_* := (t^k_{p,*})_{p\in P(\Delta^k)}$ and $\tau^{k,*} := (t^{k,*}_p)_{p\in P(\Delta^k)}$ such that

$$\underset{k\in\mathbb{N}}{\forall} \quad \left(\rho(\Delta^k, \tau^k_*, f) < r(\Delta^k, f) + \frac{1}{k} \quad \wedge \quad \rho(\Delta^k, \tau^{k,*}, f) > R(\Delta^k, f) - \frac{1}{k}\right). \tag{4.39}$$

Thus,

$$\begin{aligned} J_*(f, I) &\overset{(4.29)}{=} \lim_{k\to\infty} r(\Delta^k, f) \overset{(*)}{=} \lim_{k\to\infty} \rho(\Delta^k, \tau^k_*, f) = \alpha = \lim_{k\to\infty} \rho(\Delta^k, \tau^{k,*}, f) \\ &\overset{(**)}{=} \lim_{k\to\infty} R(\Delta^k, f) \overset{(4.29)}{=} J^*(f, I), \end{aligned} \tag{4.40}$$

where, at $(*)$ and $(**)$, we used (4.36), (4.39), and the Sandwich theorem [Phi15a, Th. 7.16]. Since (4.40) establishes both $f \in \mathcal{R}(I)$ and $\alpha = \int_I f$, the proof is complete. ■

Definition 4.11. If A is any set and $B \subseteq A$, then

$$\chi_B : A \longrightarrow \{0,1\}, \quad \chi_B(x) := \begin{cases} 1 & \text{for } x \in B, \\ 0 & \text{for } x \notin B, \end{cases} \tag{4.41}$$

is called the *characteristic function* of B.

Theorem 4.12. *Let $n \in \mathbb{N}$, $a, b \in \mathbb{R}^n$, $a < b$, $I := [a,b]$.*

(a) *The integral is linear (cf.* [Phi15a, Th. 10.11(a)]*): More precisely, if $f, g \in \mathcal{R}(I)$ and $\lambda, \mu \in \mathbb{R}$, then $\lambda f + \mu g \in \mathcal{R}(I)$ and*

$$\int_I (\lambda f + \mu g) = \lambda \int_I f + \mu \int_I g. \tag{4.42}$$

This result still holds in the $\mathbb{C}$-valued situation (see Th. D.5(a)*).*

(b) *If $c, d \in \mathbb{R}^n$, $c < d$, and $J := [c,d] \subseteq I$, then the characteristic function $\chi_{]c,d[}$ is Riemann integrable over I and*

$$\int_I \chi_{]c,d[} = |J|. \tag{4.43}$$

(c) *Monotonicity of the Integral (cf.* [Phi15a, Th. 10.11(c)]*): If $f, g : I \longrightarrow \mathbb{R}$ are bounded and $f \leq g$, then $J_*(f,I) \leq J_*(g,I)$ and $J^*(f,I) \leq J^*(g,I)$. In particular, if $f, g \in \mathcal{R}(I)$, then*

$$\int_I f \leq \int_I g. \tag{4.44}$$

(d) *Triangle Inequality (cf.* [Phi15a, Th. 10.11(d)]*): For each $f \in \mathcal{R}(I)$, one has*

$$\left| \int_I f \right| \leq \int_I |f|. \tag{4.45}$$

This result still holds for $\mathbb{C}$-valued f (see Th. D.5(b)*).*

(e) *Mean Value Theorem for Integration (cf.* [Phi15a, Th. 10.11(e)]*): If $f \in \mathcal{R}(I)$ and there exist numbers $m, M \in \mathbb{R}$ such that $m \leq f \leq M$, then*

$$m\,|I| \leq \int_I f \leq M\,|I|. \tag{4.46}$$

The theorem's name comes from the fact that $|I|^{-1} \int_I f$ is sometimes referred to as the mean value *of f on I.*

Proof. (a): Let $(\Delta^k)_{k\in\mathbb{N}}$ be a sequence of tagged partitions of I such that $\lim_{k\to\infty} |\Delta^k| = 0$. According to (4.11c), we have, for each $k \in \mathbb{N}$,

$$\rho(\Delta^k, \lambda f + \mu g) = \lambda \sum_{p\in P(\Delta^k)} f(t_p^k)|I_p^k| + \mu \sum_{p\in P(\Delta^k)} g(t_p^k)|I_p^k| = \lambda\,\rho(\Delta^k, f) + \mu\,\rho(\Delta^k, g). \tag{4.47}$$

Thus, if f and g are both Riemann integrable over I, then we obtain

$$\lim_{k\to\infty} \rho(\Delta^k, \lambda f + \mu g) \overset{(4.30b)}{=} \lambda \int_I f + \mu \int_I g. \tag{4.48}$$

Since (4.48) holds for each sequence $(\Delta^k)_{k\in\mathbb{N}}$ of tagged partitions of I with $\lim_{k\to\infty} |\Delta^k| = 0$, $\lambda f + \mu g$ is integrable and (4.42) holds by Th. 4.10(e).

(b): If we define the partition Δ of I by letting the partition Δ_j of $[a_j, b_j]$ be given by the node vector $\nu(\Delta_j) := \{a_j, c_j, d_j, b_j\}$, then there is $p \in P(\Delta)$ such that $I_p = J$. Moreover,

$$m_q(\chi_{]c,d[}) = M_q(\chi_{]c,d[}) = \begin{cases} 1 & \text{for } q = p, \\ 0 & \text{for } q \in P(\Delta) \setminus \{p\}. \end{cases} \tag{4.49}$$

Thus,

$$\begin{aligned} J_*(\chi_{]c,d[}, I) \geq r(\Delta, \chi_{]c,d[}) = \sum_{q\in P(\Delta)} m_q(\chi_{]c,d[})|I_q| = |I_p| = |J| = \sum_{q\in P(\Delta)} M_q(\chi_{]c,d[})|I_q| \\ \geq J^*(\chi_{]c,d[}, I) \geq J_*(\chi_{]c,d[}, I), \end{aligned} \tag{4.50}$$

proving $\chi_{]c,d[} \in \mathcal{R}(I)$ as well as (4.43).

(c): If $f, g : I \longrightarrow \mathbb{R}$ are bounded and $f \leq g$, then, for each partition Δ of I, $r(\Delta, f) \leq r(\Delta, g)$ and $R(\Delta, f) \leq R(\Delta, g)$ are immediate from (4.11). As these inequalities are preserved when taking the sup and the inf, respectively, all claims of (c) are established.

(d): We will see in Th. 4.15(b) below, that $f \in \mathcal{R}(I)$ implies $|f| \in \mathcal{R}(I)$. Since $f \leq |f|$ and $-f \leq |f|$, (c) implies $\int_I f \leq \int_I |f|$ and $-\int_I f \leq \int_I |f|$, i.e. (4.45).

(e): We compute

$$m\,|I| \overset{\text{Ex. 4.7(a)}}{=} \int_I m \overset{(c)}{\leq} \int_I f \overset{(c)}{\leq} \int_I M \overset{\text{Ex. 4.7(a)}}{=} M\,|I|, \tag{4.51}$$

thereby establishing the case. ∎

The following Th. 4.13 is in generalization of [Phi15a, Th. 10.12]:

Theorem 4.13 (Riemann's Integrability Criterion)**.** *Let $I = [a, b] \subseteq \mathbb{R}^n$ be an interval, $a, b \in \mathbb{R}^n$, $n \in \mathbb{N}$, $a < b$, and suppose $f : I \longrightarrow \mathbb{R}$ is bounded. Then f is Riemann integrable over I if, and only if, for each $\epsilon > 0$, there exists a partition Δ of I such that*

$$R(\Delta, f) - r(\Delta, f) < \epsilon. \tag{4.52}$$

Proof. Suppose, for each $\epsilon > 0$, there exists a partition Δ of I such that (4.52) is satisfied. Then

$$J^*(f, I) - J_*(f, I) \leq R(\Delta, f) - r(\Delta, f) < \epsilon, \tag{4.53}$$

showing $J^*(f, I) \leq J_*(f, I)$. As the opposite inequality always holds, we have $J^*(f, I) = J_*(f, I)$, i.e. $f \in \mathcal{R}(I)$ as claimed. Conversely, if $f \in \mathcal{R}(I)$ and $(\Delta^k)_{k\in\mathbb{N}}$ is a sequence of partitions of I with $\lim_{k\to\infty} |\Delta^k| = 0$, then (4.30a) implies that, for each $\epsilon > 0$, there is $N \in \mathbb{N}$ such that $R(\Delta^k, f) - r(\Delta^k, f) < \epsilon$ for each $k > N$. ∎

Theorem 4.14. *In generalization of* [Phi15a, Th. 10.15(a)]*, every continuous function on an interval $I = [a, b] \subseteq \mathbb{R}^n$, $a, b \in \mathbb{R}^n$, $n \in \mathbb{N}$, $a < b$, is Riemann integrable over I. This result still holds for $\mathbb{C}$-valued functions on I (see Th.* D.3*).*

Proof. Let $f : I \longrightarrow \mathbb{R}$ be continuous. First note that I is compact and, thus, f is bounded by Th. 3.8. As all norms on $\mathbb{R}^n$ are equivalent, in particular, f is continuous with respect to the max-norm on $\mathbb{R}^n$. Moreover, f is even uniformly continuous due to Th. 3.9. Thus, given $\epsilon > 0$, there is $\delta > 0$ such that $\|x - y\|_{\max} < \delta$ implies $|f(x) - f(y)| < \epsilon/|I|$. Then, for each partition Δ of I satisfying $|\Delta| < \delta$, we obtain

$$R(\Delta, f) - r(\Delta, f) = \sum_{p \in P(\Delta)} (M_p - m_p)|I_p| \leq \frac{\epsilon}{|I|} \sum_{p \in P(\Delta)} |I_p| = \epsilon, \tag{4.54}$$

as $|\Delta| < \delta$ implies $\|x - y\|_{\max} < \delta$ for each $x, y \in I_p$ and each $p \in P(\Delta)$. Finally, (4.54) implies $f \in \mathcal{R}(I)$ due to Riemann's integrability criterion of Th. 4.13. ■

The following Th. 4.15 generalizes the assertions of [Phi15a, Th. 10.17] from functions on 1-dimensional intervals to functions on n dimensional intervals, $n \in \mathbb{N}$:

Theorem 4.15. *Let $n \in \mathbb{N}$, $a, b \in \mathbb{R}^n$, $a < b$, $I := [a, b]$.*

(a) *If $f \in \mathcal{R}(I)$ and $\phi : f(I) \longrightarrow \mathbb{R}$ is Lipschitz continuous, then $\phi \circ f \in \mathcal{R}(I)$. For $\mathbb{C}$-valued extensions of this result, see Th.* D.4(b),(c)*.*

(b) *If $f \in \mathcal{R}(I)$, then $|f|, f^2, f^+, f^- \in \mathcal{R}(I)$. In particular, we, indeed, have* (4.2) *from the introduction (with A replaced by I). If, in addition, there exists $\delta > 0$ such that $f(x) \geq \delta$ for each $x \in I$, then $1/f \in \mathcal{R}(I)$.*

(c) *If $f, g \in \mathcal{R}(I)$, then $fg, \max(f, g), \min(f, g) \in \mathcal{R}(I)$. If, in addition, there exists $\delta > 0$ such that $g(x) \geq \delta$ for each $x \in I$, then $f/g \in \mathcal{R}(I)$. For the product and the quotient, the result remains true for $\mathbb{C}$-valued f, g (see Th.* D.4(a)*).*

Proof. (a): Let $f \in \mathcal{R}(I)$ and let $\phi : f(I) \longrightarrow \mathbb{R}$ be Lipschitz continuous. Then there exists $L \geq 0$ such that

$$\big|\phi(x) - \phi(y)\big| \leq L\,|x - y| \quad \text{for each } x, y \in f(I). \tag{4.55}$$

As $f \in \mathcal{R}(I)$, given $\epsilon > 0$, Th. 4.13 provides a partition Δ of I such that $R(\Delta, f) - r(\Delta, f) < \epsilon/L$, and we obtain

$$\begin{aligned} R(\Delta, \phi \circ f) - r(\Delta, \phi \circ f) &= \sum_{p \in P(\Delta)} \big(M_p(\phi \circ f) - m_p(\phi \circ f)\big)|I_p| \\ &\leq \sum_{p \in P(\Delta)} L\big(M_p(f) - m_p(f)\big)|I_p| \\ &= L\,\big(R(\Delta, f) - r(\Delta, f)\big) < \epsilon. \end{aligned} \tag{4.56}$$

Thus, $\phi \circ f \in \mathcal{R}(I)$ by another application of Th. 4.13.

(b): $|f|, f^2, f^+, f^- \in \mathcal{R}(I)$ follows from (a), since each of the maps $x \mapsto |x|$, $x \mapsto x^2$, $x \mapsto \max\{x, 0\}$, $x \mapsto -\min\{x, 0\}$ is Lipschitz continuous on the bounded set $f(I)$ (recall that $f \in \mathcal{R}(I)$ implies that f is bounded). Since $f = f^+ - f^-$, (4.2) is implied by (4.42). Finally, if $f(x) \geq \delta > 0$, then $x \mapsto x^{-1}$ is Lipschitz continuous on the bounded set $f(I)$, and $f^{-1} \in \mathcal{R}(I)$ follows from (a).

(c): Since

$$fg = \frac{1}{4}(f+g)^2 - (f-g)^2, \tag{4.57a}$$

$$\max(f, g) = f + (g-f)^+, \tag{4.57b}$$

$$\min(f, g) = g - (f-g)^-, \tag{4.57c}$$

everything is a consequence of (b). ■

4.2 Important Theorems

4.2.1 Fubini Theorem

In [Phi15a, Sec. 10.2], we saw several important theorems often helpful in the evaluation of one-dimensional Riemann integrals. The following Fubini Th. 4.16 allows to compute an n-dimensional Riemann integral as an iteration of n one-dimensional Riemann integrals:

Theorem 4.16 (Fubini)**.** *Let $a, b, c, d, e, f \in \mathbb{R}^n$, $n \in \mathbb{N}$, $a < b$, $c < d$, $e < f$, $I = [a, b]$, $J = [c, d]$, $K = [e, f]$. If $I = J \times K$ and $f \in \mathcal{R}(I)$, then*

$$\int_I f = \int_I f(x, y)\,\mathrm{d}(x, y) = \int_K \int_J f(x, y)\,\mathrm{d}x\,\mathrm{d}y = \int_J \int_K f(x, y)\,\mathrm{d}y\,\mathrm{d}x\,. \tag{4.58}$$

There is a slight abuse of notation in (4.58)*, as it can happen that a function $x \mapsto f(x, y)$ is not Riemann integrable over J and that a function $y \mapsto f(x, y)$ is not Riemann integrable over K. However, in that case, one can choose either the lower or the upper Riemann integral for the inner integrals in* (4.58)*. Independently of the choice, the resulting function $y \mapsto \int_J f(x, y)\,\mathrm{d}x$ is Riemann integrable over K, $x \mapsto \int_K f(x, y)\,\mathrm{d}y$ is Riemann integrable over J, and the validity of* (4.58) *is unaffected (this issue is related to the fact that changing a function's value at a "small" (for example, finite) number of points, will not change the value of its Riemann integral). By applying* (4.58) *inductively, one obtains*

$$\int_I f = \int_I f(x)\,\mathrm{d}x = \int_{a_1}^{b_1} \dots \int_{a_n}^{b_n} f(x_1, \dots, x_n)\,\mathrm{d}x_n \cdots \mathrm{d}x_1, \tag{4.59}$$

where, for the inner integrals, one can arbitrarily choose upper or lower Riemann integrals, and one can also permute their order without changing the overall value. The Fubini theorem still holds for $\mathbb{C}$-valued f (see Th. D.7*).*

Proof. Even though, in the present context, the proof of the Fubini theorem is not that difficult, we refer to [Wal02, Sec. 7.15]. ■

Example 4.17. Let $I := [0,1]^3$ and $f : I \longrightarrow \mathbb{R}$, $f(x,y,z) := xyz$. We compute the integral $\int_I f$:

$$\begin{aligned}\int_I f &= \int_0^1 \int_0^1 \int_0^1 f(x,y,z)\,\mathrm{d}x\ \mathrm{d}y\ \mathrm{d}z = \int_0^1 \int_0^1 \int_0^1 xyz\,\mathrm{d}x\ \mathrm{d}y\ \mathrm{d}z \\ &= \int_0^1 \int_0^1 \left[\frac{x^2yz}{2}\right]_{x=0}^{x=1} \mathrm{d}y\ \mathrm{d}z = \int_0^1 \int_0^1 \frac{yz}{2}\,\mathrm{d}y\ \mathrm{d}z = \int_0^1 \left[\frac{y^2z}{4}\right]_{y=0}^{y=1} \mathrm{d}z \\ &= \int_0^1 \frac{z}{4}\,\mathrm{d}z = \left[\frac{z^2}{8}\right]_0^1 = \frac{1}{8}. \end{aligned} \tag{4.60}$$

4.2.2 Change of Variables

Recall the 1-dimensional change of variables theorem [Phi15a, Th. 10.24]. The following n-dimensional version Th. 4.18 is a much deeper result. In Ex. 4.19 below, we will see that it can be very useful to compute both multi- and 1-dimensional integrals.

Theorem 4.18 (Change of Variables)**.** *Let $a,b,c,d \in \mathbb{R}^n$, $n \in \mathbb{N}$, $a < b$, $c < d$, $I := [a,b]$, $J := [c,d]$, $\phi : I \longrightarrow \mathbb{R}^n$, $f : J \longrightarrow \mathbb{R}$. If, on the interior of I, ϕ is one-to-one, Lipschitz continuous, and has continuous first partials, $\phi(I) \subseteq J$, and $(f \circ \phi)|\det J_\phi| \in \mathcal{R}(I)$, then $f\chi_{\phi(I)} \in \mathcal{R}(J)$ and the following* change of variables *formula holds:*

$$\int_J f\chi_{\phi(I)} = \int_I (f \circ \phi)|\det J_\phi|, \tag{4.61}$$

where $\chi_{\phi(I)}$ is the characteristic function of $\phi(I)$ defined according to Def. 4.11*. The change of variables theorem still holds for $\mathbb{C}$-valued f (see Th.* D.8*).*

Proof. The proof of the n-dimensional change of variables theorem is much harder than the 1-dimensional case. For example, a proof can be found in [Wal02, Sec. 7.18]. ■

The change of variables Th. 4.18 is often most effective in combination with the Fubini Th. 4.16, as illustrated by the following example:

Example 4.19. Consider the function

$$f : \mathbb{R}^2 \longrightarrow \mathbb{R}, \quad f(x,y) := e^{-(x^2+y^2)}. \tag{4.62}$$

For each $R > 0$, let $\overline{B}_R := \overline{B}_R(0) = \{(x,y) \in \mathbb{R}^2 : \|(x,y)\|_2 \leq R\}$ denote the circle with radius R and center 0, and let $J_R := [-R,R]^2$ the corresponding square with side length $2R$. We want to compute

$$\int_{\overline{B}_R} f := \int_{J_R} f\,\chi_{\overline{B}_R} = \int_{J_R} e^{-(x^2+y^2)}\,\chi_{\overline{B}_R}(x,y)\,\mathrm{d}(x,y) \tag{4.63}$$

(note $\overline{B}_R \subseteq J_R$, since $(x,y) \in \overline{B}_R$ implies $x^2+y^2 \leq R^2$, i.e. $|x| \leq R$ and $|y| \leq R$). This can be accomplished using change of variables, Fubini, and so-called *polar coordinates*, i.e.

$$\phi : \mathbb{R}_0^+ \times [0,2\pi] \longrightarrow \mathbb{R}^2, \quad \phi(r,\varphi) := (r\cos\varphi, r\sin\varphi). \tag{4.64}$$

To apply Th. 4.18 with $J = J_R$ and $I = I_R := [0,R] \times [0,2\pi]$, we need to verify that all hypotheses are satisfied. We start by observing $\phi(I_R) = \overline{B}_R \subseteq J_R$. Moreover, the map ϕ restricted to the interior $I_R^\circ =]0,R[\times]0,2\pi[$,

$$\phi : I_R^\circ \longrightarrow B_R', \quad \phi(r,\varphi) := (r\cos\varphi, r\sin\varphi), \tag{4.65}$$

where

$$B_R' := B_R \setminus \{(x,0) : 0 \leq x < R\}, \tag{4.66}$$

is bijective with inverse map

$$\phi^{-1} : B_R' \longrightarrow I_R^\circ, \quad \phi^{-1}(x,y) = \big(\phi_1^{-1}(x,y), \phi_2^{-1}(x,y)\big), \tag{4.67a}$$

where, recalling the definition of arccot from [Phi15a, Def. and Rem. 8.27],

$$\phi_1^{-1}(x,y) := \sqrt{x^2+y^2}, \tag{4.67b}$$

$$\phi_2^{-1}(x,y) := \begin{cases} \operatorname{arccot}(x/y) & \text{for } y > 0, \\ \pi & \text{for } y = 0, \\ \pi + \operatorname{arccot}(x/y) & \text{for } y < 0. \end{cases} \tag{4.67c}$$

One verifies $\phi^{-1} \circ \phi = \mathrm{Id}_{I_R^\circ}$:

$$\forall_{(r,\varphi)\in I_R^\circ} \quad (\phi^{-1}\circ\phi)(r,\varphi) = \phi^{-1}(r\cos\varphi, r\sin\varphi) = (r,\varphi),$$

since

$$\phi_1^{-1}(r\cos\varphi, r\sin\varphi) = \sqrt{r^2(\cos^2\varphi + \sin^2\varphi)} = r,$$

$$\phi_2^{-1}(r\cos\varphi, r\sin\varphi) = \begin{cases} \operatorname{arccot}(\cot\varphi) = \varphi & \text{for } 0 < \varphi < \pi, \\ \pi = \varphi & \text{for } \varphi = \pi, \\ \pi + \operatorname{arccot}(\cot\varphi) = \pi + \varphi - \pi = \varphi & \text{for } \pi < \varphi < 2\pi; \end{cases}$$

and $\phi \circ \phi^{-1} = \mathrm{Id}_{B_R'}$:

$$\begin{aligned} \forall_{(x,y)\in B_R'} \quad (\phi\circ\phi^{-1})(x,y) &= \Big(\sqrt{x^2+y^2}\cos\phi_2^{-1}(x,y),\ \sqrt{x^2+y^2}\sin\phi_2^{-1}(x,y)\Big) \\ &= (x,y), \end{aligned}$$

since

$$\cos\phi_2^{-1}(x,y) = \begin{cases} \frac{1}{\sqrt{1+(\cot\operatorname{arccot}(x/y))^{-2}}} = \frac{1}{\sqrt{1+\frac{y^2}{x^2}}} = \frac{x}{\sqrt{x^2+y^2}} & \text{for } x>0\ (\Rightarrow y\neq 0),\\ \cos(\pi/2)=0=\frac{x}{\sqrt{x^2+y^2}} & \text{for } x=0, y>0,\\ \cos(3\pi/2)=0=\frac{x}{\sqrt{x^2+y^2}} & \text{for } x=0, y<0,\\ \frac{-1}{\sqrt{1+(\cot\operatorname{arccot}(x/y))^{-2}}} = \frac{x}{\sqrt{x^2+y^2}} & \text{for } x<0, y\neq 0,\\ \cos\pi=-1=\frac{x}{\sqrt{x^2+y^2}} & \text{for } y=0\ (\Rightarrow x<0), \end{cases}$$

$$\sin\phi_2^{-1}(x,y) = \begin{cases} \frac{1}{\sqrt{1+\cot^2\operatorname{arccot}(x/y)}} = \frac{1}{\sqrt{1+\frac{x^2}{y^2}}} = \frac{y}{\sqrt{x^2+y^2}} & \text{for } y>0,\\ \sin\pi=0=\frac{y}{\sqrt{x^2+y^2}} & \text{for } y=0,\\ \frac{-1}{\sqrt{1+\cot^2\operatorname{arccot}(x/y)}} = \frac{y}{\sqrt{x^2+y^2}} & \text{for } y<0. \end{cases}$$

Next, we note that ϕ has continuous first partials on I_R°, where

$$\underset{(r,\varphi)\in I_R^\circ}{\forall} \quad J_\phi(r,\varphi) = \begin{pmatrix} \cos\varphi & -r\sin\varphi \\ \sin\varphi & r\cos\varphi \end{pmatrix}, \quad \det J_\phi(r,\varphi) = r. \tag{4.68}$$

The partials of ϕ, namely $|\partial_i\phi_j(r,\varphi)|$ are all bounded by R. Thus, ϕ_1 and ϕ_2 both are Lipschitz continuous by Th. 2.35, implying ϕ is also Lipschitz continuous. Finally,

$$\underset{(r,\varphi)\in I_R^\circ}{\forall} \quad \Big((f\circ\phi)|\det J_\phi|\Big)(r,\varphi) = re^{-(r^2\cos^2\varphi + r^2\sin^2\varphi)} = re^{-r^2}, \tag{4.69}$$

showing $(f\circ\phi)|\det J_\phi| \in \mathcal{R}(I)$.

We have thereby verified all the hypotheses of Th. 4.18 and can conclude $f\,\chi_{\overline{B}_R} \in \mathcal{R}(J)$ as well as

$$\begin{aligned} \alpha_R &:= \int_{\overline{B}_R} f = \int_{J_R} f\,\chi_{\overline{B}_R} = \int_{J_R} e^{-(x^2+y^2)}\,\chi_{\overline{B}_R}(x,y)\,\mathrm{d}(x,y) \overset{(4.61)}{=} \int_{I_R} (f\circ\phi)|\det J_\phi| \\ &= \int_{I_R} re^{-r^2}\,\mathrm{d}(r,\varphi) \overset{(*)}{=} \int_0^R\int_0^{2\pi} re^{-r^2}\,\mathrm{d}\varphi\,\mathrm{d}r = \int_0^R 2\pi\, re^{-r^2}\,\mathrm{d}r = \Big[-\pi\, e^{-r^2}\Big]_0^R \\ &= \pi\left(1-e^{-R^2}\right), \end{aligned} \tag{4.70}$$

where the Fubini Th. 4.16 was used at $(*)$.

With another application of the Fubini Th. 4.16, we can use (4.70) to prove the following important equality:

$$\int_{-\infty}^{\infty} e^{-x^2}\,\mathrm{d}x := \lim_{R\to\infty}\int_{-R}^{R} e^{-x^2}\,\mathrm{d}x = \sqrt{\pi}. \tag{4.71}$$

Indeed, we have

$$\begin{aligned} &\int_{[-R,R]^2} e^{-(x^2+y^2)}\,\mathrm{d}(x,y) = \int_{-R}^{R}\int_{-R}^{R} e^{-x^2}e^{-y^2}\,\mathrm{d}x\,\mathrm{d}y \\ \underset{R\in\mathbb{R}^+}{\forall} \quad &= \int_{-R}^{R} e^{-y^2}\int_{-R}^{R} e^{-x^2}\,\mathrm{d}x\,\mathrm{d}y = \beta_R^2, \quad \text{where} \quad \beta_R := \int_{-R}^{R} e^{-x^2}\,\mathrm{d}x\,, \end{aligned} \tag{4.72}$$

and, since, for each $R \in \mathbb{R}^+$, $\overline{B}_R \subseteq [-R,R]^2 \subseteq \overline{B}_{2R}$, monotonicity of the integral according to (4.44) provides

$$\forall_{R\in\mathbb{R}^+} \quad \alpha_R \leq \beta_R^2 \leq \alpha_{2R}, \tag{4.73}$$

i.e. the Sandwich theorem yields

$$\pi = \lim_{R\to\infty} \alpha_R \leq \lim_{R\to\infty} \beta_R^2 \leq \lim_{R\to\infty} \alpha_{2R} = \pi, \tag{4.74}$$

proving (4.71).

5 Short Introduction to Ordinary Differential Equations (ODE)

5.1 Definition and Geometric Interpretation

Definition 5.1. Let $G \subseteq \mathbb{R}\times\mathbb{R}$ and let $f : G \longrightarrow \mathbb{R}$ be continuous. An *explicit ordinary differential equation (ODE)* of *first order* is an equation of the form

$$y' = f(x,y), \tag{5.1}$$

which is an equation for the unknown *function* y. A *solution* to this ODE is a differentiable function

$$\phi : I \longrightarrow \mathbb{R}, \tag{5.2}$$

defined on a nontrivial (bounded or unbounded, open or closed or half-open) interval $I \subseteq \mathbb{R}$, satisfying the following two conditions:

(i) The graph of ϕ is contained in G, i.e. $\{(x,\phi(x)) \in I\times\mathbb{R} : x \in I\} \subseteq G$.

(ii) $\phi'(x) = f\big(x,\phi(x)\big)$ for each $x \in I$.

Note that condition (i) is necessary so that one can even formulate condition (ii).

Definition 5.2. An *initial value problem* for (5.1) consists of the ODE (5.1) plus the *initial condition*

$$y(x_0) = y_0, \tag{5.3}$$

with given $(x_0,y_0) \in G$. A solution ϕ to the initial value problem is a differentiable function ϕ as in (5.2) that is a solution to the ODE and that also satisfies (5.3) (with y replaced by ϕ) – in particular, this requires $x_0 \in I$.

—

One distinguishes between ordinary differential equations and partial differential equations (PDE). While ODE contain only derivatives with respect to one variable, PDE can contain (partial) derivatives with respect to several different variables. In general, PDE are much harder to solve than ODE. The term *first order* in Def. 5.1 indicates that only a first derivative occurs in the equation. Correspondingly, ODE of second order contain derivatives of second order etc. ODE of higher order can be equivalently formulated and solved as systems of ODE of first order (see, e.g., [Phi14, Sec. 3.1]). The *explicit* in Def. 5.1 indicates that the ODE is explicitly solved for y'. One can also consider *implicit* ODE of the form $f(x, y, y') = 0$. We will only consider explicit ODE of first order in the following.

