

EIDGENÖSSISCHE TECHNISCHE HOCHSCHULE ZÜRICH

---

## Analysis II: Several Variables

---

Lecture Notes — Spring Semester 2024

JOAQUIM SERRA

April 27, 2024

## Preface

This notes are the continuation of 'Analysis I: One variable' and follow the same format and general spirit.

The text is partially based in some notes originally crafted in German for the academic year 2016/2017 by Manfred Einsiedler and Andreas Wieser, they were designed for the Analysis I and II courses in the Interdisciplinary Natural Sciences, Physics, and Mathematics Bachelor programs. In the academic year 2019/2020, a substantial revision was undertaken by Peter Jossen.

For the academic year 2023/2024, Joaquim Serra wrote this English version. Although some chapter build on German original, with changes, the new version is completely different in several aspects: some chapters (such as the measure and integral, or most of global integral theorem) have been completely rewritten. Also the parts that have been have been reorganized: alternative proofs of some materials, rewriting and expansion in certain areas, and a more concise presentation in others. This version strictly aligns with the material presented in class, offering a streamlined educational experience.

The courses Analysis I/II and Linear Algebra I/II are fundamental to the mathematics curriculum at ETH and other universities worldwide. They lay the groundwork upon which most future studies in mathematics and physics are built.

Throughout Analysis I/II, we will delve into various aspects of differential and integral calculus. Although some topics might be familiar from high school, our approach requires minimal prior knowledge beyond an intuitive understanding of variables and basic algebraic skills. Contrary to high-school methods, our lectures emphasize the development of mathematical theory over algorithmic practice. Understanding and exploring topics such as differential equations and multidimensional integral theorems is our primary goal. However, students are encouraged to engage with numerous exercises from these notes and other resources to deepen their understanding and proficiency in these new mathematical concepts.

---

# Contents

<b>9</b>	<b>Metric spaces</b>	<b>2</b>
9.1	BASICS OF METRIC SPACES . . . . .	3
9.1.1	The Euclidean space $\mathbb{R}^n$ . . . . .	3
9.1.2	Definition of metric Space . . . . .	5
9.1.3	Sequences, limits, and completeness . . . . .	7
9.1.4	*The Reals as the Completion of Rationals (extra material; cf. Grundstrukturen) . . . . .	12
9.2	TOPOLOGY OF METRIC SPACES . . . . .	14
9.2.1	Open and closed sets . . . . .	14
9.2.2	Continuity . . . . .	18
9.2.3	Banach's Fixed-Point Theorem . . . . .	19
9.2.4	Compactness . . . . .	21
9.2.5	Compactness and continuity . . . . .	26
9.2.6	Connectedness . . . . .	28
9.3	NORMED VECTOR SPACES . . . . .	31
9.3.1	Definition of Normed Vector spaces . . . . .	31
9.3.2	Inner Product Spaces are Normed Vector Spaces . . . . .	33
9.3.3	Equivalence of norms in finite dimensional normed spaces . . . . .	35
<b>10</b>	<b>Multidimensional Differentiation</b>	<b>38</b>
10.1	THE DIFFERENTIAL . . . . .	39
10.1.1	The derivative for real valued functions . . . . .	39
10.1.2	Definition of differential . . . . .	40
10.1.3	The Chain Rule . . . . .	46
10.1.4	The Mean Value Theorem . . . . .	49
10.2	HIGHER DERIVATIVES . . . . .	52
10.2.1	Definition and basic properties . . . . .	52
10.2.2	Schwartz's Theorem and Multi-indexes notation . . . . .	53
10.2.3	Multidimensional Taylor Approximation . . . . .	57
10.2.4	Real analytic functions of several variables . . . . .	59
<b>11</b>	<b>Optimization and applications, Convexity</b>	<b>62</b>
11.1	FIRST ORDER OPTIMALITY CONDITION FOR INTERIOR POINTS . . . . .	63