It can be useful to rewrite a first-order explicit intitial value problem as an equivalent *integral equation*. We provide the details of this equivalence in the following theorem:

Theorem 5.3. *If $G \subseteq \mathbb{R}\times\mathbb{R}$ and $f : G \longrightarrow \mathbb{R}$ is continuous, then, for each $(x_0, y_0) \in G$, the explicit first-order initial value problem*

$$y' = f(x, y), \tag{5.4a}$$
$$y(x_0) = y_0, \tag{5.4b}$$

is equivalent to the integral equation

$$y(x) = y_0 + \int_{x_0}^{x} f\big(t, y(t)\big)\,\mathrm{d}t\,, \tag{5.5}$$

in the sense that a continuous function $\phi : I \longrightarrow \mathbb{R}$, with $x_0 \in I \subseteq \mathbb{R}$ being a nontrivial interval, and ϕ satisfying

$$\big\{\big(x, \phi(x)\big) \in I \times \mathbb{R} : x \in I\big\} \subseteq G, \tag{5.6}$$

is a solution to (5.4) *in the sense of Def.* 5.2 *if, and only if,*

$$\underset{x\in I}{\forall} \quad \phi(x) = y_0 + \int_{x_0}^{x} f\big(t, \phi(t)\big)\,\mathrm{d}t\,, \tag{5.7}$$

i.e. if, and only if, ϕ is a solution *to the integral equation* (5.5).

Proof. Assume $I \subseteq \mathbb{R}$ with $x_0 \in I$ to be a nontrivial interval and $\phi : I \longrightarrow \mathbb{R}$ to be a continuous function, satisfying (5.6). If ϕ is a solution to (5.4), then ϕ is differentiable and the assumed continuity of f implies the continuity of ϕ'. In other words, $\phi \in C^1(I)$. Thus, the fundamental theorem of calculus in the form [Phi15a, Th. 10.19(b)] applies, and [Phi15a, (10.50b)] yields

$$\underset{x\in I}{\forall} \quad \phi(x) = \phi(x_0) + \int_{x_0}^{x} f\big(t, \phi(t)\big)\,\mathrm{d}t \overset{(5.4b)}{=} y_0 + \int_{x_0}^{x} f\big(t, \phi(t)\big)\,\mathrm{d}t\,, \tag{5.8}$$

proving ϕ satisfies (5.7). Conversely, if ϕ satisfies (5.7), then the validity of the initial condition (5.4b) is immediate. Moreover, as f and ϕ are continuous, so is the integrand function $t \mapsto f\big(t, \phi(t)\big)$ of (5.7). Thus, [Phi15a, Th. 10.19(a)] applies to ϕ, proving $\phi'(x) = f\big(x, \phi(x)\big)$ for each $x \in I$, i.e. ϕ is a solution to (5.4). ∎

Example 5.4. Consider the situation of Th. 5.3. In the particularly simple special case, where f does not actually depend on y, but merely on x, the equivalence between (5.4) and (5.5) can be directly exploited to actually *solve* the initial value problem: If $f : I \longrightarrow \mathbb{R}$, where $I \subseteq \mathbb{R}$ is some nontrivial interval with $x_0 \in I$, then we obtain $\phi : I \longrightarrow \mathbb{R}$ to be a solution of (5.4) if, and only if,

$$\underset{x\in I}{\forall} \quad \phi(x) = y_0 + \int_{x_0}^{x} f(t)\,\mathrm{d}t\,, \tag{5.9}$$

i.e. if, and only if, ϕ is the antiderivative of f that satisfies the initial condition. In particular, in the present situation, ϕ as given by (5.9) is the *unique* solution to the initial value problem. Of course, depending on f, it can still be difficult to carry out the integral in (5.9).

A simple concrete example is

$$y' = a, \tag{5.10a}$$
$$y(0) = c, \tag{5.10b}$$

with $a, c \in \mathbb{R}$. Then, on $\mathbb{R}$, the function

$$\phi : \mathbb{R} \longrightarrow \mathbb{R}, \quad \phi(x) = c + \int_0^x a\,\mathrm{d}t \; = c + xa, \tag{5.11}$$

is the unique solution to (5.10).

—

Geometrically, the ODE (5.1) provides a slope $y' = f(x, y)$ for every point (x, y). In other words, it provides a field of directions. The task is to find a differentiable function ϕ such that its graph has the prescribed slope in each point it contains. In certain simple cases, drawing the field of directions can help to *guess* the solutions of the ODE.

Example 5.5. **(a)** Let $G := \mathbb{R}^+ \times \mathbb{R}$ and $f : G \longrightarrow \mathbb{R}$, $f(x, y) := y/x$, i.e. we consider the ODE $y' = y/x$. Drawing the field of directions leads to the idea that the solutions are functions whose graphs constitute rays, i.e. $\phi_c : \mathbb{R}^+ \longrightarrow \mathbb{R}$, $y = \phi_c(x) = c\,x$ with $c \in \mathbb{R}$. Indeed, one immediately verifies that each ϕ_c constitutes a solution to the ODE.

(b) Let $G := \mathbb{R} \times \mathbb{R}^+$ and $f : G \longrightarrow \mathbb{R}$, $f(x, y) := -x/y$, i.e. we consider the ODE $y' = -x/y$. Drawing the field of directions leads to the idea that the solutions are functions whose graphs constitute semicircles, i.e. $\phi_c :]-\sqrt{c}, \sqrt{c}[\longrightarrow \mathbb{R}$, $y = \phi_c(x) = \sqrt{c - x^2}$ with $c \in \mathbb{R}^+$. Indeed, we get

$$y' = \phi_c'(x) = \frac{-2x}{2\sqrt{c - x^2}} = \frac{-x}{\phi_c(x)} = \frac{-x}{y}, \tag{5.12}$$

i.e. each ϕ_c constitutes a solution to the ODE.

5.2 Separation of Variables

If the ODE (5.1) has the particular form

$$y' = f(x)g(y), \tag{5.13}$$

with one-dimensional continuous functions f and g, and $g(y) \neq 0$, then it can be solved by a method known as *separation of variables*:

Theorem 5.6. *Let $I, J \subseteq \mathbb{R}$ be (bounded or unbounded) open intervals and suppose that $f : I \longrightarrow \mathbb{R}$ and $g : J \longrightarrow \mathbb{R}$ are continuous with $g(y) \neq 0$ for each $y \in J$. For each $(x_0, y_0) \in I \times J$, consider the initial value problem consisting of the ODE (5.13) together with the initial condition*

$$y(x_0) = y_0. \tag{5.14}$$

Define the functions

$$F : I \longrightarrow \mathbb{R}, \quad F(x) := \int_{x_0}^{x} f(t)\,\mathrm{d}t\,, \quad G : J \longrightarrow \mathbb{R}, \quad G(y) := \int_{y_0}^{y} \frac{\mathrm{d}t}{g(t)}. \tag{5.15}$$

(a) Uniqueness: *On each open interval $I' \subseteq I$ satisfying $x_0 \in I'$ and $F(I') \subseteq G(J)$, the initial value problem consisting of (5.13) and (5.14) has a unique solution. This unique solution is given by*

$$\phi : I' \longrightarrow \mathbb{R}, \quad \phi(x) := G^{-1}\big(F(x)\big), \tag{5.16}$$

where $G^{-1} : G(J) \longrightarrow J$ is the inverse function of G on $G(J)$.

(b) Existence: *There exists an open interval $I' \subseteq I$ satisfying $x_0 \in I'$ and $F(I') \subseteq G(J)$, i.e. an I' such that* (a) *applies.*

Proof. (a): We begin by proving G has a differentiable inverse function $G^{-1} : G(J) \longrightarrow J$. According to the fundamental theorem of calculus [Phi15a, Th. 10.19(a)], G is differentiable with $G' = 1/g$. Since g is continuous and nonzero, G is even C^1. If $G'(y_0) = 1/g(y_0) > 0$, then G is strictly increasing on J (due to the intermediate value theorem [Phi15a, Th. 7.57]; $g(y_0) > 0$, the continuity of g, and $g \neq 0$ imply that $g > 0$ on J). Analogously, if $G'(y_0) = 1/g(y_0) < 0$, then G is strictly decreasing on J. In each case, G has a differentiable inverse function on $G(J)$ by [Phi15a, Th. 9.8].

In the next step, we verify that (5.16) does, indeed, define a solution to (5.13) and (5.14). The assumption $F(I') \subseteq G(J)$ and the existence of G^{-1} as shown above provide that ϕ is well-defined by (5.16). Verifying (5.14) is quite simple: $\phi(x_0) = G^{-1}(F(x_0)) = G^{-1}(0) = y_0$. To see ϕ to be a solution of (5.13), notice that (5.16) implies $F = G \circ \phi$ on I'. Thus, we can apply the chain rule to obtain the derivative of $F = G \circ \phi$ on I':

$$\underset{x \in I'}{\forall} \quad f(x) = F'(x) = G'\big(\phi(x)\big)\,\phi'(x) = \frac{\phi'(x)}{g\big(\phi(x)\big)}, \tag{5.17}$$

showing ϕ satisfies (5.13).

We now proceed to show that each solution $\phi : I' \longrightarrow \mathbb{R}$ to (5.13) that satisfies (5.14) must also satisfy (5.16). Since ϕ is a solution to (5.13),

$$\frac{\phi'(x)}{g\big(\phi(x)\big)} = f(x) \quad \text{for each } x \in I'. \tag{5.18}$$

Integrating (5.18) yields

$$\int_{x_0}^{x} \frac{\phi'(t)}{g\big(\phi(t)\big)}\,\mathrm{d}t \;= \int_{x_0}^{x} f(t)\,\mathrm{d}t \;= F(x) \quad \text{for each } x \in I'. \tag{5.19}$$

Using the change of variables formula of [Phi15a, Th. 10.24] in the left-hand side of (5.19), allows one to replace $\phi(t)$ by the new integration variable u (note that each solution $\phi : I' \longrightarrow \mathbb{R}$ to (5.13) is in $C^1(I')$ since f and g are presumed continuous). Thus, we obtain from (5.19):

$$F(x) = \int_{\phi(x_0)}^{\phi(x)} \frac{\mathrm{d}u}{g(u)} = \int_{y_0}^{\phi(x)} \frac{\mathrm{d}u}{g(u)} = G\big(\phi(x)\big) \quad \text{for each } x \in I'. \tag{5.20}$$

Applying G^{-1} to (5.20) establishes ϕ satisfies (5.16).

(b): During the proof of (a), we have already seen G to be either strictly increasing or strictly decreasing. As $G(y_0) = 0$, this implies the existence of $\epsilon > 0$ such that $]-\epsilon, \epsilon[\subseteq G(J)$. The function F is differentiable and, in particular, continuous. Since $F(x_0) = 0$, there is $\delta > 0$ such that, for $I' :=]x_0 - \delta, x_0 + \delta[$, one has $F(I') \subseteq]-\epsilon, \epsilon[\subseteq G(J)$ as desired. ■

Example 5.7. Consider the ODE

$$y' = -\frac{y}{x} \quad \text{on } I \times J := \mathbb{R}^+ \times \mathbb{R}^+ \tag{5.21}$$

with the initial condition $y(1) = c$ for some given $c \in \mathbb{R}^+$. Introducing functions

$$f : \mathbb{R}^+ \longrightarrow \mathbb{R}, \quad f(x) := -\frac{1}{x}, \quad g : \mathbb{R}^+ \longrightarrow \mathbb{R}, \quad g(y) := y, \tag{5.22}$$

one sees that Th. 5.6 applies. To compute the solution $\phi = G^{-1} \circ F$, we first have to determine F and G:

$$F : \mathbb{R}^+ \longrightarrow \mathbb{R}, \qquad F(x) = \int_1^x f(t)\,\mathrm{d}t \;= -\int_1^x \frac{\mathrm{d}t}{t} = -\ln x, \tag{5.23a}$$

$$G : \mathbb{R}^+ \longrightarrow \mathbb{R}, \qquad G(y) = \int_c^y \frac{\mathrm{d}t}{g(t)} = \int_c^y \frac{\mathrm{d}t}{t} = \ln \frac{y}{c}. \tag{5.23b}$$

Here, we can choose $I' = I = \mathbb{R}^+$, because $F(\mathbb{R}^+) = \mathbb{R} = G(\mathbb{R}^+)$. That means ϕ is defined on the entire interval I. The inverse function of G is given by

$$G^{-1} : \mathbb{R} \longrightarrow \mathbb{R}^+, \quad G^{-1}(t) = c\,e^t. \tag{5.24}$$

Finally, we get

$$\phi : \mathbb{R}^+ \longrightarrow \mathbb{R}, \quad \phi(x) = G^{-1}\big(F(x)\big) = c\, e^{-\ln x} = \frac{c}{x}. \tag{5.25}$$

The uniqueness part of Th. 5.6 further tells us the above initial value problem can have no solution different from ϕ.

—

The advantage of using Th. 5.6 as in the previous example, by computing the relevant functions F, G, and G^{-1}, is that it is mathematically rigorous. In particular, one can be sure one has found the unique solution to the ODE with initial condition. However, in practice, it is often easier to use the following heuristic (not entirely rigorous) procedure. In the end, in most cases, one can easily check by differentiation that the function found is, indeed, a solution to the ODE with initial condition. However, one does not know uniqueness without further investigations. One also has to determine on which interval the found solution is defined. On the other hand, as one is usually interested in choosing the interval as large as possible, the optimal choice is not always obvious when using Th. 5.6, either.

The heuristic procedure is as follows: Start with the ODE (5.13) written in the form

$$\frac{\mathrm{d}y}{\mathrm{d}x} = f(x)g(y). \tag{5.26a}$$

Multiply by $\mathrm{d}x$ and divide by $g(y)$ (i.e. *separate the variables*):

$$\frac{\mathrm{d}y}{g(y)} = f(x)\,\mathrm{d}x\,. \tag{5.26b}$$

Integrate:

$$\int \frac{\mathrm{d}y}{g(y)} = \int f(x)\,\mathrm{d}x\,. \tag{5.26c}$$

Change the integration variables and supply the appropriate upper and lower limits for the integrals (according to the initial condition):

$$\int_{y_0}^{y} \frac{\mathrm{d}t}{g(t)} = \int_{x_0}^{x} f(t)\,\mathrm{d}t\,. \tag{5.26d}$$

Solve this equation for y, set $\phi(x) := y$, check by differentiation that ϕ is, indeed, a solution to the ODE, and determine the largest interval I' such that $x_0 \in I'$ and such that ϕ is defined on I'. The use of this heuristic algorithm is demonstrated by the following example:

Example 5.8. Consider the ODE

$$y' = \frac{x}{y} \quad \text{on } I \times J := \mathbb{R}^+ \times \mathbb{R}^+ \tag{5.27}$$

with the initial condition $y(x_0) = y_0$ for given values $x_0, y_0 \in \mathbb{R}^+$. We manipulate (5.27) according to the heuristic algorithm described in (5.26) above:

$$\begin{aligned} \frac{\mathrm{d}y}{\mathrm{d}x} = \frac{x}{y} \quad &\rightsquigarrow \quad y\,\mathrm{d}y = x\,\mathrm{d}x \quad \rightsquigarrow \quad \int y\,\mathrm{d}y = \int x\,\mathrm{d}x \quad \rightsquigarrow \quad \int_{y_0}^{y} t\,\mathrm{d}t = \int_{x_0}^{x} t\,\mathrm{d}t \\ &\rightsquigarrow \quad y^2 - y_0^2 = x^2 - x_0^2 \quad \rightsquigarrow \quad \phi(x) = y = \sqrt{x^2 + y_0^2 - x_0^2} \end{aligned} \tag{5.28}$$

(the negative sign in front of the square root in the last manipulation is excluded by the assumption that $y_0 = \phi(x_0) \in \mathbb{R}^+$). Clearly, $\phi(x_0) = y_0$. Moreover,

$$\phi'(x) = \frac{2x}{2\sqrt{x^2 + y_0^2 - x_0^2}} = \frac{x}{\phi(x)}, \tag{5.29}$$

i.e. ϕ does, indeed, provide a solution to (5.27). For $y_0 \geq x_0$, ϕ is defined on the entire interval $I = \mathbb{R}^+$. However, if $y_0 < x_0$, then $x^2 + y_0^2 - x_0^2 > 0$ implies $x^2 > x_0^2 - y_0^2$, i.e. the maximal open interval for ϕ to be defined on is $I' =]\sqrt{x_0^2 - y_0^2}, \infty[$.

5.3 Linear ODE, Variation of Constants

Definition 5.9. Let $I \subseteq \mathbb{R}$ be an open interval and let $a, b : I \longrightarrow \mathbb{R}$ be continuous functions. An ODE of the form

$$y' = a(x)y + b(x) \tag{5.30}$$

is called a *linear ODE* of first order. It is called *homogeneous* if, and only if, $b \equiv 0$; it is called *inhomogeneous* if, and only if, it is not homogeneous.

Theorem 5.10 (Variation of Constants)**.** *Let $I \subseteq \mathbb{R}$ be an open interval and let $a, b : I \longrightarrow \mathbb{R}$ be continuous. Moreover, let $x_0 \in I$ and $c \in \mathbb{R}$. Then the linear ODE* (5.30) *has a unique solution $\phi : I \longrightarrow \mathbb{R}$ that satisfies the initial condition $y(x_0) = c$. This unique solution is given by*

$$\phi : I \longrightarrow \mathbb{R}, \quad \phi(x) = \phi_0(x) \left(c + \int_{x_0}^{x} \phi_0(t)^{-1}\, b(t)\,\mathrm{d}t \right), \tag{5.31a}$$

where

$$\phi_0 : I \longrightarrow \mathbb{R}, \quad \phi_0(x) = \exp\left(\int_{x_0}^{x} a(t)\,\mathrm{d}t \right) = e^{\int_{x_0}^{x} a(t)\,\mathrm{d}t}. \tag{5.31b}$$

Here, and in the following, ϕ_0^{-1} denotes $1/\phi_0$ and not the inverse function of ϕ_0 (which does not even necessarily exist).

Proof. We begin by noting that ϕ_0 according to (5.31b) is well-defined since a is assumed to be continuous, i.e., in particular, Riemann integrable on $[x_0, x]$. Moreover, the fundamental theorem of calculus [Phi15a, Th. 10.19(a)] applies, showing ϕ_0 is differentiable with

$$\phi_0' : I \longrightarrow \mathbb{R}, \quad \phi_0'(x) = a(x) \exp\left(\int_{x_0}^{x} a(t)\,\mathrm{d}t \right) = a(x)\phi_0(x), \tag{5.32}$$

where the chain rule [Phi15a, (9.15)] was used as well. In particular, ϕ_0 is continuous. Since $\phi_0 > 0$ as well, ϕ_0^{-1} is also continuous. Moreover, as b is continuous by hypothesis, $\phi_0^{-1}\, b$ is continuous and, thus, Riemann integrable on $[x_0, x]$. Once again, [Phi15a, Th. 10.19(a)] applies, yielding ϕ to be differentiable with

$$\begin{aligned} \phi' &: I \longrightarrow \mathbb{R}, \\ \phi'(x) &= \phi_0'(x) \left(c + \int_{x_0}^{x} \phi_0(t)^{-1}\, b(t)\, \mathrm{d}t \right) + \phi_0(x)\phi_0(x)^{-1}\, b(x) \\ &= a(x)\phi_0(x) \left(c + \int_{x_0}^{x} \phi_0(t)^{-1}\, b(t)\, \mathrm{d}t \right) + b(x) = a(x)\phi(x) + b(x), \end{aligned} \tag{5.33}$$

where the product rule [Phi15a, Th. 9.6(c)] was used as well. Comparing (5.33) with (5.30) shows that ϕ is a solution to (5.30). The computation

$$\phi(x_0) = \phi_0(x_0)\,(c + 0) = 1 \cdot c = c \tag{5.34}$$

verifies that ϕ satisfies the desired initial condition. It remains to prove uniqueness. To this end, let $\psi : I \longrightarrow \mathbb{R}$ be an arbitrary differentiable function that satisfies (5.30) as well as the initial condition $\psi(x_0) = c$. We have to show $\psi = \phi$. Since $\phi_0 > 0$, we can define $u := \psi/\phi_0$ and still have to verify

$$\underset{x \in I}{\forall} \quad u(x) = c + \int_{x_0}^{x} \phi_0(t)^{-1}\, b(t)\, \mathrm{d}t\,. \tag{5.35}$$

We obtain

$$a\,\phi_0\,u + b = a\,\psi + b = \psi' = (\phi_0\,u)' = \phi_0'\,u + \phi_0\,u' = a\,\phi_0\,u + \phi_0\,u', \tag{5.36}$$

implying $b = \phi_0\, u'$ and $u' = \phi_0^{-1}\, b$. Thus, the fundamental theorem of calculus in the form [Phi15a, (10.50b)], implies

$$\underset{x \in I}{\forall} \quad u(x) = u(x_0) + \int_{x_0}^{x} u'(t)\, \mathrm{d}t \; = c + \int_{x_0}^{x} \phi_0(t)^{-1}\, b(t)\, \mathrm{d}t\,, \tag{5.37}$$

thereby completing the proof. ■

Corollary 5.11. *Let $I \subseteq \mathbb{R}$ be an open interval and let $a : I \longrightarrow \mathbb{R}$ be continuous. Moreover, let $x_0 \in I$ and $c \in \mathbb{R}$. Then the homogeneous linear ODE* (5.30) *(i.e. with $b \equiv 0$) has a unique solution $\phi : I \longrightarrow \mathbb{R}$ that satisfies the initial condition $y(x_0) = c$. This unique solution is given by*

$$\phi(x) = c \exp\left(\int_{x_0}^{x} a(t)\, \mathrm{d}t \right) = c\, e^{\int_{x_0}^{x} a(t)\, \mathrm{d}t}. \tag{5.38}$$

Proof. One immediately obtains (5.38) by setting $b \equiv 0$ in in (5.31). ■

Remark 5.12. The name *variation of constants* for Th. 5.10 can be understood from comparing the solution (5.38) of the homogeneous linear ODE with the solution (5.31) of the general inhomogeneous linear ODE: One obtains (5.31) from (5.38) by *varying the constant c*, i.e. by replacing it with the function $x \mapsto c + \int_{x_0}^{x} \phi_0(t)^{-1}\, b(t)\, \mathrm{d}t$.

Example 5.13. (a) Applying Cor. 5.11 to the homogeneous linear ODE

$$y' = ky \tag{5.39}$$

with $k \in \mathbb{R}$ and initial condition $y(x_0) = c$ ($x_0, c \in \mathbb{R}$) yields the unique solution

$$\phi : \mathbb{R} \longrightarrow \mathbb{R}, \quad \phi(x) = c \exp\left(\int_{x_0}^{x} k \, \mathrm{d}t\right) = ce^{k(x-x_0)}. \tag{5.40}$$

(b) We can use Cor. 5.11 to recompute the solution to the ODE (5.21) from Example 5.7, since this constitutes a homogeneous linear ODE with $a(x) = -1/x$. For the initial condition $y(1) = c$, we obtain

$$\phi(x) = c \exp\left(-\int_{1}^{x} \frac{\mathrm{d}t}{t}\right) = ce^{-\ln x} = \frac{c}{x}. \tag{5.41}$$

(c) Consider the ODE

$$y' = 2xy + x^3 \tag{5.42}$$

with initial condition $y(0) = c$, $c \in \mathbb{R}$. Comparing (5.39) with Def. 5.9, we observe we are facing an inhomogeneous linear ODE with

$$a : \mathbb{R} \longrightarrow \mathbb{R}, \quad a(x) := 2x, \tag{5.43a}$$

$$b : \mathbb{R} \longrightarrow \mathbb{R}, \quad b(x) := x^3. \tag{5.43b}$$

From Cor. 5.11, we obtain the solution $\phi_{0,c}$ to the homogeneous version of (5.39):

$$\phi_{0,c} : \mathbb{R} \longrightarrow \mathbb{R}, \quad \phi_{0,c}(x) = c \exp\left(\int_{0}^{x} a(t) \, \mathrm{d}t\right) = ce^{x^2}. \tag{5.44}$$

The solution to (5.39) is given by (5.31a):

$$\begin{aligned} &\phi : \mathbb{R} \longrightarrow \mathbb{R}, \\ &\phi(x) = e^{x^2} \left(c + \int_{0}^{x} e^{-t^2} t^3 \, \mathrm{d}t\right) = e^{x^2} \left(c + \left[-\frac{1}{2}(t^2+1)e^{-t^2}\right]_{0}^{x}\right) \\ &\quad = e^{x^2} \left(c + \frac{1}{2} - \frac{1}{2}(x^2+1)e^{-x^2}\right) = \left(c + \frac{1}{2}\right) e^{x^2} - \frac{1}{2}(x^2+1). \end{aligned} \tag{5.45}$$

5.4 Change of Variables

To solve an ODE, it can be useful to transform it into an equivalent ODE, using a so-called *change of variables*. If one already knows how to solve the transformed ODE, then the equivalence allows one to also solve the original ODE. The presentation of the material in the present section is somewhat reversed as compared to the presentation in Sec. 5.2 above on separation of variables: Here, we will first present a heuristic procedure that is often used in practise in Rem. 5.14, followed by an illustrating example. Only then will we provide the rigorous Th. 5.16 that constitutes the basis of the heuristic procedure, and we will conclude with an application of Th. 5.16 to solve so-called Bernoulli differential equations.

Remark 5.14. For the initial value problem $y' = f(x,y)$, $y(x_0) = y_0$, the heuristic change of variables procedure proceeds as follows:

(1) One introduces the new variable $z := T(x,y)$ and then computes z', i.e. the derivative of the function $x \mapsto z(x) = T(x, y(x))$.

(2) In the result of (1), one eliminates all occurrences of the variable y by first replacing y' by $f(x,y)$ and then replacing y by $T_x^{-1}(z)$, where $T_x(y) := T(x,y) = z$ (i.e. one has to solve the equation $z = T(x,y)$ for y). One thereby obtains the transformed initial value problem problem $z' = g(x,z)$, $z(x_0) = T(x_0, y_0)$, with a suitable function g.

(3) One solves the transformed initial value problem to obtain a solution μ, and then $x \mapsto \phi(x) := T_x^{-1}\big(\mu(x)\big)$ yields a candidate for a solution to the original initial value problem.

(4) One checks that ϕ is, indeed, a solution to $y' = f(x,y)$, $y(x_0) = y_0$.

Example 5.15. Consider

$$f : \mathbb{R}^+ \times \mathbb{R} \longrightarrow \mathbb{R}, \quad f(x,y) := 1 + \frac{y}{x} + \frac{y^2}{x^2}, \tag{5.46}$$

and the initial value problem

$$y' = f(x,y), \quad y(1) = 0. \tag{5.47}$$

We introduce the change of variables $z := T(x,y) := y/x$ and proceed according to the steps of Rem. 5.14. According to (1), we compute, using the quotient rule,

$$z'(x) = \frac{y'(x)\,x - y(x)}{x^2}. \tag{5.48}$$

According to (2), we replace $y'(x)$ by $f(x,y)$ and then replace y by $T_x^{-1}(z) = xz$ to obtain the transformed initial value problem

$$z' = \frac{1}{x}\left(1 + \frac{y}{x} + \frac{y^2}{x^2}\right) - \frac{y}{x^2} = \frac{1}{x}\,(1 + z + z^2) - \frac{z}{x} = \frac{1+z^2}{x}, \quad z(1) = 0/1 = 0. \tag{5.49}$$

According to (3), we next solve (5.49), e.g. by seperation of variables, to obtain the solution

$$\mu : \left]e^{-\frac{\pi}{2}},\, e^{\frac{\pi}{2}}\right[\longrightarrow \mathbb{R}, \quad \mu(x) := \tan \ln x, \tag{5.50}$$

of (5.49), and

$$\phi : \left]e^{-\frac{\pi}{2}},\, e^{\frac{\pi}{2}}\right[\longrightarrow \mathbb{R}, \quad \phi(x) := x\,\mu(x) = x\,\tan \ln x, \tag{5.51}$$

as a candidate for a solution to (5.47). Finally, according to (4), we check that ϕ is, indeed, a solution to (5.47): Due to $\phi(1) = 1 \cdot \tan 0 = 0$, ϕ satisfies the initial condition, and due to

$$\begin{aligned}\phi'(x) &= \tan\ln x + x\,\frac{1}{x}\,(1+\tan^2\ln x) = 1 + \tan\ln x + \tan^2\ln x \\ &= 1 + \frac{\phi(x)}{x} + \frac{\phi^2(x)}{x^2},\end{aligned} \tag{5.52}$$

ϕ satisfies the ODE.

Theorem 5.16. *Let $G \subseteq \mathbb{R}\times\mathbb{R}$ be open, $n \in \mathbb{N}$, $f : G \longrightarrow \mathbb{R}$, and $(x_0, y_0) \in G$. Define*

$$\underset{x\in\mathbb{R}}{\forall} \quad G_x := \{y \in \mathbb{R} : (x,y) \in G\} \tag{5.53}$$

and assume the change of variables *function $T : G \longrightarrow \mathbb{R}$ is differentiable and such that*

$$\underset{G_x\neq\emptyset}{\forall} \quad \Big(T_x := T(x,\cdot) : G_x \longrightarrow T_x(G_x), \quad T_x(y) := T(x,y), \quad \textit{is a diffeomorphism}\Big), \tag{5.54}$$

i.e. T_x is invertible and both T_x and T_x^{-1} are differentiable. Then the first-order initial value problems

$$y' = f(x,y), \tag{5.55a}$$
$$y(x_0) = y_0, \tag{5.55b}$$

and

$$y' = \frac{f\big(x, T_x^{-1}(y)\big)}{(T_x^{-1})'(y)} + \partial_x T\big(x, T_x^{-1}(y)\big), \tag{5.56a}$$
$$y(x_0) = T(x_0, y_0), \tag{5.56b}$$

are equivalent in the following sense:

(a) *A differentiable function $\phi : I \longrightarrow \mathbb{R}$, where $I \subseteq \mathbb{R}$ is a nontrivial interval, is a solution to* (5.55a) *if, and only if, the function*

$$\mu : I \longrightarrow \mathbb{R}, \quad \mu(x) := (T_x \circ \phi)(x) = T\big(x, \phi(x)\big), \tag{5.57}$$

is a solution to (5.56a).

(b) *A differentiable function $\phi : I \longrightarrow \mathbb{R}$, where $I \subseteq \mathbb{R}$ is a nontrivial interval, is a solution to* (5.55) *if, and only if, the function of* (5.57) *is a solution to* (5.56).

Proof. We start by noting that the assumption of G being open clearly implies each G_x, $x \in \mathbb{R}$, to be open as well, which, in turn, implies $T_x(G_x)$ to be open. Next, for each $x \in \mathbb{R}$ such that $G_x \neq \emptyset$, we can apply the chain rule [Phi15a, Th. 9.10] to $T_x \circ T_x^{-1} = \mathrm{Id}$ to obtain

$$\underset{y\in T_x(G_x)}{\forall} \quad T_x'\big(T_x^{-1}(y)\big)\,\big(T_x^{-1}\big)'(y) = 1 \tag{5.58}$$

and, thus, each $\big(T_x^{-1}\big)'(y) \neq 0$ with

$$\underset{y\in T_x(G_x)}{\forall} \quad \Big(\big(T_x^{-1}\big)'(y)\Big)^{-1} = T_x'\big(T_x^{-1}(y)\big). \tag{5.59}$$

Consider ϕ and μ as in (a) and notice that (5.57) implies

$$\underset{x\in I}{\forall} \quad \phi(x) = T_x^{-1}(\mu(x)). \tag{5.60}$$

Moreover, the differentiability of ϕ and T imply differentiability of μ by the chain rule of Th. 2.28, which also yields

$$\begin{aligned} \underset{x\in I}{\forall} \qquad \mu'(x) &= \Big(\partial_x T\big(x,\phi(x)\big) \quad \partial_y T\big(x,\phi(x)\big)\Big) \begin{pmatrix} 1 \\ \phi'(x) \end{pmatrix} \\ &= T_x'(\phi(x))\,\phi'(x) + \partial_x T\big(x,\phi(x)\big). \end{aligned} \tag{5.61}$$

To prove (a), first assume $\phi : I \longrightarrow \mathbb{R}$ to be a solution of (5.55a). Then, for each $x \in I$,

$$\begin{aligned} \mu'(x) &\overset{(5.61),(5.55a)}{=} T_x'(\phi(x))\, f\big(x,\phi(x)\big) + \partial_x T\big(x,\phi(x)\big) \\ &\overset{(5.60)}{=} T_x'\big(T_x^{-1}(\mu(x))\big)\, f\big(x,T_x^{-1}(\mu(x))\big) + \partial_x T\big(x,T_x^{-1}(\mu(x))\big) \\ &\overset{(5.59)}{=} \frac{f\big(x,T_x^{-1}(\mu(x))\big)}{(T_x^{-1})'(\mu(x))} + \partial_x T\big(x,T_x^{-1}(\mu(x))\big), \end{aligned} \tag{5.62}$$

showing μ satisfies (5.56a). Conversely, assume μ to be a solution to (5.56a). Then, for each $x \in I$,

$$\frac{f\big(x,T_x^{-1}(\mu(x))\big)}{(T_x^{-1})'(\mu(x))} + \partial_x T\big(x,T_x^{-1}(\mu(x))\big) \overset{(5.56a)}{=} \mu'(x) \overset{(5.61)}{=} T_x'(\phi(x))\,\phi'(x) + \partial_x T\big(x,\phi(x)\big). \tag{5.63}$$

Using (5.60), one can subtract the second summand from (5.63). Multiplying the result by $(T_x^{-1})'(\mu(x))$ and taking into account (5.59) then provides

$$\underset{x\in I}{\forall} \quad \phi'(x) = f\big(x,T_x^{-1}(\mu(x))\big) \overset{(5.60)}{=} f\big(x,\phi(x)\big), \tag{5.64}$$

showing ϕ satisfies (5.55a).

It remains to prove (b). If ϕ satisfies (5.55), then μ satisfies (5.56a) by (a). Moreover, $\mu(x_0) = T\big(x_0,\phi(x_0)\big) = T(x_0,y_0)$, i.e. μ satisfies (5.56b) as well. Conversely, assume μ satisfies (5.56). Then ϕ satisfies (5.55a) by (a). Moreover, by (5.60), $\phi(x_0) = T_{x_0}^{-1}(\mu(x_0)) = T_{x_0}^{-1}(T(x_0,y_0)) = y_0$, showing ϕ satisfies (5.55b) as well. ■

As an application of Th. 5.16, we prove the following theorem about so-called *Bernoulli differential equations*:

Theorem 5.17. *Consider the* Bernoulli differential equation

$$y' = f(x,y) := a(x)\,y + b(x)\,y^\alpha, \tag{5.65a}$$

where $\alpha \in \mathbb{R}\setminus\{0,1\}$*, the functions* $a, b : I \longrightarrow \mathbb{R}$ *are continuous and defined on an open interval* $I \subseteq \mathbb{R}$*, and* $f : I \times \mathbb{R}^+ \longrightarrow \mathbb{R}$*. For* (5.65a)*, we add the initial condition*

$$y(x_0) = y_0, \quad (x_0, y_0) \in I \times \mathbb{R}^+, \tag{5.65b}$$

and, furthermore, we also consider the corresponding linear initial value problem

$$y' = (1-\alpha)\big(a(x)\,y + b(x)\big), \tag{5.66a}$$

$$y(x_0) = y_0^{1-\alpha}, \tag{5.66b}$$

with its unique solution $\psi : I \longrightarrow \mathbb{R}$ *given by Th.* 5.10.

(a) Uniqueness: *On each open interval* $I' \subseteq I$ *satisfying* $x_0 \in I'$ *and* $\psi > 0$ *on* I'*, the Bernoulli initial value problem* (5.65) *has a unique solution. This unique solution is given by*

$$\phi : I' \longrightarrow \mathbb{R}^+, \quad \phi(x) := \big(\psi(x)\big)^{\frac{1}{1-\alpha}}. \tag{5.67}$$

(b) Existence: *There exists an open interval* $I' \subseteq I$ *satisfying* $x_0 \in I'$ *and* $\psi > 0$ *on* I'*, i.e. an* I' *such that* (a) *applies.*

Proof. (b) is immediate from Th. 5.10, since $\psi(x_0) = y_0 > 0$ and ψ is continuous.

To prove (a), we apply Th. 5.16 with the change of variables

$$T : I \times \mathbb{R}^+ \longrightarrow \mathbb{R}^+, \quad T(x,y) := y^{1-\alpha}. \tag{5.68}$$

Then $T \in C^1(I \times \mathbb{R}^+, \mathbb{R})$ with $\partial_x T \equiv 0$ and $\partial_y T(x,y) = (1-\alpha)\,y^{-\alpha}$. Moreover,

$$\bigvee_{x\in I} \quad T_x = S, \quad S : \mathbb{R}^+ \longrightarrow \mathbb{R}^+, \quad S(y) := y^{1-\alpha}, \tag{5.69}$$

which is differentiable with the differentiable inverse function $S^{-1} : \mathbb{R}^+ \longrightarrow \mathbb{R}^+$, $S^{-1}(y) = y^{\frac{1}{1-\alpha}}$, $(S^{-1})'(y) = \frac{1}{1-\alpha}\,y^{\frac{\alpha}{1-\alpha}}$. Thus, (5.56a) takes the form

$$\begin{aligned} y' &= \frac{f\big(x, T_x^{-1}(y)\big)}{(T_x^{-1})'(y)} + \partial_x T\big(x, T_x^{-1}(y)\big) \\ &= (1-\alpha)\,y^{-\frac{\alpha}{1-\alpha}}\Big(a(x)\,y^{\frac{1}{1-\alpha}} + b(x)\,\big(y^{\frac{1}{1-\alpha}}\big)^\alpha\Big) + 0 \\ &= (1-\alpha)\big(a(x)\,y + b(x)\big). \end{aligned} \tag{5.70}$$

Thus, if $I' \subseteq I$ is such that $x_0 \in I'$ and $\psi > 0$ on I', then Th. 5.16 says ϕ defined by (5.67) must be a solution to (5.65) (note that the differentiability of ψ implies the differentiability of ϕ). On the other hand, if $\lambda : I' \longrightarrow \mathbb{R}^+$ is an arbitrary solution to (5.65), then Th. 5.16 states $\mu := S \circ \lambda = \lambda^{1-\alpha}$ to be a solution to (5.66). The uniqueness part of Th. 5.10 then yields $\lambda^{1-\alpha} = \psi\restriction_{I'} = \phi^{1-\alpha}$, i.e. $\lambda = \phi$. ■

Finding a suitable change of variables to transform a given ODE such that one is in a position to solve the transformed ODE is an art, i.e. it can be very difficult to spot a useful transformation, and it takes a lot of practise and experience.

A Linear Algebra

A.1 Vector Spaces

In [Phi15a], we encountered the abstract definition of a field in [Phi15a, Def. 4.4], and we studied the fields $\mathbb{Q}$, $\mathbb{R}$, and $\mathbb{C}$. Even though we will formulate the following definitions and results using a general field F as defined in [Phi15a, Def. 4.4], for the purposes of the present lecture, you may always think of F as being the field of real numbers $\mathbb{R}$ or the field of complex numbers $\mathbb{C}$.

Definition A.1. Let F be a field and let V be a nonempty set with two maps

$$\begin{aligned} + &: V \times V \longrightarrow V, \quad (x, y) \mapsto x + y, \\ \cdot &: F \times V \longrightarrow V, \quad (\lambda, x) \mapsto \lambda \cdot x \end{aligned} \tag{A.1}$$

($+$ is called *(vector) addition* and $\cdot$ is called *scalar multiplication*; often one writes xy instead of $x \cdot y$ – please take care not to confuse the vector addition on V with the addition on F and, likewise, not to confuse the scalar multiplication with the multiplication on F, the symbol $+$ is used for both additions and $\cdot$ is used for both multiplications, but you can always determine from the context which addition or multiplication is meant). Then V is called a *vector space* or a *linear space* over F (sometimes also an F-vector space) if, and only if, the following conditions are satisfied:

(i) V is a commutative group with respect to $+$. The neutral element with respect to $+$ is denoted 0 (do not confuse $0 \in V$ with $0 \in F$ – once again, the same symbol is used for different objects (both objects only coincide for $F = V$)).

(ii) Distributivity:

$$\forall_{\lambda \in F} \quad \forall_{x,y \in V} \quad \lambda(x + y) = \lambda x + \lambda y, \tag{A.2a}$$

$$\forall_{\lambda,\mu \in F} \quad \forall_{x \in V} \quad (\lambda + \mu)x = \lambda x + \mu x. \tag{A.2b}$$

(iii) Compatibility between Multiplication on F and Scalar Multiplication:

$$\forall_{\lambda,\mu \in F} \quad \forall_{x \in V} \quad (\lambda\mu)x = \lambda(\mu x). \tag{A.3}$$

(iv) The neutral element with respect to the multiplication on F is also neutral with respect to the scalar multiplication:

$$\forall_{x \in V} \quad 1x = x. \tag{A.4}$$

If V is a vector space over F, then one calls the elements of V *vectors* and the elements of F *scalars*.

Example A.2. (a) Every field F is a vector space over itself if one uses the field addition in F as the vector addition and the field multiplication in F as the scalar multiplication (as important special cases, we obtain that $\mathbb{R}$ is a vector space over $\mathbb{R}$ and $\mathbb{C}$ is a vector space over $\mathbb{C}$): All the vector space laws are immediate consequences of the corresponding field laws: Def. A.1(i) holds as every field is a commutative group with respect to addition; Def. A.1(ii) follows from the field distributivity [Phi15a, Def. (4.5)] and multiplicative commutativity on F; Def. A.1(iii) is merely the multiplicative associativity on F; and Def. A.1(iv) holds, since scalar multiplication coincides with field multiplication on F.

(b) The reasoning in (a) actually shows that every field F is a vector space over every subfield E of F (over every $E \subseteq F$ that is a field with respect to the addition and multiplication defined on F). In particular, $\mathbb{R}$ is a vector space over $\mathbb{Q}$.

(c) If A is any nonempty set, F is a field, and Y is a vector space over the field F, then we can make $V := \mathcal{F}(A, Y) = Y^A$ (the set of functions from A into Y) into a vector space over F by defining for each $f, g : A \longrightarrow Y$:

$$(f+g) : A \longrightarrow Y, \qquad (f+g)(x) := f(x) + g(x), \tag{A.5a}$$

$$(\lambda \cdot f) : A \longrightarrow Y, \qquad (\lambda \cdot f)(x) := \lambda \cdot f(x) \quad \text{for each } \lambda \in F \tag{A.5b}$$

(note that, for $Y = F = \mathbb{K}$, the above definitions are the same as the ones in [Phi15a, (6.1a)] and [Phi15a, (6.1b)], respectively).

It is an exercise to verify that $(V, +, \cdot)$ is, indeed, a vector space over F.

(d) For each $n \in \mathbb{N}$, $(\mathbb{K}^n, +, \cdot)$, with vector addition and scalar multiplication as defined in (1.1a) and (1.1c), respectively, constitutes a vector space over $\mathbb{K}$. The validity of Def. A.1(i) – Def. A.1(iv) can easily be verified directly, but $(\mathbb{K}^n, +, \cdot)$ can also be seen as a special case of (c) with $A = \{1, \ldots, n\}$ and $Y = F = \mathbb{K}$. To this end, recall that, according to [Phi15a, Ex. 2.15(c)], $\mathbb{K}^n = \mathbb{K}^{\{1,\ldots,n\}} = \mathcal{F}\big(\{1, \ldots, n\}, \mathbb{K}\big)$ is the set of functions from $\{1, \ldots, n\}$ into $\mathbb{K}$. Then $z = (z_1, \ldots, z_n) \in \mathbb{K}^n$ is the same as the function $f : \{1, \ldots, n\} \longrightarrow \mathbb{K}$, $f(j) = z_j$. Thus, (1.1a) is, indeed, the same as (A.5a), and (1.1c) is, indeed, the same as (A.5b).

Definition and Remark A.3. Let $(V, +, \cdot)$ be a vector space over the field F. A subset $U \subseteq V$ is called a *subspace* of V if, and only if, U is a vector space over F with respect to operations $+$ and $\cdot$ it inherits from V. Clearly, every law (commutativity, associativity, etc.) that holds on V must, in particular, hold on U, showing that $\emptyset \neq U \subseteq V$ is a subspace of V if, and only if,

$$\forall_{x,y \in U} \quad x + y \in U, \tag{A.6a}$$

$$\wedge \quad \forall_{\lambda \in F} \quad \forall_{x \in U} \quad \lambda x \in U. \tag{A.6b}$$

which holds if, and only if,

$$\forall_{\lambda,\mu \in F} \quad \forall_{x,y \in U} \quad \lambda x + \mu y \in U. \tag{A.7}$$

Example A.4. (a) $\mathbb{Q}$ is *not* a subspace of $\mathbb{R}$ if $\mathbb{R}$ is considered as a vector space over $\mathbb{R}$ (for example, $\sqrt{2} \cdot 2 \notin \mathbb{Q}$). However, $\mathbb{Q}$ *is* a subspace of $\mathbb{R}$ if $\mathbb{R}$ is considered as a vector space over $\mathbb{Q}$.

(b) From Ex. A.2(c), we know that, for each $\emptyset \neq A$, $\mathcal{F}(A, \mathbb{K})$ constitutes a vector space over $\mathbb{K}$. Thus, as a consequence of Def. and Rem. A.3, a subset of $\mathcal{F}(A, \mathbb{K})$ is a vector space over $\mathbb{K}$ if, and only if, it is closed under addition and scalar multiplication. By using results from [Phi15a], we obtain the following examples:

(i) The set $\mathcal{P}(\mathbb{K})$ of all polynomials mapping from $\mathbb{K}$ into $\mathbb{K}$ is a vector space over $\mathbb{K}$ by [Phi15a, Rem. 6.4]; for each $n \in \mathbb{N}_0$, the set $\mathcal{P}_n(\mathbb{K})$ of all such polynomials of degree at most n is also a vector space over $\mathbb{K}$ by [Phi15a, Rem. (6.4a),(6.4b)].

(ii) If $\emptyset \neq M \subseteq \mathbb{C}$, then the set of continuous functions from M into $\mathbb{K}$, i.e. $C(M, \mathbb{K})$, is a vector space over $\mathbb{K}$ by [Phi15a, Th. 7.38].

(iii) If $a, b \in \mathbb{R} \cup \{-\infty, \infty\}$ and $a < b$, then the set of differentiable functions from $I :=]a, b[$ into $\mathbb{K}$ is a vector space over $\mathbb{K}$ by [Phi15a, Th. 9.6(a),(b)]. Moreover, [Phi15a, Th. 9.6(a),(b)] also implies that, for each $k \in \mathbb{N}$, the set of k times differentiable functions from I into $\mathbb{K}$ is a vector space over $\mathbb{K}$, and so is each set $C^k(I, \mathbb{K})$ of k times continuously differentiable functions ([Phi15a, Th. 7.38] is also needed for the last conclusion). The set $C^\infty(I, \mathbb{K})$ is also a vector space over $\mathbb{K}$ by Th. A.5(a) below.

Theorem A.5. *Let V be a vector space over the field F.*

(a) *Let $I \neq \emptyset$ be an index set and $(U_i)_{i \in I}$ a family of subspaces of V. Then the intersection $U := \bigcap_{i \in I} U_i$ is also a subspace of V.*

(b) *In contrast to intersections, unions of subspaces are* almost never *subspaces. More precisely, if U_1 and U_2 are subspaces of V, then*

$$U_1 \cup U_2 \text{ is subspace of } V \quad \Leftrightarrow \quad \Big(U_1 \subseteq U_2 \ \vee \ U_2 \subseteq U_1 \Big). \tag{A.8}$$

Proof. See, e.g., [Str08, Th. 8.7]. ■

Definition A.6. Let V be a vector space over the field F.

(a) Let $n \in \mathbb{N}$ and $v_1, \ldots, v_n \in V$. A vector $v \in V$ is called a *linear combination* of $v_1, \ldots, v_n$ if, and only if, there exist $\lambda_1, \ldots, \lambda_n \in F$ such that

$$v = \sum_{i=1} \lambda_i \, v_i. \tag{A.9}$$

Morover, $v \in V$ is called *linearly dependent* of a subset U of V if, and only if, there exists $n \in \mathbb{N}$ and $u_1, \ldots, u_n \in U$ such that v is a linear combination of $u_1, \ldots, u_n$.

(b) A subset U of V is called *linearly independent* if, and only if,

$$\left(n \in \mathbb{N} \quad \wedge \quad u_1, \ldots, u_n \in U \quad \wedge \quad \lambda_1, \ldots, \lambda_n \in F \quad \wedge \quad \sum_{i=1} \lambda_i \, u_i = 0 \right)$$
$$\Rightarrow \quad \lambda_1 = \cdots = \lambda_n = 0. \tag{A.10a}$$

Occasionally, one also wants to have the notion available for families of vectors instead of sets, and one calls a family $(u_i)_{i \in I}$ of vectors in V *linearly independent* if, and only if,

$$\left(n \in \mathbb{N} \quad \wedge \quad i_1, \ldots, i_n \in I \quad \wedge \quad \lambda_1, \ldots, \lambda_n \in F \quad \wedge \quad \sum_{k=1} \lambda_i \, u_{i_k} = 0 \right)$$
$$\Rightarrow \quad \lambda_1 = \cdots = \lambda_n = 0. \tag{A.10b}$$

Sets and families that are not linearly independent are called *linearly dependent*.

Definition A.7. Let V be a vector space over the field F, $A \subseteq V$, and

$$\mathcal{U} := \big\{ U \in \mathcal{P}(V) : \; A \subseteq U \; \wedge \; U \text{ is subspace of } V \big\}. \tag{A.11}$$

Then the set

$$\langle A \rangle := \operatorname{span} A := \bigcap_{U \in \mathcal{U}} U \tag{A.12}$$

is called the *span* of A. Moreover A is called a *spanning set* of V if, and only if, $\langle A \rangle = V$.

Proposition A.8. *Let V be a vector space over the field F and $A \subseteq V$.*

(a) *$\langle A \rangle$ is a subspace of V, namely the smallest subspace of V containing A.*

(b) *If $A = \emptyset$, then $\langle A \rangle = \{0\}$; if $A \neq \emptyset$, then $\langle A \rangle$ is the set of all linear combinations of elements from A, i.e.*

$$\langle A \rangle = \left\{ \sum_{i=1}^{n} \lambda_i \, a_i : \; n \in \mathbb{N} \; \wedge \; \lambda_1, \ldots, \lambda_n \in F \; \wedge \; a_1, \ldots, a_n \in A \right\}. \tag{A.13}$$

Proof. (a) is immediate from Th. A.5(a).

(b): For the case $A = \emptyset$, note that $\{0\}$ is a subspace of V, and that $\{0\}$ is contained in every subspace of V. For $A \neq \emptyset$, let W denote the right-hand side of (A.13), and recall from (A.12) that $\langle A \rangle$ is the intersection of all subspaces U of V that contain A. If U is a subspace of V and $A \subseteq U$, then $W \subseteq U$, since U is closed under vector addition and scalar multiplication, showing $W \subseteq \langle A \rangle$. On the other hand, W is clearly a subspace of V that contains A, showing $\langle A \rangle \subseteq W$, completing the proof of $\langle A \rangle = W$. ■

Definition A.9. Let V be a vector space over the field F and $B \subseteq V$.

(a) B is called a *generating set* for V if, and only if, $\langle B\rangle = V$. One then also says that V is *generated* or *spanned* by B.

(b) B is called a *basis* for V if, and only if, B is a generating set for V that is also linearly independent (see Def. A.6(b)).

Theorem A.10. *Let V be a vector space over the field F and $B \subseteq V$. Then the following statements* (i) – (iii) *are equivalent:*

(i) *B is a basis for V.*

(ii) *B is a maximal linearly independent subset of V, i.e. B is linearly independent and each set $A \subseteq V$ with $B \subsetneq A$ is linearly dependent.*

(iii) *B is a minimal generating set for V, i.e. $\langle B\rangle = V$ and $\langle A\rangle \subsetneq V$ for each $A \subsetneq B$.*

Proof. See, e.g., [Str08, Th. 9.6]. ■

Theorem A.11 (Coordinates)**.** *Let V be a vector space over the field F and assume $B \subseteq V$ is a basis of V. Then each vector $v \in V$ has unique* coordinates *with respect to the basis B, i.e., for each $v \in V$, there exists a unique finite subset B_v of B and a unique map $c\colon B_v \longrightarrow F \setminus \{0\}$ such that*

$$v = \sum_{b\in B_v} c(b)\, b. \tag{A.14}$$

Note that, for $v = 0$, one has $B_v = \emptyset$, c is the empty map, and (A.14) *becomes $0 = \sum_{b\in\emptyset} c(b)\, b$, employing the useful convention that sums over the empty set are defined as* 0.

Proof. The existence of B_v and the map c follows from the fact that the basis B is a generating set, $\langle B\rangle = V$. For the uniqueness proof, consider finite sets $B_v, \tilde{B}_v \subseteq B$ and maps $c\colon B_v \longrightarrow F \setminus \{0\}$, $\tilde{c}\colon \tilde{B}_v \longrightarrow F \setminus \{0\}$ such that

$$v = \sum_{b\in B_v} c(b)\, b = \sum_{b\in \tilde{B}_v} \tilde{c}(b)\, b. \tag{A.15}$$

Extend both c and $\tilde{c}$ to $A := B_v \cup \tilde{B}_v$ by letting $c(b) := 0$ for $b \in \tilde{B}_v \setminus B_v$ and $\tilde{c}(b) := 0$ for $b \in B_v \setminus \tilde{B}_v$. Then

$$0 = \sum_{b\in A} \big(c(b) - \tilde{c}(b)\big)\, b, \tag{A.16}$$

such that the linear independence of A implies $c(b) = \tilde{c}(b)$ for each $b \in A$, which, in turn, implies $B_v = \tilde{B}_v$ and $c = \tilde{c}$. ■

Remark A.12. If the basis B of V has finitely many elements, then one often enumerates the elements $B = \{b_1, \ldots, b_n\}$, $n = \#B \in \mathbb{N}$, and writes $\lambda_i := c(b_i)$ for $b_i \in B_v$, $\lambda_i := 0$ for $b_i \notin B_v$, such that (A.14) takes the, perhaps more familiar looking, form

$$v = \sum_{i=1}^{n} \lambda_i b_i. \tag{A.17}$$

Theorem A.13. *Every vector space V over a field F has a basis $B \subseteq V$. Moreover, bases of V have a unique cardinality, i.e. if $B \subseteq V$ and $\tilde{B} \subseteq V$ are both bases of V, then $\#B = \#\tilde{B}$.*

Proof. See, e.g., [Str08, Lem. 11.3, Th. 11.5]. ■

Definition A.14. According to Th. A.13, for each vector space V over a field F, the cardinality of its bases is unique. This unique cardinality is called the *dimension* of V and is denoted $\dim V$. If $\dim V < \infty$ (i.e. $\dim V \in \mathbb{N}_0$), then V is called *finite dimensional*, otherwise *infinite dimensional*.

Example A.15. Given a field F and a nonempty set I, let F^I_{fin} denote the set of functions $f : I \longrightarrow F$ such that there exists a finite set $I_f \subseteq I$ satisfying

$$f(i) = 0 \quad \text{for each } i \in I \setminus I_f, \tag{A.18a}$$

$$f(i) \neq 0 \quad \text{for each } i \in I_f. \tag{A.18b}$$

Then $F^I_{\text{fin}} = F^I$ if, and only if, I is finite (for example $F^n_{\text{fin}} = F^n$ for $n \in \mathbb{N}$); in general F^I_{fin} is a strict subset of F^I. However, if $f, g \in F^I_{\text{fin}}$ and $\lambda \in F$, then $I_{\lambda f} = I_f$ for $\lambda \neq 0$, $I_{\lambda f} = \emptyset$ for $\lambda = 0$, and $I_{f+g} \subseteq I_f \cup I_g$, showing F^I_{fin} is always a subspace of F^I. Define

$$e_i : I \longrightarrow F, \quad e_i(j) := \begin{cases} 1 & \text{if } j = i, \\ 0 & \text{if } j \neq i. \end{cases} \tag{A.19}$$

Then $I_{e_i} = \{i\}$ for each $i \in I$, in particular, $e_i \in F^I_{\text{fin}}$ for each $i \in I$. We claim that $B := \{e_i : i \in I\}$ is a basis for F^I_{fin}. Indeed, if $f \in F^I_{\text{fin}}$, then

$$f = \sum_{i \in I_f} f(i) e_i, \tag{A.20}$$

showing $\langle B \rangle = F^I_{\text{fin}}$. If J is a finite subset of I and $(\lambda_j)_{j \in J}$ is a family in F such that

$$\sum_{j \in J} \lambda_j e_j \equiv 0, \tag{A.21}$$

then

$$\underset{j \in J}{\forall} \quad \lambda_j \overset{\text{(A.19)}}{=} \sum_{j \in J} \lambda_j e_j(j) \overset{\text{(A.21)}}{=} 0, \tag{A.22}$$

proving B is linearly independent. Clearly, $\#B = \#I$, so we have shown

$$\dim F^I_{\text{fin}} = \#I. \tag{A.23}$$

In particular, we have shown that, for each $n \in \mathbb{N}$, the set $\{e_j : j = 1, \ldots, n\}$, where

$$e_1 := (1, 0, \ldots, 0), \quad e_2 := (0, 1, \ldots, 0), \quad \ldots, \quad e_n := (0, \ldots, 0, 1), \tag{A.24}$$

forms a basis of F^n (of $\mathbb{R}^n$ if $F = \mathbb{R}$ and of $\mathbb{C}^n$ if $F = \mathbb{C}$),

$$\dim F^n = \dim \mathbb{R}^n = \dim \mathbb{C}^n = n. \tag{A.25}$$

Remark A.16. We will see in Th. A.24 below that, in a certain sense, F^I_{fin} is the only vector space of dimension $\#I$ over F. In particular, for $n \in \mathbb{N}$, you can think of $\mathbb{K}^n$ as the *standard model* of an n-dimensional vector space over $\mathbb{K}$.

A.2 Linear Maps

Definition A.17. Let V and W be vector spaces over the field F.

(a) A map $A : V \longrightarrow W$ is called *F-linear* (or merely *linear* if the field F is understood) if, and only if,

$$\forall_{v_1,v_2\in V} \quad A(v_1 + v_2) = A(v_1) + A(v_2), \tag{A.26a}$$

$$\wedge \quad \forall_{\lambda\in F} \quad \forall_{v\in V} \quad A(\lambda v) = \lambda A(v) \tag{A.26b}$$

or, equivalently, if, and only if,

$$\forall_{\lambda,\mu\in F} \quad \forall_{v_1,v_2\in V} \quad A(\lambda v_1 + \mu v_2) = \lambda A(v_1) + \mu A(v_2) \tag{A.27}$$

(note that, in general, vector addition and scalar multiplication will be different on the left-hand sides and right-hand sides of the above equations).

One also calls linear maps (vector space) *homomorphisms*. We denote the set of all F-linear maps from V into W by $\mathcal{L}(V, W)$.

(b) A linear map $I : V \longrightarrow W$ is called a (vector space or linear) *isomorphism* if, and only if, it is bijective (i.e. invertible). The vector spaces V and W are called *isomorphic* (denoted $V \cong W$) if, and only if, there exists a vector space isomorphism $I : V \longrightarrow W$.

Theorem A.18. *Let V and W be vector spaces over the field F. If $I : V \longrightarrow W$ is a linear isomorphism, then so is $I^{-1} : W \longrightarrow V$ (i.e. I^{-1} is not only bijective, but also linear).*

Proof. See, e.g., [Str08, Th. 13.5]. ■

Definition A.19. Let V and W be vector spaces over the field F, and $A \in \mathcal{L}(V, W)$. Define the *kernel* and the *image* of A by

$$\ker A := A^{-1}\{0\} = \{v \in V : A(v) = 0\}, \tag{A.28a}$$

$$\operatorname{Im} A := A(V) = \{A(v) : v \in V\}, \tag{A.28b}$$

respectively.