---

11.2	FIRST ORDER OPTIMALITY CONDITION WITH CONSTRAINTS: LAGRANGE MULTIPLIERS . . . . .	64
11.3	APPLICATION: SPECTRAL THEOREM FOR SYMMETRIC MATRICES . . . . .	66
11.4	SECOND ORDER OPTIMALITY CONDITIONS . . . . .	69
11.5	THE FUNDAMENTAL THEOREM OF ALGEBRA VIA MINIMIZATION . . . . .	72
11.6	CONVEXITY . . . . .	74
<b>12</b>	<b>Inverse and implicit function theorems and submanifolds</b>	<b>76</b>
12.1	THE INVERSE FUNCTION THEOREM . . . . .	76
12.1.1	Small Lipschitz perturbations of the identity . . . . .	76
12.1.2	Inverse of differentiable maps . . . . .	78
12.1.3	Inverse and Implicit function theorems . . . . .	79
12.2	SUBMANIFOLDS OF $\mathbb{R}^n$ . . . . .	82
12.2.1	Definition of submanifold and different representations . . . . .	82
12.2.2	Some examples of submanifolds of $\mathbb{R}^3$ . . . . .	84
<b>13</b>	<b>Multidimensional Integration</b>	<b>88</b>
13.1	THE JORDAN MEASURE IN $\mathbb{R}^n$ . . . . .	88
13.2	RIEMANN INTEGRAL, CHANGE OF VARIABLES FORMULA . . . . .	101
13.3	FUBINI THEOREM, DIFFERENTIATION UNDER THE INTEGRAL SIGN . . . . .	107
13.4	EXAMPLES: CHANGE OF VARIABLES AND FUBINI IN PRACTICE . . . . .	112
13.5	IMPROPER INTEGRALS: DEFINITION AND EXAMPLES . . . . .	115
13.6	LENGTH, AREA, AND INTEGRALS OVER SUBMANIFOLDS . . . . .	116
13.7	FROM LOCAL TO GLOBAL: PARTITION OF UNITY ON COMPACT SUBMANIFOLDS	128
<b>14</b>	<b>Global integral theorems</b>	<b>134</b>
14.1	THE INTEGRATION BY PARTS FORMULA . . . . .	134
14.1.1	Bounded, smooth domains, exterior unit normal . . . . .	134
14.1.2	Integration by parts formula: special cases . . . . .	137
14.1.3	The integration by parts formula and the divergence theorem . . . . .	141
14.1.4	Examples and application of the divergence theorem . . . . .	144
14.1.5	Rotation and the Green's Theorem . . . . .	145
14.2	LINE INTEGRALS AND THE POINCARÉ LEMMA . . . . .	148
14.2.1	The work of a Vector Field along a path . . . . .	148
14.2.2	The Poincaré Lemma . . . . .	152

---

## Chapter 9

# Metric spaces

In Analysis I, we focused primarily on functions that operate between real numbers,  $\mathbb{R}$  to  $\mathbb{R}$ . Consequently, we deeply explored the real number line,  $\mathbb{R}$ , and its properties.

Now, in Analysis II, we expand our scope to functions that map from  $\mathbb{R}^n$  to  $\mathbb{R}^m$ , where  $n$  and  $m$  are positive integers. In this context, we start by delving into the properties of  $\mathbb{R}^n$  (or  $\mathbb{R}^m$ , as we can use them interchangeably since  $n$  and  $m$  are arbitrary).

We'll discover that  $\mathbb{R}^n$  follows the axioms of a metric space when equipped with the standard Euclidean distance, which we'll define later. This means that  $\mathbb{R}^n$  is a metric space. Understanding this concept is vital because it lays the foundation for many fundamental definitions and results we want to establish for  $\mathbb{R}^n$  and can be extended to a broader range of metric spaces.

Additionally, this will allow us to revisit (and review) the crucial concept of convergence, which we have seen in Analysis I for  $\mathbb{R}$ , within the broader context of metric spaces. We'll emphasize the essential properties of metric spaces and their topology, particularly focusing on compactness.

Furthermore, we'll introduce and explore standard results related to normed vector spaces. As we will see, these spaces are more general than Euclidean space  $\mathbb{R}^n$ , but they possess a higher degree of structure (i.e., more axioms) than metric spaces.

## 9.1 Basics of Metric Spaces

In this class, our primary focus will be on  $\mathbb{R}^n$ , but we'll find that certain definitions and proofs become clearer when viewed within the broader context of metric spaces, which have fewer axioms.

However, when we're dealing with a general metric space  $X$ , it often helps to initially visualize it as our familiar 2-dimensional or 3-dimensional spaces,  $\mathbb{R}^2$  or  $\mathbb{R}^3$ . This approach can simplify the intuitive understanding of various arguments. If an argument relies on only a limited set of fundamental properties of  $\mathbb{R}^3$  (excluding aspects like angles and vector space structure), it may be applicable to general metric spaces as well.

### 9.1.1 The Euclidean space $\mathbb{R}^n$

For some integer  $n \geq 1$ , we denote by  $\mathbb{R}^n$  the set of all ordered  $n$ -tuples of real numbers. A general element  $x \in \mathbb{R}^n$  is thus of the form  $x = (x_1, \dots, x_n)$ , where the  $x_i$ 's are real numbers. (Even more rigorously  $\mathbb{R}^n := \{x: \{1, \dots, n\} \rightarrow \mathbb{R}\}$ .)

$\mathbb{R}^n$  is a vector space over the field of real numbers with the coordinate-wise addition and multiplication by a scalar. Rigorously, given  $x, y \in \mathbb{R}^n$  and  $\lambda \in \mathbb{R}$  we have

$$x + y := (x_1 + y_1, \dots, x_n + y_n), \quad \lambda x := (\lambda x_1, \dots, \lambda x_n).$$

#### DEFINITION 9.1: EUCLIDEAN STRUCTURE OF $\mathbb{R}^n$

Given  $x, y \in \mathbb{R}^n$  we define the standard **scalar product** of  $x$  and  $y$  as

$$x \cdot y = \langle x, y \rangle := \sum_{1 \leq i \leq n} x_i y_i,$$

the **Euclidean norm** of  $x$  as

$$\|x\| := \sqrt{\sum_{1 \leq i \leq n} x_i^2},$$

and the **Euclidean distance** of  $x$  and  $y$  as

$$d(x, y) := \|x - y\| = \sqrt{\sum_{1 \leq i \leq n} (x_i - y_i)^2}.$$

We key property of the Euclidean distance that we want to abstract is

PROPOSITION 9.2: TRIANGLE INEQUALITY IN  $\mathbb{R}^n$ 

For all  $x, y, z \in \mathbb{R}^n$

$$\|x - z\| \leq \|x - y\| + \|y - z\|.$$

Equivalently, for all  $x, y \in \mathbb{R}^n$ ,  $\|x + y\| \leq \|x\| + \|y\|$ .

*Proof.* To prove the equivalence of the two statements consider  $a = x - y$  and  $b = y - z$  so that  $a + b = x - z$ .

We will see later a more general proof that applies to all inner product spaces (see Corollary 9.101), based on the Cauchy-Schwarz inequality (see Proposition 9.100). Let us give here a hands-on argument.

Pick two points in  $\mathbb{R}^n$ ,  $x = (x_1, \dots, x_n)$  and  $y = (y_1, \dots, y_n)$  we would like to show

$$\left( \sum_{i=1}^n (x_i + y_i)^2 \right)^{1/2} \leq \left( \sum_{i=1}^n x_i^2 \right)^{1/2} + \left( \sum_{i=1}^n y_i^2 \right)^{1/2},$$

Taking squares, this is equivalent to

$$\sum_{i=1}^n x_i^2 + 2x_i y_i + y_i^2 \leq \sum_{i=1}^n x_i^2 + 2 \left( \sum_{i=1}^n x_i^2 \right)^{1/2} \left( \sum_{i=1}^n y_i^2 \right)^{1/2} + \sum_{i=1}^n y_i^2,$$

and, canceling terms, to

$$\sum_{i=1}^n x_i y_i \leq \left( \sum_{i=1}^n x_i^2 \right)^{1/2} \left( \sum_{i=1}^n y_i^2 \right)^{1/2}. \quad (9.1)$$