Theorem A.20. *Let V and W be vector spaces over the field F, and $A \in \mathcal{L}(V, W)$.*

(a) $\ker A$ *is a subspace of V and* $\operatorname{Im} A$ *is a subspace of W.*

(b) *A is injective if, and only if,* $\ker A = \{0\}$.

Proof. (a): See, e.g., [Str08, Lem. 12.5].

(b): Since $A(0) = 0$, A being injective implies $\ker A = \{0\}$. Conversely, assume $\ker A = \{0\}$ and $A(v_1) = A(v_2)$ for $v_1, v_2 \in V$. Then $A(-v_1 + v_2) = -A(v_1) + A(v_2) = -A(v_1) + A(v_1) = 0$, i.e. $-v_1 + v_2 \in \ker A$, i.e. $-v_1 + v_2 = 0$, showing $v_1 = v_2$ and the injectivity of A. ■

Theorem A.21 (Dimension Formulas)**.** *Let V and W be vector spaces over the field F, and let $A : V \longrightarrow W$ be linear.*

(a) *If V is finite dimensional, then* $\dim V = \dim \ker A + \dim \operatorname{Im} A$.

(b) *If V is finite dimensional, then* $\dim \operatorname{Im} A \leq \dim V$.

(c) *If W is finite dimensional, then* $\dim \operatorname{Im} A \leq \dim W$.

Proof. See, e.g., [Str08, Th. 12.12]. ■

Proposition A.22. *Let V and W be vector spaces over the field F, and let $A : V \longrightarrow W$ be linear.*

(a) *A is injective if, and only if, for each linearly independent subset S of V, $A(S)$ is a linearly independent subset of W.*

(b) *A is surjective if, and only if, for each generating subset S of V, $A(S)$ is a generating subset of W.*

(c) *A is bijective if, and only if, for each basis B of V, $A(B)$ is a basis of W.*

Proof. (a): If A is not injective, then, according to Th. A.20(b), there exists $0 \neq v \in V$ such that $A(v) = 0$. Then $S := \{v\}$ is linearly independent, whereas $A(S) = \{0\}$ is not. Conversely, if A is injective, $S \subseteq V$ is linearly independent, and $\lambda_1, \ldots, \lambda_n \in F$; $s_1, \ldots, s_n \in S$; $n \in \mathbb{N}$; such that

$$0 = \sum_{i=1}^{n} \lambda_i A(s_i) = A\left(\sum_{i=1}^{n} \lambda_i s_i\right), \tag{A.29}$$

then $\sum_{i=1}^{n} \lambda_i s_i = 0$ by Th. A.20(b), implying $\lambda_1 = \cdots = \lambda_n = 0$, showing that $A(S)$ is also linearly independent.

(b): If A is not surjective, then $\langle A(V)\rangle = \operatorname{Im} A \neq W$, since $\operatorname{Im} A$ is a subspace of W. Conversely, if A is surjective, $S \subseteq V$, $\langle S\rangle = V$, and $w \in W$, then there are $v \in V$; $\lambda_1, \ldots, \lambda_n \in F$; $s_1, \ldots, s_n \in S$; $n \in \mathbb{N}$; such that $A(v) = w$ and $v = \sum_{i=1}^{n} \lambda_i s_i$, i.e. $w = A(v) = \sum_{i=1}^{n} \lambda_i A(s_i)$, proving $w \in \langle A(S)\rangle$. Since $w \in W$ was arbitrary, we have shown $\langle A(S)\rangle = W$.

(c) follows immediately by combining (a) and (b) (recalling that a basis is a linearly independent generating set). ■

Theorem A.23. *Let V and W be vector spaces over the field F. Then each linear map $A : V \longrightarrow W$ is uniquely determined by its values on a basis of V. More precisely, if B is a basis of V, $(w_b)_{b\in B}$ is a family in W, and, for each $v \in V$, B_v and $c_v : B_v \longrightarrow F \setminus \{0\}$ are as in Th.* A.11 *(we now write c_v instead of c to underline the dependence of c on v), then the map*

$$A : V \longrightarrow W, \quad A(v) = A\left(\sum_{b\in B_v} c_v(b)\, b\right) := \sum_{b\in B_v} c_v(b)\, w_b, \tag{A.30}$$

is linear, and $\tilde{A} \in \mathcal{L}(V,W)$ *with*

$$\underset{b\in B}{\forall} \quad \tilde{A}(b) = w_b, \tag{A.31}$$

implies $A = \tilde{A}$.

Proof. We first verify A is linear. Let $v \in V$ and $\lambda \in F$. If $\lambda = 0$, then $A(\lambda v) = A(0) = 0 = \lambda A(v)$. If $\lambda \neq 0$, then $B_{\lambda v} = B_v$, $c_{\lambda v} = \lambda c_v$, and

$$A(\lambda v) = A\left(\sum_{b\in B_{\lambda v}} c_{\lambda v}(b)\, b\right) = \sum_{b\in B_v} \lambda\, c_v(b)\, w_b = \lambda\, A\left(\sum_{b\in B_v} c_v(b)\, b\right) = \lambda A(v). \tag{A.32a}$$

Now let $u, v \in V$. If $u = 0$, then $A(u+v) = A(v) = 0 + A(v) = A(u) + A(v)$, and analogously if $v = 0$. So assume $u, v \neq 0$. If $u + v = 0$, then $v = -u$ and $A(u+v) = A(0) = 0 = A(u) + A(-u) = A(u) + A(v)$. If $u + v \neq 0$, then $B_{u+v} \subseteq B_u \cup B_v$ and c_{u+v} can be extended to $B_u \cup B_v$ by letting

$$c_{u+v}(b) := \begin{cases} c_u(b) + c_v(b) & \text{if } b \in B_u \cap B_v, \\ c_u(b) & \text{if } b \in B_u \setminus B_v, \\ c_v(b) & \text{if } b \in B_v \setminus B_u. \end{cases} \tag{A.32b}$$

One then obtains

$$\begin{aligned} A(u+v) &= A\left(\sum_{b\in B_{u+v}} c_{u+v}(b)\, b\right) = \sum_{b\in B_{u+v}} c_{u+v}(b)\, w_b \\ &= \sum_{b\in B_u} c_u(b)\, w_b + \sum_{b\in B_v} c_v(b)\, w_b = A(u) + A(v). \end{aligned} \tag{A.32c}$$

If $v \in V$ and B_v and c_v are as before, then the linearity of $\tilde{A}$ and (A.31) imply

$$\begin{aligned} \tilde{A}(v) &= \tilde{A}\left(\sum_{b\in B_v} c_v(b)\, b\right) \overset{\tilde{A}\in\mathcal{L}(V,W)}{=} \sum_{b\in B_v} c_v(b)\, \tilde{A}(b) \\ &= \sum_{b\in B_v} c_v(b)\, A(b) \overset{(*)}{=} \sum_{b\in B_v} c_v(b)\, w_b = A(v), \end{aligned} \tag{A.33}$$

where, at $(*)$, it was used that, for each $b \in B$, one has $A(b) = w_b$ (note $B_b = \{b\}$, $c_b(b) = 1$). Since (A.33) establishes $\tilde{A} = A$, the proof is complete. ■

Theorem A.24. *Let* V *and* W *be vector spaces over the field* F*. Then* $V \cong W$ *(i.e.* V *and* W *are isomorphic) if, and only if,* $\dim V = \dim W$.

Proof. Suppose $\dim V = \dim W$. If B_V is a basis of V and B_W is a basis of W, then there exists a bijective map $i : B_V \longrightarrow B_W$. According to Th. A.23, i defines a unique

linear map $A : V \longrightarrow W$ with $A(b) = i(b)$ for each $b \in B_V$. More precisely, letting, once again, for each $v \in V$, B_v and $c_v : B_v \longrightarrow F \setminus \{0\}$ be as in Th. A.11 (writing c_v instead of c to underline the dependence of c on v),

$$\underset{v\in V}{\forall} \quad A(v) = A\left(\sum_{b\in B_v} c_v(b)\, b\right) = \sum_{b\in B_v} c_v(b)\, i(b). \tag{A.34}$$

It remains to show A is bijective. If $v \neq 0$, then $B_v \neq \emptyset$ and $A(v) = \sum_{b\in B_v} c_v(b)\, i(b) \neq 0$, since $c_v(b) \neq 0$ and $\{i(b) : b \in B_v\} \subseteq B_W$ is linearly independent, showing A is injective by Th. A.20(b). If $w \in W$, then there exists a finite set $\tilde{B}_w \subseteq B_W$ and $c_w : B_w \longrightarrow F$ such that $\sum_{\tilde{b}\in\tilde{B}_w} \tilde{c}_w(\tilde{b})\, \tilde{b}$. Then

$$\begin{aligned} A\left(\sum_{\tilde{b}\in\tilde{B}_w} \tilde{c}_w(\tilde{b})\, i^{-1}(\tilde{b})\right) &\overset{\tilde{A}\in\mathcal{L}(V,W)}{=} \sum_{\tilde{b}\in\tilde{B}_w} \tilde{c}_w(\tilde{b})\, A\left(i^{-1}(\tilde{b})\right) \overset{i^{-1}(\tilde{b})\in B_V}{=} \sum_{\tilde{b}\in\tilde{B}_w} \tilde{c}_w(\tilde{b})\, i\left(i^{-1}(\tilde{b})\right) \\ &= \sum_{\tilde{b}\in\tilde{B}_w} \tilde{c}_w(\tilde{b})\, \tilde{b} = w, \end{aligned} \tag{A.35}$$

showing $\operatorname{Im} A = W$, completing the proof that A is bijective.

If $A : V \longrightarrow W$ is a linear isomorphism and B is a basis for V, then, by Prop. A.22(c), $A(B)$ is a basis for W. As A is bijective, so is $A\restriction_B$, showing $\dim V = \#B = \#A(B) = \dim W$ as claimed. ■

Definition A.25. Let V and W be vector spaces over the field F. We define an addition and a scalar multiplication on $\mathcal{L}(V, W)$ by

$$(A + B) : V \longrightarrow W, \qquad (A + B)(x) := A(x) + B(x), \tag{A.36a}$$

$$(\lambda \cdot A) : V \longrightarrow W, \qquad (\lambda \cdot A)(x) := \lambda \cdot A(x) \quad \text{for each } \lambda \in F. \tag{A.36b}$$

Theorem A.26. *Let V and W be vector spaces over the field F. The addition and scalar multiplication on $\mathcal{L}(V, W)$ given by* (A.36) *are well-defined in the sense that, if $A, B \in \mathcal{L}(V, W)$ and $\lambda \in F$, then $A + B \in \mathcal{L}(V, W)$ and $\lambda A \in \mathcal{L}(V, W)$. Moreover, with the operations defined in* (A.36)*, $\mathcal{L}(V, W)$ forms a vector space over F.*

Proof. See, e.g., [Str08, Th. 13.2]. ■

Theorem A.27. *Let V and W be finite dimensional vector spaces over the field F, let $\{v_1, \ldots, v_n\}$ and $\{w_1, \ldots, w_m\}$ be bases of V and W, respectively; $m, n \in \mathbb{N}$. Using Th.* A.23*, define maps $A_{ji} \in \mathcal{L}(V, W)$ by letting*

$$\underset{(j,i,k)\in\{1,\ldots,m\}\times\{1,\ldots,n\}^2}{\forall} \quad A_{ji}(v_k) := \begin{cases} w_j & \text{for } k = i, \\ 0 & \text{for } k \neq i. \end{cases} \tag{A.37}$$

Then $\big\{A_{ji} : (j, i) \in \{1, \ldots, m\} \times \{1, \ldots, n\}\big\}$ constitutes a basis for $\mathcal{L}(V, W)$ and, in particular,

$$\dim \mathcal{L}(V, W) = \dim V \cdot \dim W = n \cdot m. \tag{A.38}$$

Proof. See, e.g., [Str08, Th. 13.11]. ■

Theorem A.28. *Let V, W, X be vector spaces over the field F.*

(a) *The composition of linear maps is linear, i.e. if $A \in \mathcal{L}(V,W)$ and $B \in \mathcal{L}(W,X)$, then $B \circ A \in \mathcal{L}(V,X)$.*

(b) *If $A \in \mathcal{L}(V,W)$ and $B, C \in \mathcal{L}(W,X)$, then*

$$A \circ (B + C) = A \circ B + A \circ C. \tag{A.39}$$

(c) *If $A, B \in \mathcal{L}(V,W)$ and $C \in \mathcal{L}(W,X)$, then*

$$(A + B) \circ C = A \circ C + B \circ C. \tag{A.40}$$

Proof. See, e.g., [Str08, Th. 13.3]. ■

A.3 Matrices

Matrices provide a convenient representation for linear maps A between finite dimensional vector spaces V and W. Recall the basis $\{A_{ji} : (j,i) \in \{1,\dots,m\} \times \{1,\dots,n\}\}$ of $\mathcal{L}(V,W)$ that, in Th. A.27, was shown to arise from bases $\{v_1,\dots,v_n\}$ and $\{w_1,\dots,w_m\}$ of V and W, respectively; $m, n \in \mathbb{N}$. Thus, each $A \in \mathcal{L}(V,W)$ can be written in the form

$$A = \sum_{i=1}^{n} \sum_{j=1}^{m} a_{ji} A_{ji}, \tag{A.41}$$

with coordinates $(a_{ji})_{(j,i)\in\{1,\dots,m\}\times\{1,\dots,n\}}$ in F. This motivates the following definition of matrices.

Definition A.29. Let F be a field and $m, n \in \mathbb{N}$. A family $(a_{ji})_{(j,i)\in\{1,\dots,m\}\times\{1,\dots,n\}}$ is called an m-by-n or an $m \times n$ *matrix* over F, where $m \times n$ is called the *size*, *dimension* or *type* of the matrix. The a_{ji} are called the *entries* or *elements* of the matrix. One also writes just (a_{ji}) instead of $(a_{ji})_{(j,i)\in\{1,\dots,m\}\times\{1,\dots,n\}}$ if the size of the matrix is understood. One usually thinks of the $m \times n$ matrix (a_{ji}) as the *rectangular array*

$$(a_{ji}) = \begin{pmatrix} a_{11} & \dots & a_{1n} \\ \vdots & \vdots & \vdots \\ a_{m1} & \dots & a_{mn} \end{pmatrix} \tag{A.42}$$

with m rows and n columns. One therefore also calls $1 \times n$ matrices *row vectors* and $m \times 1$ matrices *column vectors*, and one calls $n \times n$ matrices *quadratic*. The set of all $m \times n$ matrices over F is denoted by $\mathcal{M}(m,n,F)$, and for the set of all quadratic $n \times n$ matrices, one uses the abbreviation $\mathcal{M}(n,F) := \mathcal{M}(n,n,F)$.

Definition A.30 (Matrix Arithmetic)**.** Let F be a field and $m, n, l \in \mathbb{N}$.

(a) *Matrix Addition:* For $m \times n$ matrices (a_{ji}) and (b_{ji}) over F, define the sum

$$(a_{ji}) + (b_{ji}) := (a_{ji} + b_{ji}). \tag{A.43}$$

(b) *Scalar Multiplication:* For each $m \times n$ matrix (a_{ji}) and each $\lambda \in F$, define

$$\lambda\,(a_{ji}) := (\lambda\, a_{ji}). \tag{A.44}$$

(c) *Matrix Multiplication:* For each $m \times n$ matrix (a_{ji}) and each $n \times l$ matrix (b_{ji}) over F, define the product

$$(a_{ji})(b_{ji}) := \left(\sum_{k=1}^{n} a_{jk} b_{ki} \right)_{(j,i) \in \{1,\ldots,m\} \times \{1,\ldots,l\}}, \tag{A.45}$$

i.e. the product of an $m \times n$ matrix and an $n \times l$ matrix is an $m \times l$ matrix (cf. Th. A.34 below).

Remark A.31. We consider matrices over a field F.

(a) For each $m, n \in \mathbb{N}$, the set $\mathcal{M}(m, n, F)$ of $m \times n$ matrices over F with the operations of Def. A.30(a),(b) constitutes a vector space over F: An $m \times n$ matrix $A = (a_{ji})_{(j,i) \in \{1,\ldots,m\} \times \{1,\ldots,n\}}$ is defined as a family in F, i.e., recalling [Phi15a, Def. 2.14(a)], A is defined as the function $A : \{1, \ldots, m\} \times \{1, \ldots, n\} \longrightarrow F$, $A(j, i) = a_{ji}$; and $\mathcal{M}(m, n, F) = \mathcal{F}\big(\{1, \ldots, m\} \times \{1, \ldots, n\}, F\big)$ (so we notice that matrices are nothing new in terms of objects, but just a new way of thinking about functions from $\{1, \ldots, m\} \times \{1, \ldots, n\}$ into F, that turns out to be convenient in certain contexts). Thus, the operations defined in Def. A.30(a),(b) are precisely the same operations that were defined in (A.5) and $\mathcal{M}(m, n, F)$ is a vector space according to Ex. A.2(c). Clearly, the map

$$\begin{gathered} I : \mathcal{M}(m, n, F) \longrightarrow F^{m \cdot n}, \quad (a_{ji}) \mapsto (\lambda_1, \ldots, \lambda_{m \cdot n}), \\ \text{where } \lambda_k = a_{ji} \text{ if, and only if, } k = (j-1) \cdot n + i, \end{gathered} \tag{A.46}$$

constitutes a linear isomorphism. Other important linear isomorphisms between $\mathcal{M}(m, n, F)$ and vector spaces of linear maps will be provided in Th. A.32 below.

(b) Matrix multiplication is associative whenever all relevant multiplications are defined. More precisely, if A is an $m \times n$ matrix, B is an $n \times l$, and C is an $l \times p$ matrix, then

$$(AB)C = A(BC) : \tag{A.47}$$

Indeed, one has $m \times p$ matrices $(AB)C = (d_{ji})$ and $A(BC) = (e_{ji})$, where

$$d_{ji} = \sum_{\alpha=1}^{l} \left(\sum_{k=1}^{n} a_{jk} b_{k\alpha} \right) c_{\alpha i} = \sum_{\alpha=1}^{l} \sum_{k=1}^{n} a_{jk} b_{k\alpha} c_{\alpha i} = \sum_{k=1}^{n} a_{jk} \left(\sum_{\alpha=1}^{l} b_{k\alpha} c_{\alpha i} \right) = e_{ji}. \tag{A.48}$$

(c) Matrix multiplication is, in general, *not* commutative: If A is an $m \times n$ matrix and B is an $n \times l$ with $m \neq l$, then BA is not even defined. If $m = l$, but $m \neq n$, then AB has dimension $m \times m$, but BA has different dimension, namely $n \times n$. And even if $m = n = l > 1$, then commutativity is, in general not true – for example

$$\begin{pmatrix} 1 & 1 & \dots & 1 \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 & \dots & 0 \\ 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & \dots & 0 \end{pmatrix} = \begin{pmatrix} \lambda & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & 0 \end{pmatrix}, \tag{A.49a}$$

$$\begin{pmatrix} 1 & 0 & \dots & 0 \\ 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & \dots & 0 \end{pmatrix} \begin{pmatrix} 1 & 1 & \dots & 1 \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & 0 \end{pmatrix} = \begin{pmatrix} 1 & 1 & \dots & 1 \\ 1 & 1 & \dots & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & \dots & 1 \end{pmatrix}. \tag{A.49b}$$

Note that $\lambda = m$ for $F = \mathbb{R}$, but, in general, λ will depend on F, e.g. for $F = \{0, 1\}$, one obtains $\lambda = m \mod 2$.

—

Let us come back to the situation discussed at the beginning of the section above, resulting in (A.41). Let $v = \sum_{i=1}^{n} \lambda_i v_i \in V$ with $\lambda_1, \dots, \lambda_n \in F$. Then

$$\begin{aligned} A(v) &= \sum_{i=1}^{n} \lambda_i A(v_i) = \sum_{i=1}^{n} \lambda_i \sum_{k=1}^{n} \sum_{j=1}^{m} a_{jk} A_{jk}(v_i) \overset{(A.37)}{=} \sum_{i=1}^{n} \lambda_i \sum_{j=1}^{m} a_{ji} w_j \\ &= \sum_{j=1}^{m} \left(\sum_{i=1}^{n} a_{ji} \lambda_i \right) w_j. \end{aligned} \tag{A.50}$$

Thus, if we represent v by a column vector $\tilde{v}$ (an $n \times 1$ matrix) containing its coordinates $\lambda_1, \dots, \lambda_n$ with respect to the basis $\{v_1, \dots, v_n\}$ and $A(v)$ by a column vector $\tilde{w}$ (an $m \times 1$ matrix) containing its coordinates with respect to the basis $\{w_1, \dots, w_m\}$, then (A.50) shows

$$\tilde{w} = M\tilde{v} = M \begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_n \end{pmatrix}, \quad \text{where} \quad M := (a_{ji}). \tag{A.51}$$

For finite dimensional vector spaces, the precise relationship between linear maps, bases, and matrices is provided by the following theorem:

Theorem A.32. *Let V and W be finite dimensional vector spaces over the field F, let $\{v_1, \dots, v_n\}$ and $\{w_1, \dots, w_m\}$ be bases of V and W, respectively; $m, n \in \mathbb{N}$. Then the map*

$$I : \mathcal{L}(V, W) \longrightarrow \mathcal{M}(m, n, F), \quad A \mapsto (a_{ji}), \tag{A.52}$$

where the a_{ji} are given by (A.41) *constitutes a linear isomorphism.*

Proof. According to Th. A.27, $\big\{A_{ji} : (j,i) \in \{1,\dots,m\} \times \{1,\dots,n\}\big\}$ forms a basis of $\mathcal{L}(V,W)$. Thus, to every family of coordinates $\big\{a_{ji} : (j,i) \in \{1,\dots,m\} \times \{1,\dots,n\}\big\}$ in F, (A.41) defines a unique element of $\mathcal{L}(V,W)$, i.e. I is bijective. It remains to verify that I is linear. To this end, let $\lambda, \mu \in F$ and $A, B \in \mathcal{M}(m,n,F)$ with

$$A = \sum_{i=1}^{n} \sum_{j=1}^{m} a_{ji} A_{ji}, \qquad (a_{ji}) = I(A) \in \mathcal{M}(m,n,F), \tag{A.53a}$$

$$B = \sum_{i=1}^{n} \sum_{j=1}^{m} b_{ji} A_{ji}, \qquad (b_{ji}) = I(B) \in \mathcal{M}(m,n,F). \tag{A.53b}$$

Then

$$\lambda A + \mu B = \lambda \sum_{i=1}^{n} \sum_{j=1}^{m} a_{ji} A_{ji} + \mu \sum_{i=1}^{n} \sum_{j=1}^{m} b_{ji} A_{ji} = \sum_{i=1}^{n} \sum_{j=1}^{m} (\lambda a_{ji} + \mu b_{ji}) A_{ji}, \tag{A.54}$$

showing

$$I(\lambda A + \mu B) = (\lambda a_{ji} + \mu b_{ji}) A_{ji}) = \lambda (a_{ji}) + \mu (b_{ji}) = \lambda I(A) + \mu I(B), \tag{A.55}$$

proving the linearity of I. ■

Definition and Remark A.33. In the situation of Th. A.32, for each $A \in \mathcal{L}(V,W)$, one calls the matrix $I(A) = (a_{ji}) \in \mathcal{M}(m,n,F)$ the *(transformation) matrix corresponding* to A *with respect to* the basis $\{v_1,\dots,v_n\}$ of V and the basis $\{w_1,\dots,w_m\}$ of W. If the bases are understood, then one often tends to identify the map with its corresponding matrix.

However, as $I(A)$ depends on the bases, identifying A and $I(A)$ is only admissible as long as one keeps the bases of V and W fixed! Moreover, if one represents matrices as rectangular arrays as in (A.42) (which one usually does), then one actually considers the basis vectors of $\{v_1,\dots,v_n\}$ and $\{w_1,\dots,w_m\}$ as *ordered* from 1 to n (resp. m), i.e. $I(A)$ actually depends on the so-called *ordered bases* $(v_1,\dots,v_n)$ and $(w_1,\dots,w_m)$ (ordered bases are tuples rather than sets and the matrix corresponding to A changes if the order of the basis vectors changes).

Similarly, we had seen in (A.51) that it can be useful to identify a vector $v = \sum_{i=1}^{n} \lambda_i v_i$ with its coordinates $(\lambda_1,\dots,\lambda_n)$, typically represented as an $n \times 1$ matrix (a column vector, as in (A.51)) or a $1 \times n$ matrix (a row vector). Obviously, this identification is also only admissible as long as the basis $\{v_1,\dots,v_n\}$ and its order is kept fixed.

—

The following Th. A.34 is the justification for defining matrix multiplication according to Def. A.30(c).

Theorem A.34. *Let F be a field, let $n,m,l \in \mathbb{N}$, and let V,W,X be finite dimensional vector spaces over F such that V has basis $\{v_1,\dots,v_n\}$, W has basis $\{w_1,\dots,w_m\}$, and*

X has basis $\{x_1, \ldots, x_l\}$. If $A \in \mathcal{L}(V,W)$, $B \in \mathcal{L}(W,X)$, $M = (a_{ji}) \in \mathcal{M}(m,n,F)$ is the matrix corresponding to A with respect to $\{v_1, \ldots, v_n\}$ and $\{w_1, \ldots, w_m\}$, and $N = (b_{ji}) \in \mathcal{M}(l,m,F)$ is the matrix corresponding to B with respect to $\{w_1, \ldots, w_m\}$ and $\{x_1, \ldots, x_l\}$, then $NM = (\sum_{k=1}^m b_{jk}a_{ki}) \in \mathcal{M}(l,n,F)$ is the matrix corresponding to BA with respect to $\{v_1, \ldots, v_n\}$ and $\{x_1, \ldots, x_l\}$.

Proof. For each $i \in \{1, \ldots, n\}$, one computes

$$\begin{aligned}(BA)(v_i) = B\big(A(v_i)\big) = B\left(\sum_{k=1}^m a_{ki}w_k\right) = \sum_{k=1}^m a_{ki}B(w_k) = \sum_{k=1}^m a_{ki}\sum_{j=1}^l b_{jk}x_j \\ = \sum_{j=1}^l\sum_{k=1}^m b_{jk}a_{ki}x_j = \sum_{j=1}^l\left(\sum_{k=1}^m b_{jk}a_{ki}\right)x_j, \qquad (A.56)\end{aligned}$$

proving $NM = (\sum_{k=1}^m b_{jk}a_{ki})$ is the matrix corresponding to BA with respect to the bases $\{v_1, \ldots, v_n\}$ and $\{x_1, \ldots, x_l\}$. ∎

Definition and Remark A.35. Let F be a field, $A := (a_{ji})_{(j,i)\in\{1,\ldots,m\}\times\{1,\ldots,n\}} \in \mathcal{M}(m,n,F)$, and $m, n \in \mathbb{N}$. Then we define the *transpose* of A, denoted A^{t}, by

$$A^{\mathrm{t}} := (a_{ji})_{(i,j)\in\{1,\ldots,n\}\times\{1,\ldots,m\}}. \qquad (A.57)$$

Thus, if A is an $m \times n$ matrix, then its transpose is an $n \times m$ matrix, where one obtains A^{t} from A by switching rows and columns. One has to use care when using the notation of (A.57), as one often implicitly assumes that, when writing (a_{ji}), the first index is for rows and the second index for columns. However, this is actually determined by the order of the factors of the cartesian product that determines the domain of the family. Whereas A is the map $f : \{1, \ldots, m\} \times \{1, \ldots, n\} \longrightarrow F$, $f(j,i) = a_{ji}$, its transpose A^{t} is the map $f^{\mathrm{t}} : \{1, \ldots, n\} \times \{1, \ldots, m\} \longrightarrow F$, $f^{\mathrm{t}}(i,j) = f(j,i) = a_{ji}$. To emphasize this in the notation, one can rewrite (A.57) in the form

$$A^{\mathrm{t}} = (b_{ij})_{(i,j)\in\{1,\ldots,n\}\times\{1,\ldots,m\}}, \quad \text{where} \quad \forall_{(i,j)\in\{1,\ldots,n\}\times\{1,\ldots,m\}} \; b_{ij} := a_{ji}. \qquad (A.58)$$

For the transpose of A, one also finds the notation A' instead of A^{t}.

Theorem A.36. *Let F be a field and $m, n, l \in \mathbb{N}$.*

(a) *The map*

$$I : \mathcal{M}(m,n,F) \longrightarrow \mathcal{M}(n,m,F), \quad A \mapsto A^{\mathrm{t}}, \qquad (A.59)$$

is a linear isomorphism and

$$\forall_{A\in\mathcal{M}(m,n,F)} \; (A^{\mathrm{t}})^{\mathrm{t}} = A. \qquad (A.60)$$

(b) *If $A \in \mathcal{M}(m,n,F)$ and $B \in \mathcal{M}(n,l,F)$, then*

$$(AB)^{\mathrm{t}} = B^{\mathrm{t}}A^{\mathrm{t}}. \qquad (A.61)$$

Proof. (a): It is immediate from (A.57) that (A.60) is valid, showing I has an inverse map and is, hence, bijective. So it just remains to verify I is linear. However, if $A, B \in \mathcal{M}(m,n,F)$, $A = (a_{ji})$, $B = (b_{ji})$, and $\mu, \lambda \in F$, then

$$\begin{aligned}(\lambda A + \mu B)^{\mathrm{t}} &= (\lambda a_{ji} + \mu b_{ji})^{\mathrm{t}}_{(j,i)\in\{1,\ldots,m\}\times\{1,\ldots,n\}} = (\lambda a_{ji} + \mu b_{ji})_{(i,j)\in\{1,\ldots,n\}\times\{1,\ldots,m\}} \\ &= \lambda (a_{ji})_{(i,j)\in\{1,\ldots,n\}\times\{1,\ldots,m\}} + \mu (b_{ji})_{(i,j)\in\{1,\ldots,n\}\times\{1,\ldots,m\}} \\ &= \lambda (a_{ji})^{\mathrm{t}}_{(j,i)\in\{1,\ldots,m\}\times\{1,\ldots,n\}} + \mu (b_{ji})^{\mathrm{t}}_{(j,i)\in\{1,\ldots,m\}\times\{1,\ldots,n\}} \\ &= \lambda A^{\mathrm{t}} + \mu B^{\mathrm{t}}, \end{aligned} \tag{A.62}$$

thereby establishing the case.

(b): Let $A = (a_{ji})$, $B = (b_{ji})$, $A^{\mathrm{t}} = (a^{\mathrm{t}}_{ji})$, $B^{\mathrm{t}} = (b^{\mathrm{t}}_{ji})$. Then

$$\begin{aligned}(AB)^{\mathrm{t}} &\overset{(A.45)}{=} \left(\sum_{k=1}^{n} a_{jk} b_{ki}\right)^{\mathrm{t}}_{(j,i)\in\{1,\ldots,m\}\times\{1,\ldots,l\}} = \left(\sum_{k=1}^{n} b_{ki} a_{jk}\right)_{(i,j)\in\{1,\ldots,l\}\times\{1,\ldots,m\}} \\ &= \left(\sum_{k=1}^{n} b_{kj} a_{ik}\right)_{(j,i)\in\{1,\ldots,l\}\times\{1,\ldots,m\}} = \left(\sum_{k=1}^{n} b^{\mathrm{t}}_{jk} a^{\mathrm{t}}_{ki}\right)_{(j,i)\in\{1,\ldots,l\}\times\{1,\ldots,m\}} \\ &\overset{(A.45)}{=} B^{\mathrm{t}} A^{\mathrm{t}}, \end{aligned} \tag{A.63}$$

proving (A.61). ■

A.4 Determinants

For each quadratic matrix $A \in \mathcal{M}(n,F)$, one can define its determinant $\det(A) \in F$, resulting in a function $\det : \mathcal{M}(n,F) \longrightarrow F$ that is often useful when studying matrices and linear maps. One can characterize the determinant function axiomatically (see Def. A.38 below), and, with some preparation, one can also provide an explicit formula (see (A.82) below).

One important feature of the determinant is its being nonzero if, and only if, the matrix is invertible (cf. Def. A.46 and Th. A.48(a) below). Another is the fact that the determinant's value only depends on the linear map defined by A and an arbitrary basis of F^n. This allows to define $\det : \mathcal{L}(V,V) \longrightarrow F$ as in Def. and Rem. A.53 below. One can show that, if $A \in \mathcal{L}(V,V) \longrightarrow F$, then $\det(A)$ is a measure of the n-dimensional volume distortion caused by applying A, but, here, we will not pursue this aspect.

Definition A.37. Let F be a field, $n \in \mathbb{N}$. Then the $n \times n$ matrix

$$\mathrm{Id} := \begin{pmatrix} 1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 1 \end{pmatrix} = (e_{ji}), \quad \text{where} \quad e_{ji} := \begin{cases} 1 & \text{for } j = i, \\ 0 & \text{for } j \neq i, \end{cases} \tag{A.64}$$

is called *identity matrix* or just *identity* or *unit matrix*. The dependence of Id on n is suppressed in the notation, but n should always be clear from the context. In the literature, one also finds the notation E or I instead of Id.

Definition A.38. Let F be a field, $n \in \mathbb{N}$. A map $\det : \mathcal{M}(n,F) \longrightarrow F$ is called *determinant* if, and only if, it satisfies the following conditions (i) – (iii):

(i) det is *multilinear* with regard to matrix columns, i.e., for each $A \in \mathcal{M}(n,F)$, $b \in \mathcal{M}(n,1,F)$, $i \in \{1,\ldots,n\}$, and $\lambda, \mu \in F$:

$$\begin{aligned} &\det(a_1,\ldots,\lambda\, a_i + \mu\, b,\ldots,a_n) \\ &= \lambda\, \det(A) + \mu\, \det(a_1,\ldots,a_{i-1},b,a_{i+1}\ldots,a_n), \end{aligned} \tag{A.65}$$

where $a_1,\ldots,a_n$ denote the columns of A.

(ii) If the columns $a_1,\ldots,a_n$ of $A = (a_1,\ldots,a_n) \in \mathcal{M}(n,F)$ are linearly dependent, then $\det(A) = 0$.

(iii) $\det(\mathrm{Id}) = 1$.

Notation A.39. If F is a field, $n \in \mathbb{N}$, and $\det : \mathcal{M}(n,F) \longrightarrow F$ is a determinant, then, for $A = (a_{ji}) \in \mathcal{M}(n,F)$, one commonly uses the notation

$$|A| := \begin{vmatrix} a_{11} & \ldots & a_{1n} \\ \vdots & \vdots & \vdots \\ a_{m1} & \ldots & a_{mn} \end{vmatrix} := \det(A). \tag{A.66}$$

—

In Th. A.45 below, it will be stated that, for each $n \in \mathbb{N}$, there exists a unique determinant. To also state an explicit formula for this determinant, we need to know a few things about permutations.

Definition and Remark A.40. Let $n \in \mathbb{N}$. Each bijective map $\pi : \{1,\ldots,n\} \longrightarrow \{1,\ldots,n\}$ is called a *permutation* of $\{1,\ldots,n\}$. The set of permutations of $\{1,\ldots,n\}$ forms a group with respect to the composition of maps, the so-called *symmetric group* S_n: Indeed, the composition of maps is associative by [Phi15a, Prop. 2.9(a)]; the neutral element is the identity map $e : \{1,\ldots,n\} \longrightarrow \{1,\ldots,n\}$, $e(i) = i$; and, for each $\sigma \in S_n$, its inverse map σ^{-1} is also its inverse element in the group S_n. Caveat: Simple examples show that S_n is *not* commutative.

Definition A.41. Let $k,n \in \mathbb{N}$, $k \leq n$. A permutation $\pi \in S_n$ is called a k-cycle if, and only if, there exist k distinct numbers $i_1,\ldots,i_k \in \{1,\ldots,n\}$ such that

$$\pi(i) = \begin{cases} i_{j+1} & \text{if } j \in \{1,\ldots,k-1\}, \\ i_1 & \text{if } i = i_k, \\ i & \text{if } i \notin \{i_1,\ldots,i_k\}. \end{cases} \tag{A.67}$$

If π is a cycle as in (A.67), then one writes

$$\pi = (i_1\ i_2\ \ldots\ i_k). \tag{A.68}$$

Each 2-cycle is also known as a *transposition*.

Theorem A.42. *Let $n \in \mathbb{N}$.*

(a) *Each permutation can be decomposed into finitely many disjoint cycles: For each $\pi \in S_n$ there exists a decomposition of $\{1, \dots, n\}$ into disjoint sets $A_1, \dots, A_N$, $N \in \mathbb{N}$, i.e.*

$$\{1, \dots, n\} = \bigcup_{i=1}^{N} A_i \quad \text{and} \quad A_i \cap A_j = \emptyset \quad \text{for } i \neq j, \tag{A.69}$$

such that A_i consists of the distinct elements $a_{i1}, \dots, a_{i,N_i}$ and

$$\pi = (a_{N1} \ \dots \ a_{N,N_N}) \cdots (a_{11} \ \dots \ a_{1,N_1}). \tag{A.70}$$

The decomposition (A.70) *is unique up to the order of the cycles.*

(b) *If $n \geq 2$, then every permutation $\pi \in S_n$ is the composition of finitely many transpositions, where each transposition permutes two juxtaposed elements, i.e.*

$$\forall_{\pi \in S_n} \quad \exists_{N \in \mathbb{N}} \quad \exists_{\tau_1, \dots, \tau_N \in T} \quad \pi = \tau_N \circ \cdots \circ \tau_1, \tag{A.71}$$

where $T := \big\{(i \ \ i+1) : i \in \{1, \dots, n-1\}\big\}$.

Proof. (a): We prove the statement by induction on n. For $n = 1$, there is nothing to prove. Let $n > 1$ and choose $i \in \{1, \dots, n\}$. We claim that

$$\exists_{k \in \mathbb{N}} \quad \left(\pi^k(i) = i \ \wedge \ \forall_{l \in \{1, \dots, k-1\}} \quad \pi^l(i) \neq i \right). \tag{A.72}$$

Indeed, since $\{1, \dots, n\}$ is finite, there must be a smallest $k \in \mathbb{N}$ such that $\pi^k(i) \in A_1 := \{i, \pi(i), \dots, \pi^{k-1}(i)\}$. Since π is bijective, it must be $\pi^k(i) = i$ and $(i \ \pi(i) \ \dots, \pi^{k-1}(i))$ is a k-cycle. We are already done in case $k = n$. If $k < n$, then consider $B := \{1, \dots, n\} \setminus A_1$. Then, again using the bijectivity of π, $\pi \restriction_B$ is a permutation on B with $1 \leq \# B < n$. By induction, there are disjoint sets $A_2, \dots, A_N$ such that $B = \bigcup_{j=2}^{N} A_j$, A_j consists of the distinct elements $a_{j1}, \dots, a_{j,N_j}$ and

$$\pi \restriction_B = (a_{N1} \ \dots \ a_{N,N_N}) \cdots (a_{21} \ \dots \ a_{2,N_2}).$$

Since $\pi = (i \ \pi(i) \ \dots, \pi^{k-1}(i)) \circ \pi \restriction_B$, this finishes the proof of (A.70). If there were another, different, decomposition of π into cycles, say, given by disjoint sets $B_1, \dots, B_M$, $\{1, \dots, n\} = \bigcup_{i=1}^{M} B_i$, $M \in \mathbb{N}$, then there were $A_i \neq B_j$ and $k \in A_i \cap B_j$. But then k were in the cycle given by A_i and in the cycle given by B_j, implying $A_i = \{\pi^l(k) : l \in \mathbb{N}\} = B_j$, in contradiction to $A_i \neq B_j$.

(b): We first show that every $\pi \in S_n$ is a composition of finitely many transpositions (not necessarily transpositions from the set T): According to (a), it suffices to show that every cycle is a composition of finitely many transpositions. Since each 1-cycle is the identity, it is $(i) = \mathrm{Id} = (1\ 2)\,(1\ 2)$ for each $i \in \{1, \dots, n\}$. If $(i_1 \ \dots \ i_k)$ is a k-cycle, $k \in \{2, \dots, n\}$, then

$$(i_1 \ \dots \ i_k) = (i_1 \ i_2)\,(i_2 \ i_3) \cdots (i_{k-1} \ i_k) : \tag{A.73}$$

Indeed,

$$\underset{i\in\{1,\dots,n\}}{\forall} \quad (i_1\ i_2)\,(i_2\ i_3)\cdots(i_{k-1}\ i_k)(i) = \begin{cases} i_1 & \text{for } i = i_k, \\ i_{l+1} & \text{for } i = i_l,\ l \in \{1,\dots,k-1\}, \\ i & \text{for } i \notin \{i_1,\dots,i_k\}, \end{cases} \tag{A.74}$$

proving (A.73). To finish the proof of (b), we observe that every transposition is a composition of finitely many elements of T: If $i, j \in \{1,\dots,n\}$, $i < j$, then

$$(i\ j) = (i\ \ i+1)\cdots(j-2\ j-1)(j-1\ \ j)\cdots(i+1\ \ i+2)(i\ \ i+1)\,: \tag{A.75}$$

Indeed,

$$\begin{aligned} &\underset{k\in\{1,\dots,n\}}{\forall} \quad (i\ \ i+1)\cdots(j-2\ j-1)(j-1\ \ j)\cdots(i+1\ \ i+2)(i\ \ i+1)(k) \\ &\qquad = \begin{cases} j & \text{for } k = i, \\ i & \text{for } k = j, \\ k & \text{for } i < k < j, \\ k & \text{for } k \notin \{i, i+1,\dots,j\}, \end{cases} \end{aligned} \tag{A.76}$$

proving (A.75). ■

Definition A.43. Let $n \in \mathbb{N}$. For each permutation $\pi \in S_n$, one defines its *sign*, $\operatorname{sgn}(\pi)$, via the map

$$\operatorname{sgn} :\ S_n \longrightarrow \{-1, 1\}, \quad \operatorname{sgn}(\pi) := \prod_{1\le i<j\le n} \frac{\pi(i) - \pi(j)}{i - j}. \tag{A.77}$$

Note that, for $n = 1$, $\operatorname{sgn} :\ S_1 = \{e\} \longrightarrow \{-1, 1\}$, $\operatorname{sgn}(e) = 1$, as the product in (A.77) is empty.

Proposition A.44. *Let $n \in \mathbb{N}$.*

(a) *The sign is well-defined by* (A.77)*, i.e. the map is, indeed, $\{-1, 1\}$-valued.*

(b) *The function* $\operatorname{sgn} :\ S_n \longrightarrow \{-1, 1\}$ *is a group homomorphism (note that $\{-1, 1\}$ forms a multiplicative subgroup of $\mathbb{R}$), i.e.*

$$\underset{\pi_1,\pi_2\in S_n}{\forall} \quad \operatorname{sgn}(\pi_1 \circ \pi_2) = \operatorname{sgn}(\pi_1)\operatorname{sgn}(\pi_2). \tag{A.78}$$

(c) *For $n \ge 2$, if a permutation $\pi \in S_n$ is the composition of k transpositions, then the parity of k is uniquely determined by π (i.e., for a given π, k is either always even or always odd) and*

$$\operatorname{sgn}(\pi) = (-1)^k = \begin{cases} 1 & \textit{if } k \textit{ is even}, \\ -1 & \textit{if } k \textit{ is odd}. \end{cases} \tag{A.79}$$

Proof. (a): The map sgn is $\{-1,1\}$-valued, since the bijectivity of $\pi \in S_n$ implies that the factor $i-j$ appears in the denominator of $\operatorname{sgn}(\pi)$ as defined in (A.77) if, and only if, the factor $i-j$ or the factor $j-i$ appears in the numerator.

(b): Let $\pi_1, \pi_2 \in S_n$. One computes

$$\begin{aligned} \operatorname{sgn}(\pi_1 \circ \pi_2) &= \prod_{1\le i<j\le n} \frac{\pi_1(\pi_2(i)) - \pi_1(\pi_2(j))}{i-j} \\ &= \prod_{1\le i<j\le n} \left(\frac{\pi_1(\pi_2(i)) - \pi_1(\pi_2(j))}{\pi_2(i)-\pi_2(j)} \cdot \frac{\pi_2(i)-\pi_2(j)}{i-j} \right) \\ &\overset{\pi_2 \text{ bij.}}{=} \operatorname{sgn}(\pi_1)\operatorname{sgn}(\pi_2). \end{aligned} \tag{A.80}$$

(c): If $\tau \in S_n$ is a transposition, then there are elements $i, j \in \{1,\dots,n\}$ such that $i<j$ and $\tau = (i\ j)$. Thus,

$$\operatorname{sgn}(\tau) = \frac{\tau(i)-\tau(j)}{i-j} = \frac{j-i}{i-j} = -1 \tag{A.81}$$

holds for every transposition τ. In consequence, if $\pi \in S_n$ is the composition of k transpositions, $k \in \mathbb{N}$, then (A.79) must hold and, in particular, k is always even if $\operatorname{sgn}(\pi) = 1$ and k is always odd if if $\operatorname{sgn}(\pi) = -1$. ■

Theorem A.45. *Let F be a field. For each $n \in \mathbb{N}$, there exists a unique determinant, i.e. there is a unique map* $\det : \mathcal{M}(n,F) \longrightarrow F$, *satisfying* (i) – (iii) *of Def.* A.38. *Moreover, this map is given by*

$$\det : \mathcal{M}(n,F) \longrightarrow F, \quad \det\big((a_{ji})\big) := \sum_{\pi\in S_n} \operatorname{sgn}(\pi)\, a_{1\pi(1)} \cdots a_{n\pi(n)}. \tag{A.82}$$

Proof. See, e.g., [Str08, Th. 17.5, Th. 17.11(a)]. ■

Definition A.46. Let F be a field, $n \in \mathbb{N}$. A quadratic matrix $A \in \mathcal{M}(n,F)$ is called *invertible* or *regular* if, and only if,

$$\exists_{B\in\mathcal{M}(n,F)} \quad AB = \mathrm{Id}\,. \tag{A.83}$$

One then usually writes A^{-1} instead of B and calls A^{-1} the *inverse matrix* of A. If A is not regular, then it is called *singular*.

Remark A.47. If V is a finite dimensional vector space over a field F, and $\{v_1,\dots,v_n\}$ is a basis of V, $n \in \mathbb{N}$, then, due to Th. A.32, a linear map $A \in \mathcal{L}(V,V)$ is bijective if, and only if, its transformation matrix $I(V)$ with respect to the given basis is invertible.

—

Important properties of the determinant are compiled in the following Th. A.48.

Theorem A.48. *Let F be a field, $n \in \mathbb{N}$, let $A \in \mathcal{M}(n, F)$, and let $c_1, \ldots, c_n$ denote the columns of A, whereas $r_1, \ldots, r_n$ denote the rows of A, i.e.*

$$A = (c_1, \ldots, c_n) = \begin{pmatrix} r_1 \\ \vdots \\ r_n \end{pmatrix}. \tag{A.84}$$

(a) $\det(A) = 0$ *if, and only if, A is singular. If A is invertible, then* $\det(A^{-1}) = (\det(A))^{-1}$.

(b) *If $B \in \mathcal{M}(n, F)$, then* $\det(AB) = \det(A)\det(B)$.

(c) $\det(A^{\mathrm{t}}) = \det(A)$.

(d) *If $\lambda \in F$, then* $\det(\lambda A) = \lambda^n \det(A)$.

(e) *The value of the determinant remains the same if one column of a matrix is replaced by the sum of that column and a scalar multiple of another column. More generally, the determinant remains the same if one column of a matrix is replaced by the sum of that column and a linear combination of the other columns, i.e., if $\lambda_1, \ldots, \lambda_n \in F$ and $i \in \{1, \ldots, n\}$, then*

$$\det(A) = \det(c_1, \ldots, c_n) = \det\left(c_1, \ldots, c_{i-1}, c_i + \sum_{\substack{j=1 \\ j \neq i}}^{n} \lambda_j \, c_j, c_{i+1}, \ldots, c_n\right). \tag{A.85}$$

(f) *Switching columns i and j, where $i, j \in \{1, \ldots, n\}$, $i \neq j$, changes the sign of the determinant, i.e.*

$$\det(c_1, \ldots, c_i, \ldots, c_j, \ldots, c_n) = -\det(c_1, \ldots, c_j, \ldots, c_i, \ldots, c_n). \tag{A.86}$$

(g) det *is* multilinear *with regard to matrix rows, i.e., for each $b \in \mathcal{M}(1, n, F)$, $i \in \{1, \ldots, n\}$, and $\lambda, \mu \in F$:*

$$\det\begin{pmatrix} r_1 \\ \vdots \\ r_{i-1} \\ \lambda \, r_i + \mu \, b \\ r_{i+1} \\ \vdots \\ r_n \end{pmatrix} = \lambda \, \det(A) + \mu \, \det\begin{pmatrix} r_1 \\ \vdots \\ r_{i-1} \\ b \\ r_{i+1} \\ \vdots \\ r_n \end{pmatrix}. \tag{A.87}$$

(h) *The value of the determinant remains the same if one row of a matrix is replaced by the sum of that row and a scalar multiple of another row. More generally, the determinant remains the same if one row of a matrix is replaced by the sum of*

that row and a linear combination of the other rows, i.e., if $\lambda_1, \dots, \lambda_n \in F$ *and* $i \in \{1, \dots, n\}$*, then*

$$\det(A) = \det \begin{pmatrix} r_1 \\ \vdots \\ r_n \end{pmatrix} = \det \begin{pmatrix} r_1 \\ \vdots \\ r_{i-1} \\ r_i + \sum_{\substack{j=1 \\ j \neq i}}^{n} \lambda_j \, r_j \\ r_{i+1} \\ \vdots \\ r_n \end{pmatrix} . \tag{A.88}$$

(i) *Switching rows* i *and* j*, where* $i, j \in \{1, \dots, n\}$*,* $i \neq j$*, changes the sign of the determinant, i.e.*

$$\det \begin{pmatrix} r_1 \\ \vdots \\ r_i \\ \vdots \\ r_j \\ \vdots \\ r_n \end{pmatrix} = -\det \begin{pmatrix} r_1 \\ \vdots \\ r_j \\ \vdots \\ r_i \\ \vdots \\ r_n \end{pmatrix} . \tag{A.89}$$

Proof. (a): See, e.g., [Str08, Th. 17.7(b), Th. 17.11(a)].

(b): See, e.g., [Str08, Th. 17.11(b)].

(c): See, e.g., [Str08, Lem. 18.1].

(d) is an immediate consequence of Def. A.38(i).

(e): One computes, for $i < j$,

$$\begin{aligned} &\det\Big(c_1, \dots, c_{i-1}, c_i + \lambda_j \, c_j, c_{i+1}, \dots, c_n\Big) \\ &\overset{\text{Def. A.38(i)}}{=} \lambda_j^{-1} \det\Big(c_1, \dots, c_{i-1}, c_i + \lambda_j \, c_j, c_{i+1}, \dots, \lambda_j \, c_j, \dots, c_n\Big) \\ &\overset{\text{Def. A.38(i)}}{=} \lambda_j^{-1} \det\Big(c_1, \dots, c_{i-1}, c_i, c_{i+1}, \dots, \lambda_j \, c_j, \dots, c_n\Big) \\ &\qquad + \lambda_j^{-1} \det\Big(c_1, \dots, c_{i-1}, \lambda_j \, c_j, c_{i+1}, \dots, \lambda_j \, c_j, \dots, c_n\Big) \\ &\overset{\text{Def. A.38(ii)}}{=} \lambda_j^{-1} \det\Big(c_1, \dots, c_{i-1}, c_i, c_{i+1}, \dots, \lambda_j \, c_j, \dots, c_n\Big) + 0 \\ &\overset{\text{Def. A.38(i)}}{=} \det\Big(c_1, \dots, c_n\Big) = \det(A). \end{aligned} \tag{A.90}$$

The general case of (A.85) then follows by induction.

(f): We compute

$$\begin{aligned}
&\det(c_1,\dots,c_i,\dots,c_j,\dots,c_n)+\det(c_1,\dots,c_j,\dots,c_i,\dots,c_n)\\
&\overset{\text{(e)}}{=}\quad \det(c_1,\dots,c_i+c_j,\dots,c_j,\dots,c_n)+\det(c_1,\dots,c_j+c_i,\dots,c_i,\dots,c_n)\\
&\overset{\text{Def. A.38(i)}}{=}\quad \det(c_1,\dots,c_i+c_j,\dots,c_i+c_j,\dots,c_n)\overset{\text{Def. A.38(ii)}}{=}0, \qquad \text{(A.91)}
\end{aligned}$$

proving (f).

(g) is inferred by combining Def. A.38(i) with (c).

(h) is inferred by combining (e) with (c).

(i) is inferred by combining (f) with (c). ■

Theorem A.49 (Block Matrices)**.** *The determinant of so-called block matrices, where one block is a zero matrix (all entries* 0*), can be computed as the product of the determinants of the corresponding blocks. More precisely, if* $n, m \in \mathbb{N}$*, then*

$$\begin{vmatrix} a_{11} & \dots & a_{1n} & & & \\ \vdots & \vdots & \vdots & & * & \\ a_{n1} & \dots & a_{nn} & & & \\ 0 & \dots & 0 & b_{11} & \dots & b_{1m} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & 0 & b_{m1} & \dots & b_{mm} \end{vmatrix} = \det(a_{ji})\det(b_{ji}). \qquad \text{(A.92)}$$

Proof. See, e.g., [Str08, Th. 18.3]. ■

Definition A.50. Let F be a field, $n \in \mathbb{N}$, $n \geq 2$, $A = (a_{ji}) \in \mathcal{M}(n, F)$. For each $j, i \in \{1, \dots, n\}$, let M_{ji} the $n \times n$ submatrix of A obtained by deleting the jth row and the ith column of A – the M_{ji} are sometimes called the *minor matrices* of A; define

$$A_{ji} := (-1)^{i+j} \det(M_{ij}), \qquad \text{(A.93)}$$

where the A_{ji} are called *cofactors* of A and the $\det(M_{ij})$ are called the *minors* of A. Let $\tilde{A} := (A_{ji})$ denote the matrix of cofactors.

Theorem A.51. *Let* F *be a field,* $n \in \mathbb{N}$*,* $n \geq 2$*,* $A = (a_{ji}) \in \mathcal{M}(n, F)$*. Moreover, let* $\tilde{A} := (A_{ji})$ *be the matrix of cofactors according to Def.* A.50*.*

(a) $A\tilde{A} = (\det A)\,\mathrm{Id}$*.*

(b) *If* $\det A \neq 0$*, then* $\det \tilde{A} = (\det A)^{n-1}$*.*

(c) *If* $\det A \neq 0$*, then* $A^{-1} = (\det A)^{-1}\,\tilde{A}$*.*

(d) Laplace Expansion by Rows: $\det A = \sum_{j=1}^{n} a_{ij} A_{ji}$ *(expansion with respect to the ith row).*

(e) Laplace Expansion by Columns: $\det A = \sum_{j=1}^{n} A_{ij} a_{ji}$ *(expansion with respect to the ith column).*

Proof. See, e.g., [Str08, Th. 18.6]. ■

Theorem A.52. *Let F be a field, $n \in \mathbb{N}$, let V be an n-dimensional vector space over F, and $A \in \mathcal{L}(V,V)$. Moreover, let $B_1 = \{v_1, \ldots, v_n\}$ and $B_2 = \{w_1, \ldots, w_n\}$ be bases of V. If $M = (m_{ji})$ is the transformation matrix corresponding to A with respect to B_1 and $N = (n_{ji})$ is the transformation matrix corresponding to A with respect to B_2 (i.e., for each $i \in \{1, \ldots, n\}$, $A(v_i) = \sum_{j=1}^{n} m_{ji} v_j$ and $A(w_i) = \sum_{j=1}^{n} n_{ji} w_j$, cf. Def. and Rem.* A.33*), then* $\det(M) = \det(N)$.

Proof. See, e.g., [Str08, Th. 17.11(a)]. ■

Definition and Remark A.53. Let F be a field, $n \in \mathbb{N}$, and let V be an n-dimensional vector space over F. Then Th. A.52 allows to define a determinant function for linear maps by

$$\det : \mathcal{L}(V,V) \longrightarrow F, \quad \det(A) := \det(M), \tag{A.94}$$

where M is a transformation matrix for A with respect to an arbitrary basis of V. Then Th. A.32 shows that Th. A.48(a),(b),(d) yield the following properties of the new determinant function defined in (A.94):

(a) If $A \in \mathcal{L}(V,V)$, then $\det(A) = 0$ if, and only if, A is not bijective. If A is bijective, then $\det(A^{-1}) = (\det(A))^{-1}$.

(b) If $A, B \in \mathcal{M}(n,F)$, then $\det(AB) = \det(A)\det(B)$.

(c) If $A \in \mathcal{L}(V,V)$ and $\lambda \in F$, then $\det(\lambda A) = \lambda^n \det(A)$. In particular, det is *not* linear for $n > 1$.

In [Str08, §17], the author actually defines the determinant first for linear maps $A \in \mathcal{L}(V,V)$, establishes properties including the above properties, and only then defines the determinant function for square matrices. Several alternative, but equivalent, approaches are possible and can be found in the literature.

B Metric Spaces

B.1 Metric Subspaces

Definition B.1. If (X,d) is a metric space, $M \subseteq X$, then (M,d) is called a *metric subspace* of (X,d) (if d is understood, one also speaks of M as a metric subspace of X). Thus, the metric on the subspace M is just the metric on X restricted to M.

Remark B.2. One sees immediately that a metric subspace (M,d) of a metric space (X,d) is, indeed, a metric space: Since d satisfies the laws (i) – (iii) from Def. 1.17 for all $x, y, z \in X$, in particular, d satisfies the same laws for all $x, y, z \in M \subseteq X$.

Definition B.3. Let (X, d) be a metric space, and let (M, d) be a metric subspace of (X, d). Is $A \subseteq M$ open with respect to (M, d), then one says that A is *open* in M or *M-open* or *relatively open*. For $A \subseteq M$ closed with respect to (M, d), one introduces analogous terms. Moreover, for $x \in M$, $r > 0$, call

$$B_{r,M}(x) := M \cap B_r(x) = \{y \in M : d(x, y) < r\}, \tag{B.1}$$

the *open M-ball* with radius r and center x.

Caveat B.4. One has to use care when working with a subspace (M, d) of a metric space (X, d): As will be seen in Ex. B.5, the notions and properties with respect to M are in general very different from the corresponding notions and properties with respect to X. For example, a set that is M-open might not be X-open and a set that is M-closed might not be X-closed!

Example B.5. (a) If (M, d) is a metric subspace of a metric space (X, d), then, according to Lem. 1.27(b), M is always both M-open and M-closed (irrespective of M being X-open or X-closed).

(b) Let $X = \mathbb{R}$ with the usual metric, i.e. $d(x, y) = |x - y|$ for each $x, y \in \mathbb{R}$. Let $M = [0, 1]$. According to (a), is both M-closed and M-open, even though $[0, 1]$ is not open in X. When noting before that $\mathbb{Q}$ and $]0, 1]$ are metric spaces that are not complete, we already considered metric subspaces of $\mathbb{R}$ without making use of the term subspace. If $M =]0, 1]$, then, again, M is both M-closed and M-open, even though $]0, 1]$ is neither closed nor open in X. Moreover, $]0, \frac{1}{2}]$ is M-closed (but not X-closed) and $]\frac{1}{2}, 1]$ is M-open (but not X-open).

Proposition B.6. *Let (M, d) be a metric subspace of a metric space (X, d).*

(a) *A subset A of M is M-open if, and only if, there is a set $O \subseteq X$ which is X-open and $A = O \cap M$.*

(b) *A subset A of M is M-closed if, and only if, there is a set $C \subseteq X$ which is X-closed and $A = C \cap M$.*

Proof. (a): Suppose A is M-open. Then, for each $a \in A$, there is $\epsilon_a > 0$ such that the open M-ball $B_{\epsilon_a,M}(a)$ is contained in A, i.e.

$$B_{\epsilon_a,M}(a) = M \cap B_{\epsilon_a}(a) \subseteq A. \tag{B.2}$$

Let $O := \bigcup_{a \in A} B_{\epsilon_a}(a)$. Then O is X-open by Th. 1.29(a). Moreover,

$$O \cap M = \bigcup_{a \in A} \big(M \cap B_{\epsilon_a}(a)\big) = \bigcup_{a \in A} B_{\epsilon_a,M}(a) = A, \tag{B.3}$$

where the last equality is due to (B.2) and the fact that $a \in A$ implies $a \in B_{\epsilon_a,M}(a)$.

Conversely, if $O \subseteq X$ is X-open and $A = O \cap M$, then each $a \in A$ is an X-interior point of O, i.e. there is $\epsilon > 0$ such that the open X-ball $B_\epsilon(a)$ is contained in O, i.e.

$B_\epsilon(a) \subseteq O$. Intersecting with M yields $B_{\epsilon,M}(a) = M \cap B_\epsilon(a) \subseteq M \cap O = A$, i.e. the open M-ball $B_{\epsilon,M}(a)$ is contained in A, showing that a is an M-interior point of A. As a was an arbitrary point of A, A is M-open.

(b): If A is M-closed, then $M \setminus A$ is M-open. According to (a), there is an X-open set $O \subseteq X$ such that $M \setminus A = M \cap O$. Then $C := X \setminus O$ is an X-closed set and $M \cap C = M \cap (X \setminus O) = M \setminus (M \cap O) = M \setminus (M \setminus A) = A$.

Conversely, if there is an X-closed set $C \subseteq X$ with $A = C \cap M$, then $O := X \setminus C$ is an X-open set satisfying $O \cap M = M \cap (X \setminus C) = M \setminus (C \cap M) = M \setminus A$. Thus, according to (a), $M \setminus A$ is M-open, i.e. A is M-closed. ■

B.2 Norm-Preserving and Isometric Maps

Definition B.7. (a) Given normed vector spaces $(X, \|\cdot\|_X)$ and $(Y, \|\cdot\|_Y)$ over $\mathbb{K}$, a function $f : X \longrightarrow Y$ is called *norm-preserving* if, and only if,

$$\big\|f(x)\big\|_Y = \|x\|_X \quad \text{for each } x \in X. \tag{B.4}$$

(b) Given metric spaces (X, d_X) and (Y, d_Y), a function $f : X \longrightarrow Y$ is called *distance-preserving* or *isometric* if, and only if,

$$d_Y\big(f(x), f(y)\big) = d_X(x, y) \quad \text{for each } x, y \in X. \tag{B.5}$$

Lemma B.8. *Given normed vector spaces $(X, \|\cdot\|_X)$ and $(Y, \|\cdot\|_Y)$ over $\mathbb{K}$, a $\mathbb{K}$-linear function $f : X \longrightarrow Y$ is norm-preserving if, and only if, f is isometric with respect to the induced metrics.*

Proof. The function f is norm-preserving if, and only if, $\big\|f(x)\big\|_Y = \|x\|_X$ for each $x \in X$. This, in turn is the case if, and only if,

$$\big\|f(x) - f(y)\big\|_Y = \big\|f(x - y)\big\|_Y = \|x - y\|_X \quad \text{for each } x, y \in X, \tag{B.6}$$

where it was used that f is linear. As (B.6) states that f is isometric with respect to the induced metrics, the proof is complete. ■

The following examples show that the assertion of Lem. B.8 becomes false if the word "linear" is omitted.

Example B.9. (a) Let $(X, \|\cdot\|_X)$ be a normed vector space over $\mathbb{K}$, and $f : X \longrightarrow \mathbb{K}$, $f(x) := \|x\|_X$. If we take $\|\cdot\|_Y$ to be the usual norm on $\mathbb{K}$, i.e. $\|y\|_Y := |y|$, then, for each $x \in X$, $\big\|f(x)\big\|_Y = \big|\|x\|_X\big| = \|x\|_X$, i.e. f is norm-preserving. However, if $\dim X > 0$ (i.e. if $X \neq \{0\}$), then f is *not* isometric with respect to the induced metrics: Take any $0 \neq x \in X$. One computes

$$\big\|f(x) - f(-x)\big\|_Y = \big|\|x\|_X - \|x\|_X\big| = 0 \neq \|x - (-x)\|_X = 2\|x\|_X. \tag{B.7}$$

(b) Consider $(X, \|\cdot\|_X)$, $(Y, \|\cdot\|_Y)$, where $X = Y = \mathbb{K}$ and $\|x\|_X = \|x\|_Y = |x|$ for each $x \in \mathbb{K}$. Then $f : X \longrightarrow Y$, $f(x) := 1 + x$, is isometric due to $|f(x) - f(y)| = |1 + x - (1 + y)| = |x - y|$, but f is *not* norm-preserving, since $0 = |0| \neq |f(0)| = 1$.

Lemma B.10. *Isometric functions between metric spaces are one-to-one (in particular, isometric functions between normed spaces are one-to-one).*

Proof. Let (X, d_X) and (Y, d_Y) be metric spaces, and let $f : X \longrightarrow Y$ be an isometric function, i.e. $d_Y\big(f(x), f(y)\big) = d_X(x, y)$ for each $x, y \in X$. If $x \neq y$, then $0 \neq d_X(x, y) = d_Y\big(f(x), f(y)\big)$. Thus, $f(x) \neq f(y)$, showing that f is one-to-one. ■

Example B.11. If a function between normed spaces is just norm-preserving, but not isometric, then this function is not necessarily one-to-one: To see this, we reemploy the function f from Ex. B.9(a), i.e. let $(X, \|\cdot\|_X)$ be a normed vector space over $\mathbb{K}$, $\dim X > 0$, and $f : X \longrightarrow \mathbb{K}$, $f(x) := \|x\|_X$. In Ex. B.9(a), we saw that f is norm-preserving, but not isometric. Since, for $x \neq 0$, one has $x \neq -x$, but $f(x) = \|x\|_X = f(-x)$, f is not one-to-one.

Remark B.12. If $(X, \|\cdot\|)$ is a normed space, d is the induced metric, and $M \subseteq X$, then (M, d) can be considered as the metric subspace of (X, d) according to Def. B.3. Thus, every subset of a normed space is turned into a metric space in a natural way. It is quite remarkable that actually *every* metric space arises in this way. That means, given any metric space (M, d), there exists a normed space $(X, \|\cdot\|)$ and an isometric (in particular, one-to-one) function $f : M \longrightarrow X$: One can choose X as the $\mathbb{R}$-vector space of bounded functions from M into $\mathbb{R}$ with the sup-norm (for $F \in X$, define $\|F\| := \sup\{|F(x)| : x \in M\}$) and $f : M \longrightarrow X$, $f(x) = f_x$, where $f_x : M \longrightarrow \mathbb{R}$, $f_x(y) = d(x, y) - d(x_0, y)$ with some fixed $x_0 \in M$. However, the normed space X can be very large (i.e. much larger than M), and, thus, in practice, it is not always useful to study X in order to learn more about the metric space M.

B.3 Uniform Continuity and Lipschitz Continuity

This section provides some additional important results regarding uniformly continuous functions (see Def. 1.49(b)) and Lipschitz continuous functions (see Def. 1.49(c)). We start with an auxiliary result:

Lemma B.13. *If f, g are real-valued functions on a set X, i.e. if $f, g : X \longrightarrow \mathbb{R}$, then, for each $x, y \in X$,*

$$\big|\max(f, g)(x) - \max(f, g)(y)\big| \leq \max\big\{|f(x) - f(y)|, |g(x) - g(y)|\big\}, \tag{B.8a}$$

$$\big|\min(f, g)(x) - \min(f, g)(y)\big| \leq \max\big\{|f(x) - f(y)|, |g(x) - g(y)|\big\}. \tag{B.8b}$$

Proof. By possibly switching the names of f and g, one can assume, without loss of generality, that $\max(f, g)(x) = f(x)$, i.e. $g(x) \leq f(x)$. If $g(y) \leq f(y)$ as well, then

$\big|\max(f,g)(x)-\max(f,g)(y)\big| = |f(x)-f(y)|$ and $\big|\min(f,g)(x)-\min(f,g)(y)\big| = |g(x)-g(y)|$, i.e. (B.8) is true. If $g(y) > f(y)$, then

$$\big|\max(f,g)(x)-\max(f,g)(y)\big| = |f(x)-g(y)| \leq \begin{cases} \leq |g(x)-g(y)| & \text{for } f(x) \leq g(y), \\ < f(x)-f(y) & \text{for } f(x) > g(y), \end{cases} \tag{B.9a}$$

$$\big|\min(f,g)(x)-\min(f,g)(y)\big| = |g(x)-f(y)| \leq \begin{cases} < |g(x)-g(y)| & \text{for } g(x) \leq f(y), \\ \leq f(x)-f(y) & \text{for } g(x) > f(y), \end{cases} \tag{B.9b}$$

showing that (B.8) holds in all cases. ■

Theorem B.14. *Let (X,d) be a metric space (e.g. a normed space), $(Y,\|\cdot\|)$ a normed vector space over $\mathbb{K}$, and assume that $f,g : X \longrightarrow Y$ are uniformly continuous. Then $f+g$ and λf are uniformly continuous for each $\lambda \in \mathbb{K}$, i.e. the set of all uniformly continuous functions from X into Y constitutes a subspace of the vector space $\mathcal{F}(X,Y)$ over $\mathbb{K}$. Moreover, if $Y = \mathbb{K} = \mathbb{R}$, then $\max(f,g)$, $\min(f,g)$, f^+, f^-, $|f|$ are all uniformly continuous.*

Proof. As f and g are uniformly continuous, given $\epsilon > 0$, there exist $\delta_f > 0$ and $\delta_g > 0$ such that, for each $x, y \in X$,

$$d(x,y) < \delta_f \;\Rightarrow\; \big\|f(x)-f(y)\big\| < \epsilon/2, \tag{B.10a}$$

$$d(x,y) < \delta_g \;\Rightarrow\; \big\|g(x)-g(y)\big\| < \epsilon/2. \tag{B.10b}$$

Thus, if $d(x,y) < \min\{\delta_f,\delta_g\}$, then

$$\big\|(f+g)(x)-(f+g)(y)\big\| \leq \big\|f(x)-f(y)\big\| + \big\|g(x)-g(y)\big\| < \frac{\epsilon}{2}+\frac{\epsilon}{2} = \epsilon, \tag{B.10c}$$

showing that $f+g$ is uniformly continuous. Next, if $\lambda = 0$, then $\lambda f \equiv 0$, and obviously uniformly continuous. For $\lambda \neq 0$, choose $\delta > 0$ such that $d(x,y) < \delta$ implies $\|f(x)-f(y)\| < \epsilon/|\lambda|$. Then

$$\big\|(\lambda f)(x)-(\lambda f)(y)\big\| = |\lambda|\big\|f(x)-f(y)\big\| < |\lambda|\,\frac{\epsilon}{|\lambda|} = \epsilon, \tag{B.10d}$$

showing that λf is uniformly continuous. If $Y = \mathbb{K} = \mathbb{R}$, then $d(x,y) < \min\{\delta_f,\delta_g\}$ together with Lem. B.13 implies

$$\big|\max(f,g)(x)-\max(f,g)(y)\big| < \epsilon/2 < \epsilon, \tag{B.10e}$$

$$\big|\min(f,g)(x)-\min(f,g)(y)\big| < \epsilon/2 < \epsilon, \tag{B.10f}$$

showing the uniform continuity of $\max(f,g)$ and $\min(f,g)$ and, in turn, also of f^+, f^-, and $|f|$. ■

Theorem B.15. *Let (X, d) be a metric space (e.g. a normed space), $(Y, \|\cdot\|)$ a normed vector space over $\mathbb{K}$, and assume that $f, g : X \longrightarrow Y$ are Lipschitz continuous. Then $f+g$ and λf are Lipschitz continuous for each $\lambda \in \mathbb{K}$, i.e. the set $\operatorname{Lip}(X, Y)$ constitutes a subspace of the vector space $\mathcal{F}(X, Y)$ over $\mathbb{K}$. Moreover, if $Y = \mathbb{K} = \mathbb{R}$, then $\max(f, g)$, $\min(f, g)$, f^+, f^-, $|f|$ are all Lipschitz continuous.*

Proof. As f and g are Lipschitz continuous, there exist $L_f \geq 0$ and $L_g \geq 0$ such that, for each $x, y \in X$,

$$\big\|f(x) - f(y)\big\| \leq L_f\, d(x, y), \tag{B.11a}$$
$$\big\|g(x) - g(y)\big\| \leq L_g\, d(x, y). \tag{B.11b}$$

Thus,

$$\begin{aligned}\big\|(f+g)(x) - (f+g)(y)\big\| &\leq \big\|f(x) - f(y)\big\| + \big\|g(x) - g(y)\big\| \\ &\leq L_f\, d(x, y) + L_g\, d(x, y) = (L_f + L_g) d(x, y),\end{aligned} \tag{B.11c}$$

showing that $f + g$ is Lipschitz continuous with Lipschitz constant $L_f + L_g$. Next, for $\lambda \in \mathbb{K}$,

$$\big\|(\lambda f)(x) - (\lambda f)(y)\big\| = |\lambda| \big\|f(x) - f(y)\big\| \leq |\lambda| L_f\, d(x, y), \tag{B.11d}$$

showing that λf is Lipschitz continuous with Lipschitz constant $|\lambda| L_f$. For $Y = \mathbb{K} = \mathbb{R}$, Lem. B.13 shows $\max(f, g)$ and $\min(f, g)$ are Lipschitz continuous with Lipschitz constant $\max\{L_f, L_g\}$, f^+ and f^- are Lipschitz continuous with Lipschitz constant L_f, and $|f|$ is Lipschitz continuous with Lipschitz constant $2L_f$. ■

Caveat B.16. Products and quotients of uniformly continuous functions are not necessarily uniformly continuous; products and quotients of Lipschitz continuous functions are not necessarily Lipschitz continuous: Even though $f \equiv 1$ and $g(x) = x$ are Lipschitz continuous, it was shown in Examples 1.52(a),(b), respectively, that f/g and g^2 are not even uniformly continuous on $\mathbb{R}^+$.

B.4 Viewing $\mathbb{C}^n$ as $\mathbb{R}^{2n}$

Remark B.17. Recall that the set of complex numbers $\mathbb{C}$ is *defined* to be $\mathbb{R}^2$, where the imaginary unit is $i := (0, 1) \in \mathbb{R}^2$, which allows to write each $z = (x, y) \in \mathbb{C} = \mathbb{R}^2$ as $z = x + iy$, where $x = \operatorname{Re} z$ and $y = \operatorname{Im} z$. This, for each $n \in \mathbb{N}$, gives rise to the $\mathbb{R}$-linear bijective map

$$I : \mathbb{C}^n \longrightarrow \mathbb{R}^{2n}, \quad I\big((x_1, y_1), \ldots, (x_n, y_n)\big) := (x_1, y_1, \ldots, x_n, y_n), \tag{B.12}$$

allowing to canonically identify $\mathbb{C}^n$ with $\mathbb{R}^{2n}$.

—

The identification (B.12) allows the identification of metric structures on $\mathbb{C}^n$ and $\mathbb{R}^{2n}$ due to the following general result:

Proposition B.18. *Let X, Y be sets, let $d : X \times X \longrightarrow \mathbb{R}_0^+$ be a metric on X, and let $I : X \longrightarrow Y$ be bijective. Then*

$$d_Y : Y \times Y \longrightarrow \mathbb{R}_0^+, \quad d_Y(x,y) := d\big(I^{-1}(x), I^{-1}(y)\big), \tag{B.13}$$

defines a metric on Y such that (X,d) and (Y,d_Y) are isometric (with the map I providing the isometry).

Proof. Let $x, y, z \in Y$. Then

$$d_Y(x,y) = 0 \quad \Leftrightarrow \quad d\big(I^{-1}(x), I^{-1}(y)\big) = 0 \quad \Leftrightarrow \quad I^{-1}(x) = I^{-1}(y) \quad \Leftrightarrow \quad x = y, \tag{B.14}$$

showing that d_Y is positive definite. Moreover,

$$d_Y(x,y) = d\big(I^{-1}(x), I^{-1}(y)\big) = d\big(I^{-1}(y), I^{-1}(x)\big) = d_Y(y,x), \tag{B.15}$$

showing d_Y is symmetric. Finally,

$$\begin{aligned} d_Y(x,z) &= d\big(I^{-1}(x), I^{-1}(z)\big) \leq d\big(I^{-1}(x), I^{-1}(y)\big) + d\big(I^{-1}(y), I^{-1}(z)\big) \\ &= d_Y(x,y) + d_Y(y,z), \end{aligned} \tag{B.16}$$

proving the triangle inequality for d_Y and completing the proof that d_Y constitutes a metric. That I provides an isometry between (X,d) and (Y,d_Y) is immediate from (B.13). ■

Corollary B.19. *Let $n \in \mathbb{N}$, let $d : \mathbb{C}^n \times \mathbb{C}^n \longrightarrow \mathbb{R}_0^+$ be a metric, and let I be the map from* (B.12)*. Then*

$$d_{\mathrm{r}} : \mathbb{R}^{2n} \times \mathbb{R}^{2n} \longrightarrow \mathbb{R}_0^+, \quad d_{\mathrm{r}}(x,y) := d\big(I^{-1}(x), I^{-1}(y)\big), \tag{B.17}$$

defines a metric on $\mathbb{R}^{2n}$ such that $(\mathbb{C}^n, d)$ and $(\mathbb{R}^{2n}, d_{\mathrm{r}})$ are isometric (with the map I providing the isometry). Moreover, the map $d \mapsto d_{\mathrm{r}}$ is bijective between the set of metrics on $\mathbb{C}^n$ and the set of metrics on $\mathbb{R}^{2n}$. ■

Proposition B.20. *Let $n \in \mathbb{N}$. If $\|\cdot\|$ constitutes a norm on the vector space $\mathbb{C}^n$ over $\mathbb{C}$ in the sense of Def.* 1.19*, then*

$$\|\cdot\|_{\mathrm{r}} : \mathbb{R}^{2n} \longrightarrow \mathbb{R}_0^+, \quad \|(x_1, y_1, \ldots, x_n, y_n)\|_{\mathrm{r}} := \big\|\big((x_1, y_1), \ldots, (x_n, y_n)\big)\big\| \tag{B.18}$$

defines a norm on the vector space $\mathbb{R}^{2n}$ over $\mathbb{R}$ such that $(\mathbb{C}^n, \|\cdot\|)$ and $(\mathbb{R}^{2n}, \|\cdot\|_{\mathrm{r}})$ are isometric (with the map I from (B.12) *providing the isometry – even more precisely, if d and d_{r} denote the respective induced metrics, then the relation between d and d_{r} is given by* (B.17)*).*

Proof. Exercise. ■

Example B.21. Let $n \in \mathbb{N}$, $p \in [1, \infty]$, and let $\|\cdot\|$ denote the p-norm on the vector space $\mathbb{R}^n$ over $\mathbb{R}$, i.e. $\|x\| := (\sum_{j=1}^n |x_j|^p)^{1/p}$ for $p < \infty$ and $\|x\| = \max\{|x_j| : j = 1, \ldots, n\}$ for $p = \infty$. Then it is an exercise to show

$$\|\cdot\|_{\mathrm{c}} : \mathbb{C}^n \longrightarrow \mathbb{R}_0^+, \quad \|(z_1, \ldots, z_n)\|_{\mathrm{c}} := \|(|z_1|, \ldots, |z_n|)\| \tag{B.19}$$

defines a norm on the vector space $\mathbb{C}^n$ over $\mathbb{C}$.

Remark B.22. As a consequence of Th. 1.95, every norm on the normed vector space $\mathbb{C}^n$ over $\mathbb{C}$ generates precisely the same open subsets of $\mathbb{C}^n$ – in other words, there is only one *norm topology* on $\mathbb{C}^n$. Analogously, there is only one norm topology on $\mathbb{R}^n$ as every norm on the normed vector space $\mathbb{R}^n$ over $\mathbb{R}$ generates precisely the same open subsets of $\mathbb{R}^n$ Moreover, Prop. B.20 shows that the open sets of the norm topology on $\mathbb{C}^n$ are actually precisely the same as the open sets of the norm topology on $\mathbb{R}^{2n}$.

Theorem B.23. *Let $n \in \mathbb{N}$, $A \subseteq \mathbb{C}^n$. Then A is bounded in the normed vector space $\mathbb{C}^n$ over $\mathbb{C}$ if, and only if, A is bounded in the normed vector space $\mathbb{R}^{2n}$ over $\mathbb{R}$.*

Proof. Exercise. ■

B.5 Banach Fixed Point Theorem a.k.a. Contraction Mapping Principle

Definition B.24. Let $\emptyset \neq A$ be a subset of a metric space (X, d). A map $\varphi : A \longrightarrow A$ is called a *contraction* if, and only if, there exists $0 \leq L < 1$ satisfying

$$d\big(\varphi(x), \varphi(y)\big) \leq L\, d(x, y) \quad \text{for each } x, y \in A. \tag{B.20}$$

Remark B.25. According to Def. B.24, $\varphi : A \longrightarrow A$ is a contraction if, and only if, φ is Lipschitz continuous with Lipschitz constant $L < 1$.

The following Th. B.26 constitutes the Banach fixed point theorem. It is also known as the contraction mapping principle. Its proof is surprisingly simple, e.g. about an order of magnitude easier than the proof of the Brouwer fixed point theorem.

Theorem B.26 (Banach Fixed Point Theorem)**.** *Let $\emptyset \neq A$ be a closed subset of a complete metric space (X, d) (for example, a Banach space). If $\varphi : A \longrightarrow A$ is a contraction with Lipschitz constant $0 \leq L < 1$, then φ has a unique fixed point $x_* \in A$. Moreover, for each initial value $x_0 \in A$, the sequence $(x_n)_{n \in \mathbb{N}_0}$, defined by*

$$x_{n+1} := \varphi(x_n) \quad \textit{for each } n \in \mathbb{N}_0, \tag{B.21}$$

converges to x_*:

$$\lim_{n \to \infty} \varphi^n(x_0) = x_*. \tag{B.22}$$

Furthermore, for each such sequence, we have the error estimate

$$d(x_n, x_*) \leq \frac{L}{1 - L}\, d(x_n, x_{n-1}) \leq \frac{L^n}{1 - L}\, d(x_1, x_0) \tag{B.23}$$

for each $n \in \mathbb{N}$.

Proof. We start with uniqueness: Let $x_*, x_{**} \in A$ be fixed points of φ. Then

$$d(x_*, x_{**}) = d\big(\varphi(x_*), \varphi(x_{**})\big) \leq L\, d(x_*, x_{**}), \tag{B.24}$$

which implies $1 \leq L$ for $d(x_*, x_{**}) > 0$. Thus, $L < 1$ implies $d(x_*, x_{**}) = 0$ and $x_* = x_{**}$. Next, we turn to existence. A simple induction on $m - n$ shows

$$\begin{gathered} d(x_{m+1}, x_m) \leq L\, d(x_m, x_{m-1}) \leq L^{m-n}\, d(x_{n+1}, x_n) \\ \text{for each } m, n \in \mathbb{N}_0,\ m > n. \end{gathered} \tag{B.25}$$

This, in turn, allows us to estimate, for each $n, k \in \mathbb{N}_0$:

$$\begin{aligned} d(x_{n+k}, x_n) &\leq \sum_{m=n}^{n+k-1} d(x_{m+1}, x_m) \overset{\text{(B.25)}}{\leq} \sum_{m=n}^{n+k-1} L^{m-n}\, d(x_{n+1}, x_n) \\ &\leq \frac{1}{1-L}\, d(x_{n+1}, x_n) \overset{\text{(B.25)}}{\leq} \frac{L^n}{1-L}\, d(x_1, x_0) \to 0 \quad \text{for } n \to \infty, \end{aligned} \tag{B.26}$$

establishing that $(x_n)_{n\in\mathbb{N}_0}$ constitutes a Cauchy sequence. Since X is complete, this Cauchy sequence must have a limit $x_* \in X$, and since the sequence is in A and A is closed, $x_* \in A$. The continuity of φ allows to take limits in (B.21), resulting in $x_* = \varphi(x_*)$, showing that x_* is a fixed point and proving existence.

Finally, the error estimate (B.23) follows from (B.26) by fixing n and taking the limit for $k \to \infty$. ■

Example B.27. Suppose, we are looking for a fixed point of the map $\varphi(x) = \cos x$ (or, equivalently, for a zero of $f(x) = \cos x - x$). To apply the Banach fixed point theorem, we need to restrict φ to a set A such that $\varphi(A) \subseteq A$. This is the case for $A := [0, 1]$. Moreover, $\varphi : A \longrightarrow A$ is a contraction, due to $\sin 1 < 1$ and the mean value theorem providing $\tau \in]0, 1[$, satisfying

$$\big|\varphi(x) - \varphi(y)\big| = \big|\varphi'(\tau)\big|\, |x - y| < (\sin 1)|x - y| \tag{B.27}$$

for each $x, y \in A$. Since $\mathbb{R}$ is complete and A is closed in $\mathbb{R}$, the Banach fixed point theorem yields the existence of a unique fixed point $x_* \in [0, 1]$ and $\lim \varphi^n(x_0) = x_*$ for each $x_0 \in [0, 1]$.

B.6 Unit Balls in Normed Spaces

The goal of this section is to prove that a normed vector space is finite-dimensional if, and only if, its closed unit ball is compact (see Th. B.29). In preparation, we show that finite-dimensional subspaces of normed vector spaces are always closed:

Theorem B.28. *Let $(X, \|\cdot\|)$ be a normed vector space over $\mathbb{K}$. If $U \subseteq X$ is a subspace such that $\dim U = n \in \mathbb{N}$, then U is closed.*

Proof. Let $(b_1, \dots, b_n)$ be a basis of U. Then

$$A : U \longrightarrow \mathbb{K}^n, \quad A\left(\sum_{k=1}^{n} \alpha_k\, b_k\right) := (\alpha_1, \dots, \alpha_n), \tag{B.28}$$

defines a linear isomorphism (cf. Th. A.24). We define a norm on $\mathbb{K}^n$ by letting

$$\|\cdot\| : \mathbb{K}^n \longrightarrow \mathbb{R}_0^+, \quad \|z\| := \big\|A^{-1}(z)\big\|. \tag{B.29}$$

Indeed, (B.29) defines a norm: $\|0\| = \big\|A^{-1}(0)\big\| = \|0\| = 0$; if $z \in \mathbb{K}^n$ and $\|z\| = \big\|A^{-1}(z)\big\| = 0$, then $A^{-1}(z) = 0$, i.e. $z = 0$, showing $\|\cdot\|$ to be positive definite. Moreover

$$\forall_{z\in\mathbb{K}^n} \quad \forall_{\lambda\in\mathbb{K}} \quad \|\lambda z\| = \big\|A^{-1}(\lambda z)\big\| = |\lambda|\,\big\|A^{-1}(z)\big\| = |\lambda|\,\|z\|, \tag{B.30}$$

showing $\|\cdot\|$ to be homogeneous of degree 1. Finally,

$$\forall_{z,w\in\mathbb{K}^n} \quad \|z+w\| = \big\|A^{-1}(z+w)\big\| \le \big\|A^{-1}(z)\big\| + \big\|A^{-1}(w)\big\| = \|z\| + \|w\|, \tag{B.31}$$

showing the triangle inequality to hold for $\|\cdot\|$.

Let $(u^k)_{k\in\mathbb{N}}$ be a sequence in U such that $\lim_{k\to\infty} u^k = x \in X$. Then $(u^k)_{k\in\mathbb{N}}$ is a Cauchy sequence and, as A is norm-preserving in consequence of (B.29), $(Au^k)_{k\in\mathbb{N}}$ is a Cauchy sequence in $\mathbb{K}^n$. Since $\mathbb{K}^n$ is complete, there is $z \in \mathbb{K}^n$ such that $\lim_{k\to\infty} Au^k = z$ and $\lim_{k\to\infty} u^k = A^{-1}z$, showing $x = A^{-1}z \in U$, i.e. U is closed. ■

Theorem B.29. *A normed vector space $(X, \|\cdot\|)$ over $\mathbb{K}$ is finite-dimensional if, and only if, its closed unit ball $\overline{B}_1(0)$ is compact.*

Proof. Let X be finite-dimensional. If $(b_1, \dots, b_n)$ denotes a basis of X, then (B.28) defines a linear isomorphism $A : X \longrightarrow \mathbb{K}^n$. If we define a norm on $\mathbb{K}^n$ via (B.29), then A^{-1} becomes norm-preserving and, in particular, continuous. Then $\overline{B}_1(0)$ in X must be compact as the countinuous image (under A^{-1}) of $\overline{B}_1(0)$ in $\mathbb{K}^n$.

Conversely, let X be infinite-dimensional. To show that $\overline{B}_1(0)$ is not compact, we construct, via recursion, a sequence $(x^k)_{k\in\mathbb{N}}$ in $\overline{B}_1(0)$ (actually in the sphere $S_1(0)$) that does not have a convergent subsequence: Fix $n \in \mathbb{N}$ and assume $(x^1, \dots, x^n)$ to be already constructed such that

$$\forall_{k\in\{1,\dots,n\}} \quad \|x^k\| = 1, \tag{B.32a}$$

$$\forall_{\substack{k,l\in\{1,\dots,n\},\\ k\neq l}} \quad \|x^k - x^l\| \ge \frac{1}{2}. \tag{B.32b}$$

Let $U := \operatorname{span}\{x^1, \dots, x^n\}$. Since X is infinite-dimensional, we have $U \neq X$. Let $x \in X \setminus U$. Since U is closed by Th. B.28, it is

$$d := \inf\big\{\|x - u\| : u \in U\big\} > 0. \tag{B.33}$$

Moreover, there exists $u_0 \in U$ such that $\|x - u_0\| \leq 2d$. Set

$$x^{n+1} := \frac{x - u_0}{\|x - u_0\|}. \tag{B.34}$$

Then $\|x^{n+1}\| = 1$ and, for each $u \in U$ is $\|x - u_0\|u + u_0 \in U$, implying

$$\|u - x^{n+1}\| = \frac{\big\|\|x - u_0\|u - x + u_0\big\|}{\|x - u_0\|} \geq \frac{d}{\|x - u_0\|} \geq \frac{1}{2}. \tag{B.35}$$

Thus, (B.32) holds with n replaced by $n+1$, where (B.32b) means that $(x^k)_{k\in\mathbb{N}}$ can not have a convergent subsequence. ■

C Differential Calculus in $\mathbb{R}^n$

C.1 Proof of the Chain Rule

Proof of Th. 2.28. As usual, we first consider the case $\mathbb{K} = \mathbb{R}$. Since f is differentiable at ξ and g is differentiable at $f(\xi)$, according to Lem. 2.21, there are functions $r_f : \mathbb{R}^n \longrightarrow \mathbb{R}^m$ and $r_g : \mathbb{R}^m \longrightarrow \mathbb{R}^p$ satisfying

$$r_f(h) = f(\xi + h) - f(\xi) - Df(\xi)(h), \tag{C.1a}$$

$$r_g(h) = g\big(f(\xi) + h\big) - g\big(f(\xi)\big) - Dg\big(f(\xi)\big)(h) \tag{C.1b}$$

for each h such that $\|h\|_2$ is sufficiently small, as well as

$$\lim_{h\to 0} \frac{r_f(h)}{\|h\|_2} = 0, \quad \lim_{h\to 0} \frac{r_g(h)}{\|h\|_2} = 0. \tag{C.2}$$

Defining $r_{g\circ f} : \mathbb{R}^n \longrightarrow \mathbb{R}^p$ by

$$r_{g\circ f}(h) := \begin{cases} (g \circ f)(\xi + h) - (g \circ f)(\xi) - \Big(Dg\big(f(\xi)\big) \circ Df(\xi)\Big)(h) & \text{for } \xi + h \in G_f, \\ 0 & \text{otherwise,} \end{cases} \tag{C.3}$$

it remains to show

$$\lim_{h\to 0} \frac{r_{g\circ f}(h)}{\|h\|_2} = 0. \tag{C.4}$$

For each $h \in \mathbb{R}^n$ with $\|h\|_2$ sufficiently small, we use (C.1) to compute

$$\begin{aligned}(g \circ f)(\xi + h) &= g\Big(f(\xi) + Df(\xi)(h) + r_f(h)\Big) \\ &= g\big(f(\xi)\big) + Dg\big(f(\xi)\big)\Big(Df(\xi)(h) + r_f(h)\Big) + r_g\Big(Df(\xi)(h) + r_f(h)\Big),\end{aligned} \tag{C.5a}$$

implying

$$r_{g\circ f}(h) = Dg\big(f(\xi)\big)\Big(r_f(h)\Big) + r_g\Big(Df(\xi)(h) + r_f(h)\Big). \tag{C.5b}$$

From Th. 1.67, we know that $Dg(f(\xi))$ is Lipschitz continuous with some Lipschitz constant $L_g \in \mathbb{R}_0^+$. Thus, for each $0 \neq h \in \mathbb{R}^n$,

$$0 \leq \frac{\Big\|Dg\big(f(\xi)\big)\Big(r_f(h)\Big)\Big\|_2}{\|h\|_2} \leq \frac{L_g \,\|r_f(h)\|_2}{\|h\|_2}, \tag{C.6a}$$

implying

$$\lim_{h\to 0} \frac{\Big\|Dg\big(f(\xi)\big)\Big(r_f(h)\Big)\Big\|_2}{\|h\|_2} = 0 \tag{C.6b}$$

due to (C.2). Thus, to prove (C.4), it merely remains to show

$$\lim_{h\to 0} \frac{\Big\|r_g\Big(Df(\xi)(h) + r_f(h)\Big)\Big\|_2}{\|h\|_2} = 0. \tag{C.7}$$

To that end, we rewrite, for $Df(\xi)(h) + r_f(h) \neq 0$,

$$\frac{\Big\|r_g\Big(Df(\xi)(h) + r_f(h)\Big)\Big\|_2}{\|h\|_2} = \frac{\big\|Df(\xi)(h) + r_f(h)\big\|_2 \Big\|r_g\Big(Df(\xi)(h) + r_f(h)\Big)\Big\|_2}{\|h\|_2 \big\|Df(\xi)(h) + r_f(h)\big\|_2}. \tag{C.8a}$$

Next, note

$$\lim_{h\to 0} \big\|Df(\xi)(h) + r_f(h)\big\|_2 = 0 \quad \overset{\text{(C.2)}}{\Rightarrow} \quad \lim_{h\to 0} \frac{\Big\|r_g\Big(Df(\xi)(h) + r_f(h)\Big)\Big\|_2}{\big\|Df(\xi)(h) + r_f(h)\big\|_2} = 0. \tag{C.8b}$$

Once again, from Th. 1.67, we know that $Df(\xi)$ is Lipschitz continuous with some Lipschitz constant $L_f \in \mathbb{R}_0^+$, implying

$$\frac{\big\|Df(\xi)(h) + r_f(h)\big\|_2}{\|h\|_2} \leq \frac{\big\|Df(\xi)(h)\big\|_2 + \big\|r_f(h)\big\|_2}{\|h\|_2} \leq L_f + 1 \tag{C.8c}$$

for $0 \neq \|h\|_2$ sufficiently small. Combining (C.8a) – (C.8c) proves (C.7) and, thus, (C.4). Together with (C.3) and Lem. 2.21, this shows that $g \circ f$ is differentiable at ξ with $D(g \circ f)(\xi) = Dg\big(f(\xi)\big) \circ Df(\xi)$.

In the case $\mathbb{K} = \mathbb{C}$, we can apply the case $\mathbb{K} = \mathbb{R}$ to obtain the differentiability of $\operatorname{Re}(g \circ f) = (\operatorname{Re} g) \circ f$ and of $\operatorname{Im}(g \circ f) = (\operatorname{Im} g) \circ f$ at ξ, and, in consequence, the differentiability of $g \circ f$ at ξ. Moreover, to verify the chain rule, we use the chain rule of the case $\mathbb{K} = \mathbb{R}$ to compute

$$\begin{aligned} D(g \circ f)(\xi) &= D\operatorname{Re}(g \circ f)(\xi) + i\, D\operatorname{Im}(g \circ f)(\xi) \\ &= D\big((\operatorname{Re} g) \circ f\big)(\xi) + i\, D\big((\operatorname{Im} g) \circ f\big)(\xi) \\ &= D\operatorname{Re} g\big(f(\xi)\big) \circ Df(\xi) + i\, D\operatorname{Im} g\big(f(\xi)\big) \circ Df(\xi) \\ &= Dg\big(f(\xi)\big) \circ Df(\xi), \end{aligned} \tag{C.9}$$

thereby completing the proof. ■

C.2 Bounded Derivatives Imply Lipschitz Continuity

First, we provide an $\mathbb{R}^m$-valued variant of Th. 2.35:

Theorem C.1. *Let $m, n \in \mathbb{N}$, let $G \subseteq \mathbb{R}^n$ be open, and let $f : G \longrightarrow \mathbb{R}^m$ be differentiable. Suppose there exists $M \in \mathbb{R}_0^+$ such that $|\partial_k f_l(\xi)| \leq M$ for each $k \in \{1, \ldots, n\}$, each $l \in \{1, \ldots, m\}$, and each $\xi \in G$. If G is convex, then f is Lipschitz continuous with Lipschitz constant $L := mM$ with respect to the 1-norms on $\mathbb{R}^n$ and $\mathbb{R}^m$ and with Lipschitz constant cL, $c > 0$, with respect to arbitrary norms on $\mathbb{R}^n$ and $\mathbb{R}^m$.*

Proof. According to Th. 2.35, each f_l is M-Lipschitz with respect to the 1-norm on $\mathbb{R}^n$. Thus, we obtain, for each $x, y \in G$,

$$\|f(y) - f(x)\|_1 = \sum_{l=1}^{m} |f_l(y) - f_l(x)| \leq mM\|y - x\|_1, \tag{C.10}$$

showing that, with respect to the 1-norms on $\mathbb{R}^n$ and $\mathbb{R}^m$, f is Lipschitz continuous with Lipschitz constant mM. Since all norms on $\mathbb{R}^n$ and $\mathbb{R}^m$ are equivalent, we also get that f is Lipschitz continuous with Lipschitz constant cL, $c > 0$, with respect to all other norms on $\mathbb{R}^n$ and $\mathbb{R}^m$. ■

It is sometimes useful if the bound on the derivatives is the same as the resulting Lipschitz constant (which, for $m > 1$, is not the case in the above Th. C.1). The following Th. C.3 provides a variant, where the constants are the same, formulated for functions $f : I \longrightarrow \mathbb{R}^n$, defined on open intervals $I \subseteq \mathbb{R}$, and making use of the Euclidean norm $\|\cdot\|_2$ on $\mathbb{R}^n$. We will start with some auxiliary results regarding the Euclidean norm and the Euclidean inner product:

Proposition C.2. *Let $I \subseteq \mathbb{R}$ be an open interval, and let $g, h : I \longrightarrow \mathbb{R}^n$ be differentiable, $n \in \mathbb{N}$.*

(a) *The function*

$$f : I \longrightarrow \mathbb{R},\ f(x) := g(x) \bullet h(x) = \sum_{j=1}^{n} g_j(x) h_j(x), \tag{C.11}$$

is differentiable and

$$f' : I \longrightarrow \mathbb{R},\ f'(x) = g'(x) \bullet h(x) + h(x) \bullet h'(x). \tag{C.12}$$

(b) *The function*

$$\alpha : I \longrightarrow \mathbb{R}, \quad \alpha(x) := \|g(x)\|_2 = \sqrt{g(x) \bullet g(x)}, \tag{C.13}$$

is differentiable at each $x \in I$ such that $g(x) \neq 0$. Moreover,

$$\underset{\substack{x \in I,\\ g(x) \neq 0}}{\forall} \quad \alpha'(x) = \frac{g(x) \bullet g'(x)}{\alpha(x)} = \frac{g(x) \bullet g'(x)}{\|g(x)\|_2}. \tag{C.14}$$

Proof. (a) is immediate from the product rule.

(b) is an easy consequence of (a), as (a) implies α to be differentiable at each $x \in I$ such that $g(x) \neq 0$, and

$$\underset{\substack{x\in I,\\ g(x)\neq 0}}{\forall} \quad \alpha'(x) = \frac{2\, g(x) \bullet g'(x)}{2\,\sqrt{g(x)\bullet g(x)}} = \frac{g(x)\bullet g'(x)}{\alpha(x)}, \tag{C.15}$$

completing the proof. ■

Theorem C.3. *Let $a, b \in \mathbb{R}$ with $a < b$ and let $f :]a, b[\longrightarrow \mathbb{R}^n$ be differentiable with uniformly bounded derivative, i.e. with*

$$\underset{M\in\mathbb{R}_0^+}{\exists} \quad \underset{x\in]a,b[}{\forall} \quad \|f'(x)\|_2 = \sqrt{\sum_{j=1}^{n} |f_j'(x)|^2} \leq M. \tag{C.16}$$

Then f is M-Lipschitz, i.e.

$$\underset{x_1,x_2\in]a,b[}{\forall} \quad \|f(x_1) - f(x_2)\|_2 \leq M\,|x_1 - x_2|. \tag{C.17}$$

Proof. For $x_1 = x_2$, there is nothing to prove. Thus, assume $x_1 \neq x_2$ and define the auxiliary function

$$g : [0,1] \longrightarrow \mathbb{R}^n, \quad g(t) := f\big(x_1 + t(x_2 - x_1)\big) - f(x_1). \tag{C.18}$$

According to the chain rule of Th. 2.28, g is differentiable on $]0,1[$ and

$$\underset{t\in]0,1[}{\forall} \quad g'(t) = (x_2 - x_1) f'\big(x_1 + t(x_2 - x_1)\big), \tag{C.19}$$

implying

$$\underset{t\in]0,1[}{\forall} \quad \|g'(t)\|_2 \leq M\,|x_1 - x_2|. \tag{C.20}$$

We now introduce another auxiliary function, namely

$$\alpha : [0,1] \longrightarrow \mathbb{R}, \quad \alpha(t) := \|g(t)\|_2. \tag{C.21}$$

Then α is continuous (as f and the norm are both continuous), satisfying $\alpha(0) = \|g(0)\|_2 = 0$ and $\alpha(1) = \|g(1)\|_2 = \|f(x_2) - f(x_1)\|_2$. If $\alpha(1) = 0$, then (C.17) is trivially true, and, thus, we proceed to assume $\alpha(1) > 0$. Then the continuity of α implies

$$s := \sup\big\{t \in [0,1] :\, \alpha(t) = 0\big\} < 1, \quad \alpha(s) = 0. \tag{C.22}$$

In consequence, α is positive on $]s, 1[$ and, thus, differentiable on $]s, 1[$ by Prop. C.2(b). The mean value theorem [Phi15a, Th. 9.17] implies the existence of $\sigma \in]s, 1[$ such that

$$\begin{aligned} \alpha(1) \quad &= \quad \alpha(1) - \alpha(s) = (1-s)\,\alpha'(\sigma) \overset{(C.14)}{=} (1-s)\,\frac{g(\sigma)\bullet g'(\sigma)}{\alpha(\sigma)} \\ &\overset{(1.81)}{\leq} \quad (1-s)\,\frac{\|g(\sigma)\|_2 \|g'(\sigma)\|_2}{\|g(\sigma)\|_2} \overset{(C.20)}{\leq} (1-s)\,M\,|x_1 - x_2| \\ &\leq \quad M\,|x_1 - x_2|, \end{aligned} \tag{C.23}$$

which establishes the case. ■

C.3 Surjectivity of Directional Derivatives

We finish the proof of Th. 2.38 by showing that, for $n \geq 2$, the map

$$D:\, S_1(0) \longrightarrow [-\alpha,\alpha], \quad D(e) := \nabla f(\xi)\cdot e = \sum_{j=1}^{n} \epsilon_j \partial_j f(\xi), \quad \alpha = \|\nabla f(\xi)\|_2, \tag{C.24}$$

is surjective (we already know from (2.65) that $D(e) \in [-\alpha,\alpha]$ for each $e \in S_1(0)$). We also recall $e_{\max} = \nabla f(\xi)/\alpha$, $e_{\min} = -e_{\max}$, $D(e_{\max}) = \alpha$, $D(e_{\min}) = -\alpha$.

The idea is to rotate $e_{\max}$ into $e_{\min}$. This can be achieved using a suitable function

$$\rho:\, [0,\pi] \longrightarrow S_1(0) \subseteq \mathbb{R}^n, \quad \rho = (\rho_1,\ldots,\rho_n).$$

We have to define ρ differently, depending on $n \geq 2$ being even or odd. To this end, let $(\epsilon_1,\ldots,\epsilon_n) := e_{\max}$. If n is even, then define

$$\forall_{j\in\{1,\ldots,n\}} \quad \rho_j:\, [0,\pi] \longrightarrow [-1,1], \quad \rho_j(\theta) := \begin{cases} \epsilon_j \cos\theta + \epsilon_{j+1}\sin\theta & \text{if } j \text{ is odd,} \\ -\epsilon_{j-1}\sin\theta + \epsilon_j \cos\theta & \text{if } j \text{ is even;} \end{cases} \tag{C.25a}$$

if n is odd (note $n \geq 3$ in this case), then define

$$\rho_j:\, [0,\pi] \longrightarrow [-1,1],$$
$$\forall_{j\in\{1,\ldots,n\}} \quad \rho_j(\theta) := \begin{cases} \epsilon_j \cos\theta + \epsilon_{j+1}\sin\theta & \text{if } j < n-2 \text{ is odd,} \\ -\epsilon_{j-1}\sin\theta + \epsilon_j\cos\theta & \text{if } j < n-2 \text{ is even,} \\ \epsilon_{n-2}\cos\theta + \sqrt{\epsilon_{n-1}^2+\epsilon_n^2}\,\sin\theta & \text{if } j = n-2, \\ \epsilon_{n-1}\cos\theta - \frac{\epsilon_{n-2}\,\epsilon_{n-1}}{\sqrt{\epsilon_{n-1}^2+\epsilon_n^2}}\,\sin\theta & \text{if } j = n-1, \\ \epsilon_n\cos\theta - \frac{\epsilon_{n-2}\,\epsilon_n}{\sqrt{\epsilon_{n-1}^2+\epsilon_n^2}}\,\sin\theta & \text{if } j = n. \end{cases} \tag{C.25b}$$

For the sake of readability, we assumed $\epsilon_{n-1} \neq 0$ or $\epsilon_n \neq 0$ in (C.25b). There is always at least one $j_0 \in \{1,\ldots,n\}$ such that $\epsilon_{j_0} \neq 0$. If $j_0 \notin \{n-1,n\}$, then one merely needs to interchange the roles of j_0 and n in (C.25b).

Clearly, for every $n \geq 2$, each ρ_j is continuous, i.e. ρ is continuous.

Next, we verify that ρ, indeed, maps into $S_1(0)$ (which, in particular, implies each ρ_j maps into $[-1,1]$): If $n \geq 2$ is even, then, for each odd $j \leq n-1$, one has

$$\begin{aligned} &(\rho_j(\theta))^2 + (\rho_{j+1}(\theta))^2 \\ &= (\epsilon_j\cos\theta + \epsilon_{j+1}\sin\theta)^2 + (-\epsilon_j\sin\theta + \epsilon_{j+1}\cos\theta)^2 \\ &= \epsilon_j^2\cos^2\theta + 2\epsilon_j\epsilon_{j+1}\cos\theta\sin\theta + \epsilon_{j+1}^2\sin^2\theta \\ &\quad + \epsilon_j^2\sin\theta - 2\epsilon_j\epsilon_{j+1}\cos\theta\sin\theta + \epsilon_{j+1}^2\cos^2\theta \\ &= \epsilon_j^2(\cos^2\theta+\sin^2\theta) + \epsilon_{j+1}^2(\cos^2\theta+\sin^2\theta) = \epsilon_j^2 + \epsilon_{j+1}^2, \end{aligned} \qquad \forall_{\theta\in[0,\pi]} \tag{C.26}$$

implying

$$\forall_{\theta\in[0,\pi]} \quad \|\rho(\theta)\|_2^2 = \sum_{j=1}^{n} (\rho_j(\theta))^2 = \sum_{j=1}^{n} \epsilon_j^2 = 1. \tag{C.27}$$

If $n \geq 3$ is odd, then (C.26) still holds for each odd $j \leq n-4$. Additionally,

$$\begin{aligned}
&(\rho_{n-2}(\theta))^2 + (\rho_{n-1}(\theta))^2 + (\rho_n(\theta))^2 \\
&= \epsilon_1^2 \cos^2\theta + 2\epsilon_1\sqrt{\epsilon_2^2+\epsilon_3^2}\,\sin\theta\cos\theta + (\epsilon_2^2+\epsilon_3^2)\,\sin^2\theta \\
&\quad + \epsilon_2^2\cos^2\theta - 2\frac{\epsilon_1\,\epsilon_2^2}{\sqrt{\epsilon_2^2+\epsilon_3^2}}\,\sin\theta\cos\theta + \frac{\epsilon_1^2\,\epsilon_2^2}{\epsilon_2^2+\epsilon_3^2}\,\sin^2\theta \\
&\quad + \epsilon_3^2\cos^2\theta - 2\frac{\epsilon_1\,\epsilon_3^2}{\sqrt{\epsilon_2^2+\epsilon_3^2}}\,\sin\theta\cos\theta + \frac{\epsilon_1^2\,\epsilon_3^2}{\epsilon_2^2+\epsilon_3^2}\,\sin^2\theta \\
&\underset{\theta\in[0,\pi]}{\forall} \\
&= (\epsilon_1^2+\epsilon_2^2+\epsilon_3^2)\,\cos^2\theta \\
&\quad + \frac{2\epsilon_1\,(\epsilon_2^2+\epsilon_3^2-\epsilon_2^2-\epsilon_3^2)}{\sqrt{\epsilon_2^2+\epsilon_3^2}}\,\sin\theta\cos\theta \\
&\quad + \left(\epsilon_2^2+\epsilon_3^2+\frac{\epsilon_1^2\,(\epsilon_2^2+\epsilon_3^2)}{\epsilon_2^2+\epsilon_3^2}\right)\sin^2\theta \\
&= (\epsilon_1^2+\epsilon_2^2+\epsilon_3^2)\,(\cos^2\theta+\sin^2\theta) = \epsilon_{n-2}^2+\epsilon_{n-1}^2+\epsilon_n^2,
\end{aligned} \tag{C.28}$$

i.e. (C.27) is true also for each $n \geq 3$ odd.

Clearly, D is also continuous and, thus, so is $D \circ \rho : [0,\pi] \longrightarrow [-\alpha,\alpha]$. Moreover, as $\sin(0) = \sin(\pi) = 0$, $\cos(0) = 1$, $\cos(\pi) = -1$, we obtain

$$\underset{n\geq 2}{\forall}\quad \underset{j\in\{1,\ldots,n\}}{\forall}\quad \Big(\rho_j(0) = \epsilon_j \quad\wedge\quad \rho_j(\pi) = -\epsilon_j\Big), \tag{C.29}$$

implying

$$\underset{n\geq 2}{\forall}\quad \Big(\rho(0) = e_{\max} \quad\wedge\quad (D\circ\rho)(0) = \alpha \quad\wedge\quad \rho(\pi) = e_{\min} \quad\wedge\quad (D\circ\rho)(\pi) = -\alpha\Big). \tag{C.30}$$

The continuity of $D \circ \rho$ and the intermediate value theorem [Phi15a, Th. 7.57] imply $D \circ \rho$ to be surjective, i.e. D must be surjective as well.

C.4 Implicit Function Theorem

We start with a preparatory proposition.

Proposition C.4. *Let $\|\cdot\|$ be some norm on $\mathbb{R}^n$, $n \in \mathbb{N}$. Moreover, let $a \in \mathbb{R}^n$, $r > 0$, and let $f : B_r(a) \longrightarrow \mathbb{R}^n$ be defined on the open r-ball with center a with respect to $\|\cdot\|$. If A is an invertible $n \times n$ matrix over $\mathbb{R}$ such that*

$$\left\|A^{-1}f(a)\right\| < \frac{r}{2} \tag{C.31}$$

and such that the map

$$F : B_r(a) \longrightarrow \mathbb{R}^n, \quad F(x) := x - A^{-1}f(x), \tag{C.32}$$

is Lipschitz continuous with Lipschitz constant $L = 1/2$, then f has a unique zero $\xi \in B_r(a)$. Moreover, for each $x_0 \in B_r(a)$, ξ is the limit of the sequence $(x_k)_{k\in\mathbb{N}_0}$, recursively defined by

$$\forall_{k\in\mathbb{N}_0} \quad x_{k+1} := F(x_k). \tag{C.33}$$

Proof. Set

$$s_0 := \max\left\{2\left\|A^{-1}f(a)\right\|, \|x_0 - a\|\right\} \overset{(C.31)}{\in} [0, r[. \tag{C.34}$$

The idea is to show that, for each $s_0 < s < r$, the Banach fixed point Th. B.26 applies to the contraction

$$F_s : \overline{B}_s(a) \longrightarrow \overline{B}_s(a), \quad F_s(x) := F(x). \tag{C.35}$$

We verify that F_s, indeed, maps $\overline{B}_s(a)$ into $\overline{B}_s(a)$: If $x \in \overline{B}_s(a)$, then

$$\begin{aligned}\left\|F(x) - a\right\| &\le \left\|F(x) - F(a)\right\| + \left\|F(a) - a\right\| \le \frac{1}{2}\|x - a\| + \left\|A^{-1}f(a)\right\| \\ &\le \frac{s}{2} + \frac{s_0}{2} < s,\end{aligned} \tag{C.36}$$

showing $F_s(x) \in \overline{B}_s(a)$ (in particular, this shows the x_k are well-defined by (C.33)). As F is Lipschitz continuous with Lipschitz constant $L = 1/2$, so is F_s, i.e. F_s is, indeed, a contraction. As $\overline{B}_s(a)$ is closed, the Banach fixed point Th. B.26 yields that F_s has a unique fixed point ξ and, moreover, $\xi = \lim_{k\to\infty} x_k$. Since this holds for each $s \in]s_0, r[$, ξ must also be the unique fixed point of F. The proof is concluded by noting

$$\forall_{y\in B_r(a)} \quad f(y) = 0 \quad \Leftrightarrow \quad F(y) = y - A^{-1}f(y) = y, \tag{C.37}$$

that means y is a zero of f if, and only if, y is a fixed point of F. ∎

Remark C.5. If the map f in Prop. C.4 is differentiable with invertible derivatives $Df(x)$, and if, instead of using a constant matrix A in the definition of (C.33), one uses $(Df(x_k))^{-1}$, then the iteration defined by (C.33) is known as *Newton's method* (in n dimensions, cf. [Phi15b, Sec. 6.3]). In consequence, if $A \approx (Df(x_k))^{-1}$ in (C.33), then the defined iteration is sometimes referred to as a *simplified Newton's method*.

Notation C.6. Let $k, m, n \in \mathbb{N}$, let $G \subseteq \mathbb{R}^n \times \mathbb{R}^m$ be open, and consider a map $f : G \longrightarrow \mathbb{R}^k$. If $(\xi, \eta) \in G$ and f is differentiable at (ξ, η), then let $D_y f(\xi, \eta)$ and $D_x f(\xi, \eta)$ denote the linear maps

$$D_y f(\xi,\eta) : \mathbb{R}^m \longrightarrow \mathbb{R}^k, \qquad \big(D_y f(\xi,\eta)\big)(h) := \big(Df(\xi,\eta)\big)(0,h), \tag{C.38a}$$

$$D_x f(\xi,\eta) : \mathbb{R}^n \longrightarrow \mathbb{R}^k, \qquad \big(D_x f(\xi,\eta)\big)(h) := \big(Df(\xi,\eta)\big)(h,0), \tag{C.38b}$$

respectively.

Theorem C.7 (Implicit Function Theorem)**.** *Let $m, n \in \mathbb{N}$, let $G \subseteq \mathbb{R}^n \times \mathbb{R}^m$ be open, and let $f : G \longrightarrow \mathbb{R}^m$ be continuously differentiable, i.e. $f \in C^1(G, \mathbb{R}^m)$. If $(\xi, \eta) \in G$ is such that*

$$f(\xi,\eta) = 0 \quad \textit{and} \quad A := D_y f(\xi,\eta) \textit{ is invertible,} \tag{C.39}$$

then there exist open neighborhoods $U_\xi \subseteq \mathbb{R}^n$ of ξ and $V_\eta \subseteq \mathbb{R}^m$ of η, and a continuously differentiable map $g : U_\xi \longrightarrow V_\eta$ such that the zeros of f in $U_\xi \times V_\eta$ are given precisely by the graph of g, i.e.

$$(U_\xi \times V_\eta) \cap f^{-1}\{0\} = \{(x, g(x)) : x \in U_\xi\}, \tag{C.40a}$$

which can be restated as

$$\underset{(x,y)\in U_\xi \times V_\eta}{\forall} \quad \Big(f(x,y) = 0 \quad \Leftrightarrow \quad y = g(x)\Big). \tag{C.40b}$$

Moreover,

$$\underset{x\in U_\xi}{\forall} \quad Dg(x) = -\big(D_y f(x, g(x))\big)^{-1} D_x f(x, g(x)) \tag{C.41}$$

and, if $f \in C^\alpha(G, \mathbb{R}^m)$, $\alpha \in \mathbb{N} \cup \{\infty\}$, then $g \in C^\alpha(U_\xi, \mathbb{R}^m)$.

Proof. Fix some arbitrary norms on $\mathbb{R}^n$ and on the set $\mathcal{M}(m, \mathbb{R})$ of real $m \times m$ matrices (for readability's sake, we will denote both norms by $\| \cdot \|$). On $\mathbb{R}^m$, we will use the 1-norm $\| \cdot \|_1$ to apply Th. C.1. According to the hypothesis, A is invertible. Thus, $\det(A) > 0$. Since the map $B \mapsto \det(B)$ is continuous (cf. Ex. 1.66(a)), and the map $D_y f : G \longrightarrow \mathcal{M}(m, \mathbb{R})$ is continuous due to the assumed continuous differentiability of f, the set

$$G_0 := \big\{(x, y) \in G : \det\big(D_y f(x, y)\big) > 0\big\} \subseteq G \tag{C.42}$$

is an open neighborhood of (ξ, η). Next, we consider the map

$$F : G_0 \longrightarrow \mathbb{R}^m, \quad F(x, y) := y - A^{-1} f(x, y). \tag{C.43}$$

Then F is continuously differentiable with

$$D_y F : G_0 \longrightarrow \mathcal{M}(m, \mathbb{R}), \quad D_y F(x, y) = \mathrm{Id} - A^{-1} D_y f(x, y), \tag{C.44}$$

being continuous as well. Thus, since $D_y F(\xi, \eta) = \mathrm{Id} - A^{-1}A = 0$, there exists $r > 0$ such that the open r-balls $B_r(\xi) \subseteq \mathbb{R}^n$ and $B_r(\eta) \subseteq \mathbb{R}^m$ satisfy

$$(\xi, \eta) \in B_r(\xi) \times B_r(\eta) \subseteq \left\{(x, y) \in G_0 : \underset{k,l=1,\ldots,m}{\forall} \big|\partial_{y_k} F_l(x, y)\big| < \frac{1}{2m}\right\} \subseteq G_0. \tag{C.45}$$

As we assume f to be continuous with $f(\xi, \eta) = 0$, there exists $s \in]0, r]$ such that

$$B_s(\xi) \subseteq \left\{x \in \mathbb{R}^n : \big\|A^{-1} f(x, \eta)\big\|_1 < \frac{r}{2}\right\} \subseteq B_r(\xi). \tag{C.46}$$

To construct the map $g : B_s(\xi) \longrightarrow B_r(\eta)$, we fix $x \in B_s(\xi)$ and apply Prop. C.4 to the function

$$f_x : B_r(\eta) \longrightarrow \mathbb{R}^m, \quad f_x(y) := f(x, y). \tag{C.47}$$

To verify that the hypotheses of Prop. C.4 are satisfied, we observe $\|A^{-1} f_x(\eta)\|_1 < \frac{r}{2}$ holds due to $x \in B_s(\xi)$ and (C.46), the map $F_x : B_r(\eta) \longrightarrow \mathbb{R}^m$, $F_x(y) := y - A^{-1} f_x(y) = F(x, y)$, is Lipschitz continuous with Lipschitz constant $L = m\frac{1}{2m} = \frac{1}{2}$ due to (C.45)

and Th. C.1. Thus, according to Prop. C.4, the function f_x has a unique zero $g(x)$ in $B_r(\eta)$, which defines the function g.

Note that, in the above argument, for each $0 < \rho < r$, one can choose $s(\rho) < s$ such that (C.46) holds with s replaced by $s(\rho)$ and r replaced by ρ, then showing that g maps $B_{s(\rho)}$ into B_ρ. We now choose some arbitrary $\rho \in]0, r[$ and set

$$U_\xi := B_{s(\rho)}(\xi), \quad V_\eta := B_\rho(\eta) \tag{C.48}$$

for the desired neighborhoods of the theorem. We verify g to be continuous on U_ξ: Let $x \in U_\xi$ and let $(x_k)_{k\in\mathbb{N}}$ be a sequence in U_ξ with $\lim_{k\to\infty} x_k = x$. We have to show $\lim_{k\to\infty} g(x_k) = g(x)$. If $\lim_{k\to\infty} g(x_k) = g(x)$ does not hold, then, without loss of generality, we may assume that there exists $\epsilon > 0$ such that $\|g(x_k) - g(x)\| > \epsilon$ for each $k \in \mathbb{N}$ (after having replaced $(x_k)_{k\in\mathbb{N}}$ with a suitable subsequence). Moreover, we may replace $(x_k)_{k\in\mathbb{N}}$ with another subsequence such that there exists $y \in \overline{B}_\rho(\eta) \subseteq B_r(\eta)$ satisfying $y = \lim_{k\to\infty} g(x_k)$ (this is due to the Bolzano-Weierstrass Th. 1.16(b), as $g(x_k) \in B_\rho(\eta)$ for each $k \in \mathbb{N}$). Then the continuity of f implies

$$f(x, y) = \lim_{k\to\infty} f\big(x_k, g(x_k)\big) = 0, \tag{C.49}$$

showing $g(x) = y = \lim_{k\to\infty} g(x_k)$ (due to (C.40b) – here we need $y \in B_r(\eta)$, which was the reason for choosing $\rho < r$), which is in contradiction to the choice of the x_k, and proves the continuity of g.

Next, we show that g is differentiable at each $x \in U_\xi$, where the derivative is given by (C.41): To this end, let $x \in U_\xi$ and note the existence of $\big(D_y f(x, g(x))\big)^{-1}$ due to $(x, g(x)) \in U_\xi \times V_\eta \subseteq G_0$. According to Def. 2.19, we have to show

$$\lim_{h\to 0} \frac{g(x+h) - g(x) + \big(D_y f(x, g(x))\big)^{-1} D_x f(x, g(x))\, h}{\|h\|} = 0. \tag{C.50}$$

Let $0 \neq h \in \mathbb{R}^n$ be sufficiently small such that $x+h \in U_\xi$. Using the notation of the mean value Th. 2.32, for each $l \in \{1, \dots, m\}$, there exist $x_{h,l} \in S_{x,x+h}$ and $y_{h,l} \in S_{g(x),g(x+h)}$ such that

$$\begin{aligned} 0 &= f_l\big(x+h, g(x+h)\big) - f_l\big(x, g(x)\big) \\ &= f_l\big(x+h, g(x+h)\big) - f_l\big(x, g(x+h)\big) + f_l\big(x, g(x+h)\big) - f_l\big(x, g(x)\big) \\ &= D_x f_l\big(x_{h,l}, g(x+h)\big)(h) + D_y f_l\big(x, y_{h,l}\big)\big(g(x+h) - g(x)\big). \end{aligned} \tag{C.51}$$

Note that the two derivatives occurring in (C.51) have the form of gradients, which, according to our usual convention, we can interpret as row vectors. Joining m row vectors into a matrix, we can write the m equations of (C.51) in matrix form as

$$0 = X_h h + Y_h\big(g(x+h) - g(x)\big), \tag{C.52}$$

where

$$X_h := \begin{pmatrix} D_x f_1\big(x_{h,1}, g(x+h)\big) \\ \vdots \\ D_x f_m\big(x_{h,m}, g(x+h)\big) \end{pmatrix}, \quad Y_h := \begin{pmatrix} D_y f_1\big(x, y_{h,1}\big) \\ \vdots \\ D_y f_m\big(x, y_{h,m}\big) \end{pmatrix}. \tag{C.53}$$

As we already know g to be continuous, $h \to 0$ implies $g(x+h) \to g(x)$. Thus, since $y_{h,l} \in S_{g(x),g(x+h)}$, $h \to 0$ implies $y_{h,l} \to g(x)$ for each $l \in \{1,\dots,m\}$, and, as all partials of f are continuous as well, $Y_h \to D_y f\big(x, g(x)\big)$. Since the maps $B \mapsto \det(B)$ and $B \mapsto \|B^{-1}\|$ are continuous (cf. Ex. 1.53 and Ex. 1.66(a),(b)), $h \to 0$ implies $\det(Y_h) \to \det\big(D_y f\big(x, g(x)\big)\big) \neq 0$ and Y_h is invertible for sufficiently small h with $(Y_h)^{-1} \to \big(D_y f\big(x, g(x)\big)\big)^{-1}$. For such sufficiently small h, we can rewrite (C.52) as

$$g(x+h) - g(x) = -(Y_h)^{-1} X_h h. \tag{C.54}$$

Also, since $x_{h,l} \in S_{x,x+h}$, $h \to 0$ implies $x_{h,l} \to x$ and, then, the continuity of g together with the continuity of the partials of f implies $X_h \to D_x f(x, g(x))$. Thus, we can finish the proof of (C.41) by noting

$$\begin{aligned} &\lim_{h\to 0} \frac{\big\| g(x+h) - g(x) + \big(D_y f(x, g(x))\big)^{-1} D_x f(x, g(x))\, h \big\|_1}{\|h\|} \\ &= \lim_{h\to 0} \frac{\big\| -(Y_h)^{-1} X_h h + \big(D_y f(x, g(x))\big)^{-1} D_x f(x, g(x))\, h \big\|_1}{\|h\|} = 0. \end{aligned} \tag{C.55}$$

It remains to prove that g is C^α if f is C^α, $\alpha \in \mathbb{N} \cup \{\infty\}$. To this end, for $\alpha \in \mathbb{N}$, we will show by induction on $\beta = 1, \dots, \alpha$ that each partial derivative of g at $x \in U_\xi$ of order β is a rational function of partials of f of order $\leq \beta$, all taken at $(x, g(x))$, and of partials of g of order $\leq \beta - 1$, all taken at x (in particular, the denominator of this rational function does not have any zeros in U_ξ): For $\beta = 1$, the claim follows from (C.41): The entries of $D_x f(x, g(x))$ are polynomials of first partials of f taken at $(x, g(x))$; the entries of $\big(D_y f(x, g(x))\big)^{-1}$ are, according to Th. A.51(c), rational functions, where both the numerator and the denominator polynomial are polynomials of first partials of f taken at $(x, g(x))$ (in particular, the entries of the right-hand side of (C.41) do not involve any first partials of g). For the induction step, let $1 < \beta \leq \alpha$. By induction, we know the partials of g of order $\beta - 1$ are rational functions of partials of f of order $\leq \beta - 1$, all taken at $(x, g(x))$, and of partials of g of order $\leq \beta - 2$, all taken at x. Taking the derivative of partials of g of order $\leq \beta - 2$ evaluated at x, yields partials of g of order $\leq \beta - 1$ still evaluated at x; according to the chain rule of Th. 2.28, taking the derivative of partials of f of order $\leq \beta - 1$ evaluated at $(x, g(x))$, yields polynomials of partials of f of order $\leq \beta$ evaluated at $(x, g(x))$ and of first partials of g evaluated at x. Thus, applying the product and the quotient rule establishes the case. ■

Theorem C.8 (Inverse Function Theorem)**.** *Let $n \in \mathbb{N}$, let $G \subseteq \mathbb{R}^n$ be open, and let $f : G \longrightarrow \mathbb{R}^n$ be continuously differentiable, i.e. $f \in C^1(G, \mathbb{R}^n)$. If $\xi \in G$ is such that*

$$Df(\xi) \text{ is invertible}, \tag{C.56}$$

then there exists an open neighborhood $U \subseteq G$ of ξ such that $V := f(U)$ is open and the restriction $f : U \longrightarrow V$ is bijective with continuously differentiable inverse function $f^{-1} : V \longrightarrow U$. Moreover,

$$\forall_{y \in V} \quad D(f^{-1})(y) = \Big(Df\big(f^{-1}(y)\big)\Big)^{-1} \tag{C.57}$$

and, if $f \in C^\alpha(U, \mathbb{R}^n)$, $\alpha \in \mathbb{N} \cup \{\infty\}$, then $f^{-1} \in C^\alpha(V, \mathbb{R}^n)$.

Proof. The idea is to apply the implicit function Th. C.7 to the continuously differentiable map

$$F : G \times \mathbb{R}^n \longrightarrow \mathbb{R}^n, \quad F(x,y) := f(x) - y. \tag{C.58}$$

Here, as compared to Th. C.7, the roles of the variables x and y are switched. Letting $\eta := f(\xi)$, we have

$$F(\xi,\eta) = f(\xi) - \eta = 0, \quad \text{and} \quad D_x F(\xi,\eta) = Df(\xi) \text{ is invertible.} \tag{C.59}$$

Thus, Th. C.7 applies and yields an open neighborhood $\tilde{U} \subseteq G$ of ξ, an open neighborhood $V \subseteq \mathbb{R}^n$ of η, and a C^1 map $g : V \longrightarrow \tilde{U}$ such that

$$\underset{(x,y)\in\tilde{U}\times V}{\forall} \quad \Big(F(x,y) = f(x) - y = 0 \quad \Leftrightarrow \quad x = g(y)\Big). \tag{C.60}$$

If we let $U := g(V)$, then $U \subseteq \tilde{U}$ is a neighborhood of $\xi = g(f(\xi))$, and (C.60) implies that $f : U \longrightarrow V$ and $g : V \longrightarrow U$ are inverse to each other, in particular, they are both bijective with $f^{-1} = g$. To verify that U is open, consider the (still continuous) map $f : \tilde{U} \longrightarrow \mathbb{R}^n$ and observe $U = f^{-1}(V)$. As V is open, Th. 1.54(ii) implies the existence of $O \subseteq \mathbb{R}^n$ open with $U = O \cap \tilde{U}$. Since both O and $\tilde{U}$ are open, U must be open as well.

Using (C.41), we obtain, for each $y \in V$,

$$\begin{aligned} Dg(y) &= -\big(D_x F(g(y),y)\big)^{-1} D_y F(g(y),y) \\ &= -\Big(Df\big(g(y)\big)\Big)^{-1} (-\operatorname{Id}) = \Big(Df\big(g(y)\big)\Big)^{-1}, \end{aligned} \tag{C.61}$$

proving (C.57). Finally, if f is C^α on U, then F is C^α on $U \times \mathbb{R}^n$, such that Th. C.7 implies $g = f^{-1}$ to be C^α as well. ∎

Corollary C.9. *Let $n \in \mathbb{N}$, let $G \subseteq \mathbb{R}^n$ be open, and let $f : G \longrightarrow \mathbb{R}^n$ be continuously differentiable, i.e. $f \in C^1(G,\mathbb{R}^n)$. If $Df(x)$ is invertible for each $x \in G$, then f maps open sets to open sets, i.e. if $O \subseteq G$ is open, then $f(O)$ is open as well.*

Proof. Let $O \subseteq G$ be open. We have to show that each point $\eta \in f(O)$ is an interior point of $f(O)$. To this end, let $\eta \in f(O)$ and let $\xi \in O$ be such that $f(\xi) = \eta$. Since $Df(\xi)$ is invertible by hypothesis, we can apply the inverse function Th. C.8 to the restriction of f to O, obtaining open neighborhoods $U \subseteq O$ of ξ and $V \subseteq f(O)$ of η such that $f : U \longrightarrow V$ is bijective. In particular, η is an interior point of $f(O)$, proving $f(O)$ to be open. ∎

D Riemann Integral for $\mathbb{C}$-Valued Functions

D.1 Riemann Integrability

Notation D.1. Let $I := [a,b] \subseteq \mathbb{R}^n$ be an interval, $a, b \in \mathbb{R}^n$, $a < b$. By $\mathcal{R}(I,\mathbb{R}) := \mathcal{R}(I)$ we denote the set of all Riemann integrable functions $f : I \longrightarrow \mathbb{R}$ (cf. Def. 4.5(b)).

Definition D.2. Let $I := [a,b] \subseteq \mathbb{R}^n$ be an interval, $a,b \in \mathbb{R}^n$, $a < b$. We call a function $f : I \longrightarrow \mathbb{C}$ *Riemann integrable* if, and only if, both $\operatorname{Re} f$ and $\operatorname{Im} f$ are Riemann integrable. The set of all Riemann integrable functions $f : I \longrightarrow \mathbb{C}$ is denoted by $\mathcal{R}(I,\mathbb{C})$. If $f \in \mathcal{R}(I,\mathbb{C})$, then

$$\int_I f := \left(\int_I \operatorname{Re} f, \int_I \operatorname{Im} f \right) = \int_I \operatorname{Re} f + i \int_I \operatorname{Im} f \in \mathbb{C} \tag{D.1}$$

is called the Riemann integral of f over I.

Theorem D.3. *Let $I := [a,b] \subseteq \mathbb{R}^n$, $a,b \in \mathbb{R}^n$, $n \in \mathbb{N}$, $a < b$, $f : I \longrightarrow \mathbb{C}$. If f is continuous, then f is Riemann integrable over I.*

Proof. If f is continuous, then $\operatorname{Re} f$ and $\operatorname{Im} f$ are both continuous, and, thus, the statement follows from the real-valued case of Th. 4.14. ■

Theorem D.4. *Let $I := [a,b] \subseteq \mathbb{R}^n$, $a,b \in \mathbb{R}^n$, $n \in \mathbb{N}$, $a < b$.*

(a) *If $f,g \in \mathcal{R}(I,\mathbb{C})$, then $\bar{f}, fg, \in \mathcal{R}(I,\mathbb{C})$. If, in addition, there exists $\delta > 0$ such that $|g(x)| \geq \delta$ for each $x \in I$, then $f/g \in \mathcal{R}(I,\mathbb{C})$.*

(b) *If $f \in \mathcal{R}(I,\mathbb{R})$ and $\phi : f(I) \longrightarrow \mathbb{C}$ is Lipschitz continuous, then $\phi \circ f \in \mathcal{R}(I,\mathbb{C})$.*

(c) *If $f \in \mathcal{R}(I,\mathbb{C})$ and $\phi : f(I) \longrightarrow \mathbb{R}$ is Lipschitz continuous, then $\phi \circ f \in \mathcal{R}(I,\mathbb{R})$.*

Proof. All the following proofs are completely analogous to the respective 1-dimensional case in [Phi15a, Th. G.4].

(a): Since

$$\bar{f} = (\operatorname{Re} f, -\operatorname{Im} f), \tag{D.2a}$$
$$fg = (\operatorname{Re} f \operatorname{Re} g - \operatorname{Im} f \operatorname{Im} g,\ \operatorname{Re} f \operatorname{Im} g + \operatorname{Im} f \operatorname{Re} g), \tag{D.2b}$$
$$1/g = (\operatorname{Re} g/|g|^2, -\operatorname{Im} g/|g|^2), \tag{D.2c}$$

everything follows from the real-valued case of Th. 4.12(a) and of Th. 4.15(b),(c), where $|g| \geq \delta > 0$ guarantees $|g|^2 \geq \delta^2 > 0$).

(b): Assume ϕ to be L-Lipschitz, $L \geq 0$. For each $x,y \in f(I)$, one has

$$|\operatorname{Re}\phi(x) - \operatorname{Re}\phi(y)| \overset{\text{[Phi15a, Th. 5.11(d)]}}{\leq} |\phi(x) - \phi(y)| \leq L|x-y|, \tag{D.3a}$$
$$|\operatorname{Im}\phi(x) - \operatorname{Im}\phi(y)| \overset{\text{[Phi15a, Th. 5.11(d)]}}{\leq} |\phi(x) - \phi(y)| \leq L|x-y|, \tag{D.3b}$$

showing $\operatorname{Re}\phi$ and $\operatorname{Im}\phi$ are L-Lipschitz, such that $\operatorname{Re}(\phi \circ f)$ and $\operatorname{Im}(\phi \circ f)$ are Riemann integrable by Th. 4.15(a).

(c): Assume ϕ to be L-Lipschitz, $L \geq 0$. If $f \in \mathcal{R}(I,\mathbb{C})$, then $\operatorname{Re} f, \operatorname{Im} f \in \mathcal{R}(I,\mathbb{R})$, and, given $\epsilon > 0$, Riemann's integrability criterion of Th. 4.13 provides partitions Δ_1, Δ_2 of I such that $R(\Delta_1, \operatorname{Re} f) - r(\Delta_1, \operatorname{Re} f) < \epsilon/2L$, $R(\Delta_2, \operatorname{Im} f) - r(\Delta_2, \operatorname{Im} f) < \epsilon/2L$, where

R and r denote upper and lower Riemann sums, respectively (cf. (4.11)). Letting Δ be a joint refinement of Δ_1 and Δ_2, we have (cf. Def. 4.8(a),(b) and Th. 4.10(a))

$$R(\Delta, \operatorname{Re} f) - r(\Delta, \operatorname{Re} f) < \epsilon/2L, \quad R(\Delta, \operatorname{Im} f) - r(\Delta, \operatorname{Im} f) < \epsilon/2L. \tag{D.4}$$

Recalling that, for each $g : I \longrightarrow \mathbb{R}$, it is

$$r(\Delta, g) = \sum_{p \in P(\Delta)} m_p(g)|I_p|, \tag{D.5a}$$

$$R(\Delta, g) = \sum_{p \in P(\Delta)} M_p(g)|I_p|, \tag{D.5b}$$

where $P(\Delta)$ is according to Def. 4.2,

$$m_p(g) := \inf\{g(x) : x \in I_p\}, \quad M_p(g) := \sup\{g(x) : x \in I_p\}, \tag{D.5c}$$

we obtain, for each $\xi_p, \eta_p \in I_p$,

$$\begin{aligned}
&\big|(\phi \circ f)(\xi_p) - (\phi \circ f)(\eta_p)\big| \\
&\leq L\,\big|f(\xi_p) - f(\eta_p)\big| \overset{\text{[Phi15a, Th. 5.11(d)]}}{\leq} L\,\big|\operatorname{Re} f(\xi_p) - \operatorname{Re} f(\eta_p)\big| + L\,\big|\operatorname{Im} f(\xi_p) - \operatorname{Im} f(\eta_p)\big| \\
&\leq L\,\big(M_p(\operatorname{Re} f) - m_p(\operatorname{Re} f)\big) + L\,\big(M_p(\operatorname{Im} f) - m_p(\operatorname{Im} f)\big),
\end{aligned} \tag{D.6}$$

and, thus,

$$\begin{aligned}
&R(\Delta, \phi \circ f) - r(\Delta, \phi \circ f) = \sum_{p \in P(\Delta)} \big(M_p(\phi \circ f) - m_p(\phi \circ f)\big)|I_p| \\
&\overset{(D.6)}{\leq} \sum_{p \in P(\Delta)} L\,\big(M_p(\operatorname{Re} f) - m_p(\operatorname{Re} f)\big)|I_p| + \sum_{p \in P(\Delta)} L\,\big(M_p(\operatorname{Im} f) - m_p(\operatorname{Im} f)\big)|I_p| \\
&= L\,\big(R(\Delta, \operatorname{Re} f) - r(\Delta, \operatorname{Re} f)\big) + L\,\big(R(\Delta, \operatorname{Im} f) - r(\Delta, \operatorname{Im} f)\big) \overset{(D.4)}{<} \epsilon.
\end{aligned} \tag{D.7}$$

Thus, $\phi \circ f \in \mathcal{R}(I, \mathbb{R})$ by Th. 4.13. ■

Theorem D.5. *Let $n \in \mathbb{N}$, $a, b \in \mathbb{R}^n$, $a < b$, $I := [a, b]$.*

(a) *The integral is linear: More precisely, if $f, g \in \mathcal{R}(I, \mathbb{C})$ and $\lambda, \mu \in \mathbb{C}$, then $\lambda f + \mu g \in \mathcal{R}(I, \mathbb{C})$ and*

$$\int_I (\lambda f + \mu g) = \lambda \int_I f + \mu \int_I g. \tag{D.8}$$

(b) *For each $f \in \mathcal{R}(I, \mathbb{C})$, one has $|f| \in \mathcal{R}(I, \mathbb{R})$ and*

$$\left| \int_I f \right| \leq \int_I |f|. \tag{D.9}$$

Proof. (a): One computes, using the real-valued case of Th. 4.12(a),

$$\begin{aligned}\int_I(\lambda f) &= \left(\int_I(\operatorname{Re}\lambda\operatorname{Re}f-\operatorname{Im}\lambda\operatorname{Im}f),\ \int_I(\operatorname{Re}\lambda\operatorname{Im}f+\operatorname{Im}\lambda\operatorname{Re}f)\right)\\ &= \left(\operatorname{Re}\lambda\int_I\operatorname{Re}f-\operatorname{Im}\lambda\int_I\operatorname{Im}f,\ \operatorname{Re}\lambda\int_I\operatorname{Im}f+\operatorname{Im}\lambda\int_I\operatorname{Re}f\right)\\ &= \lambda\int_I f \end{aligned}\tag{D.10a}$$

and

$$\begin{aligned}\int_I(f+g) &= \left(\int_I\operatorname{Re}(f+g),\ \int_I\operatorname{Im}(f+g)\right)=\left(\int_I\operatorname{Re}f+\int_I\operatorname{Re}g,\ \int_I\operatorname{Im}g+\int_I\operatorname{Im}g\right)\\ &= \left(\int_I\operatorname{Re}f,\ \int_I\operatorname{Im}f\right)+\left(\int_I\operatorname{Re}g,\ \int_I\operatorname{Im}g\right)=\int_I f+\int_I g.\end{aligned}\tag{D.10b}$$

(b): As the modulus is 1-Lipschitz by the inverse triangle inequality, $|f|\in\mathcal{R}(I,\mathbb{R})$ by Th. D.4(c). Let Δ be an arbitrary partition of I. Then, using the notation from the proof of Th. D.4(c) above, we obtain the following estimate of intermediate Riemann sums (cf. (4.11c)):

$$\begin{aligned}\Big|\big(\rho(\Delta,\operatorname{Re}f),\,\rho(\Delta,\operatorname{Im}f)\big)\Big| &:= \left|\left(\sum_{p\in P(\Delta)}\operatorname{Re}f(\xi_p)\,|I_p|,\ \sum_{p\in P(\Delta)}\operatorname{Im}f(\xi_p)\,|I_p|\right)\right|\\ &\le \sum_{p\in P(\Delta)}\Big|\big(\operatorname{Re}f(\xi_p),\,\operatorname{Im}f(\xi_p)\big)\Big|\,|I_p|\\ &= \sum_{p\in P(\Delta)}|f(\xi_p)|\,|I_p|=\rho(\Delta,|f|).\end{aligned}\tag{D.11}$$

Since the intermediate Riemann sums in (D.11) converge to the respective integrals by (4.30b), one obtains

$$\left|\int_I f\right| = \lim_{|\Delta|\to0}\Big|\big(\rho(\Delta,\operatorname{Re}f),\,\rho(\Delta,\operatorname{Im}f)\big)\Big| \overset{\text{(D.11)}}{\le} \lim_{|\Delta|\to0}\rho(\Delta,|f|)=\int_I|f|,\tag{D.12}$$

proving (D.9). ∎

D.2 Fubini Theorem

Definition D.6. Let $I=[a,b]\subseteq\mathbb{R}^n$ be an interval, $a,b\in\mathbb{R}^n$, $n\in\mathbb{N}$, $a<b$, and suppose $f:I\longrightarrow\mathbb{C}$ is bounded (i.e. both $\operatorname{Re}f$ and $\operatorname{Im}f$ are bounded). Define

$$J_*(f,I) := J_*(\operatorname{Re}f,I)+i\,J_*(\operatorname{Im}f,I),\tag{D.13a}$$

$$J^*(f,I) := J^*(\operatorname{Re}f,I)+i\,J^*(\operatorname{Im}f,I).\tag{D.13b}$$

As in the $\mathbb{R}$-valued case, we call $J_*(f,I)$ the *lower Riemann integral* of f over I and $J^*(f,I)$ the *upper Riemann integral* of f over I.

Theorem D.7. *Let $a, b, c, d, e, f \in \mathbb{R}^n$, $n \in \mathbb{N}$, $a < b$, $c < d$, $e < f$, $I = [a, b]$, $J = [c, d]$, $K = [e, f]$. If $I = J \times K$ and $f \in \mathcal{R}(I, \mathbb{C})$, then*

$$\int_I f = \int_I f(x, y)\, \mathrm{d}(x, y) = \int_K \int_J f(x, y)\, \mathrm{d}x\, \mathrm{d}y = \int_J \int_K f(x, y)\, \mathrm{d}y\, \mathrm{d}x\,. \tag{D.14}$$

As in the real-valued Th. 4.16*, there is a slight abuse of notation in* (D.14)*, as it can happen that a function $x \mapsto f(x, y)$ is not Riemann integrable over J and that a function $y \mapsto f(x, y)$ is not Riemann integrable over K. As in Th.* 4.16*, in that case, one can choose either the lower or the upper Riemann integral for the inner integrals in* (D.14)*. Independently of the choice, the resulting function $y \mapsto \int_J f(x, y)\, \mathrm{d}x$ is Riemann integrable over K, $x \mapsto \int_K f(x, y)\, \mathrm{d}y$ is Riemann integrable over J, and the validity of* (D.14) *is unaffected. By applying* (D.14) *inductively, one obtains*

$$\int_I f = \int_I f(x)\, \mathrm{d}x = \int_{a_1}^{b_1} \dots \int_{a_n}^{b_n} f(x_1, \dots, x_n)\, \mathrm{d}x_n \cdots \mathrm{d}x_1, \tag{D.15}$$

where, for the inner integrals, one can choose the upper or lower Riemann integral, and one can also permute their order without changing the overall value.

Proof. We show how to obtain the present $\mathbb{C}$-valued case from the $\mathbb{R}$-valued case of Th. 4.16. One computes, using lower Riemann integrals for the inner integrals,

$$\begin{aligned} \int_I f &= \int_I \operatorname{Re} f + i \int_I \operatorname{Im} f \\ &\overset{\text{Th. 4.16}}{=} \int_K J_*\big(\operatorname{Re} f(\cdot, y), J\big)\, \mathrm{d}y + i \int_K J_*\big(\operatorname{Im} f(\cdot, y), J\big)\, \mathrm{d}y \\ &\overset{\text{(D.13a)}}{=} \int_K J_*\big(f(\cdot, y), J\big)\, \mathrm{d}y\,, \end{aligned} \tag{D.16}$$

proving $\int_I f = \int_K \int_J f(x, y)\, \mathrm{d}x\, \mathrm{d}y$ with the inner integral interpreted as lower Riemann integral. Clearly, the same calculation works if the inner integral is interpreted as upper Riemann integral, and it also still works if J and K are switched, completing the proof of (D.14). As mentioned in the statement, (D.15) follows from (D.14) by induction. ■

D.3 Change of Variables

Theorem D.8. *Let $a, b, c, d \in \mathbb{R}^n$, $n \in \mathbb{N}$, $a < b$, $c < d$, $I := [a, b]$, $J := [c, d]$, $\phi : I \longrightarrow \mathbb{R}^n$, $f : J \longrightarrow \mathbb{C}$. If, on the interior of I, ϕ is one-to-one, Lipschitz continuous, and has continuous first partials, $\phi(I) \subseteq J$, and $(f \circ \phi)|\det J_\phi| \in \mathcal{R}(I)$, then $f\chi_{\phi(I)} \in \mathcal{R}(J)$ and the following* change of variables *formula holds:*

$$\int_J f\chi_{\phi(I)} = \int_I (f \circ \phi)|\det J_\phi|. \tag{D.17}$$

Proof. We show how to obtain the present $\mathbb{C}$-valued case from the $\mathbb{R}$-valued case of Th. 4.18. One computes

$$\begin{aligned}\int_J f\chi_{\phi(I)} &= \int_J \operatorname{Re} f\chi_{\phi(I)} + i\int_J \operatorname{Im} f\chi_{\phi(I)} \\ &\overset{(4.61)}{=} \int_I \big((\operatorname{Re} f)\circ\phi\big)|\det J_\phi| + i\int_I \big((\operatorname{Im} f)\circ\phi\big)|\det J_\phi| \\ &= \int_I (f\circ\phi)|\det J_\phi|, \end{aligned} \tag{D.18}$$

thereby establishing the case. ■

References

[Phi14] P. Philip. *Ordinary Differential Equations*. Lecture Notes, Ludwig-Maximilians-Universität, Germany, 2014, available in PDF format at `http://www.math.lmu.de/~philip/publications/lectureNotes/ODE.pdf`.

[Phi15a] P. Philip. *Calculus I for Computer Science and Statistics Students*. Lecture Notes, Ludwig-Maximilians-Universität, Germany, 2014/2015, available in PDF format at `http://www.math.lmu.de/~philip/publications/lectureNotes/calc1_forInfAndStatStudents.pdf`.

[Phi15b] P. Philip. *Numerical Analysis I*. Lecture Notes, Ludwig-Maximilians-Universität, Germany, 2014/2015, available in PDF format at `http://www.math.lmu.de/~philip/publications/lectureNotes/numericalAnalysis.pdf`.

[Str08] Gernot Stroth. *Lineare Algebra*, 2nd ed. Berliner Studienreihe zur Mathematik, Vol. 7, Heldermann Verlag, Lemgo, Germany, 2008 (German).

[Wal02] Wolfgang Walter. *Analysis 2*, 5th ed. Springer-Verlag, Berlin, 2002 (German).