Therefore, the Proposition will follow if we can establish the validity of (9.1), or (squaring it) of:

$$\left( \sum_{i=1}^n x_i y_i \right) \left( \sum_{j=1}^n x_j y_j \right) = \sum_{i,j=1}^n x_i x_j y_i y_j \leq \sum_{i,j=1}^n x_i^2 y_j^2.$$

But this last inequality is easily established summing over all pairs  $i, j \in \{1, \dots, n\}$  the inequalities

$$2x_i x_j y_i y_j \leq x_i^2 y_j^2 + x_j^2 y_i^2 \quad \Leftrightarrow \quad (x_i y_j - x_j y_i)^2 \geq 0;$$

and observing that  $\sum_{i,j=1}^n x_i^2 y_j^2 = \sum_{i,j=1}^n x_j^2 y_i^2$ .  $\square$

### 9.1.2 Definition of metric Space

#### DEFINITION 9.3: METRIC SPACE

A **metric space**  $(X, d)$  is a nonempty set  $X$  together with a nonnegative function  $d: X \times X \rightarrow [0, \infty)$ , called the **distance** (or **metric**) on  $X$ , which satisfies:

- (1) For all  $x, y \in X$ ,  $d(x, y) = 0$  if and only if  $x = y$  (Definiteness).
- (2) For all  $x, y \in X$ ,  $d(x, y) = d(y, x)$  (Symmetry).
- (3) For all  $x, y, z \in X$ ,  $d(x, z) \leq d(x, y) + d(y, z)$  (Triangle inequality).

9.4. — A metric  $d$  on a set  $X$  assigns to each pair of points their **distance**. In this interpretation, the definiteness condition states that the only point at zero distance from a given point  $x \in X$  is  $x$  itself. The symmetry condition states that the distance from  $x \in X$  to  $y \in X$  is the same as from  $y$  to  $x$ . Interpreting the distance between two points as the length of a shortest path from one point to the other, the triangle inequality states that the length of a shortest path from  $x$  to  $y$  is at most the length of a path one takes by first going to  $y$  and then from there to  $z$ .

EXERCISE 9.5. — Prove using induction over the number of points that  $N \geq 3$  that if  $d$  satisfies the triangle inequality then:

$$d(x_1, x_N) \leq \sum_{i=1}^{N-1} d(x_i, x_{i+1}).$$

9.6. — When there is no possible confusion, we will often say “Let  $X$  be a metric space...”, leaving the distance function unspecified. This is a shorter version of the more precise sentence “Let  $(X, d)$  be a metric space...”.

Furthermore, we may refer to the set  $X$  as a **space** and the elements of  $X$  as **points**. This is because we have in mind that  $X$  is some sort of geometric space, like a subset of the plane or the surface of a sphere. In this setting, “spaces” and “points” will be synonymous with “sets” and “elements”.

9.7. — Notice that the Euclidean space  $(\mathbb{R}^n, d)$ , with  $d(x, y) := \|x - y\|$ , is a metric space. In particular  $\mathbb{R}$ , equipped with the absolute value distance  $|x - y|$  is a metric space.

EXERCISE 9.8. — Let  $(X, d)$  be a metric space and let  $\phi: [0, \infty) \rightarrow [0, \infty)$  a function which is concave, increasing,  $\phi(0) = 0$ , and not identically zero. Show that  $(X, \phi \circ d)$  is again a metric space. For example one can take  $\phi(t) = \sqrt{t}$ ,  $\phi(t) = \arctan t$  or  $\phi(t) = \frac{t}{1+t}$ . Notice that the last two choices always give bounded distances.



EXAMPLE 9.9. — When  $X \subset \mathbb{R}$ , we can take the **standard metric**  $d$  defined by

$$d(x, y) = |x - y| \text{ for all } x, y \in X.$$

EXAMPLE 9.10. — Let  $X$  be a set, and  $d : X \times X \rightarrow \mathbb{R}$  defined by

$$d(x, y) = \begin{cases} 1 & \text{if } x \neq y \\ 0 & \text{if } x = y \end{cases}$$

for  $x, y \in X$ . Then,  $(X, d)$  is a metric space. Indeed,  $d$  is definite and symmetric by definition. Furthermore,  $d$  satisfies the triangle inequality: Let  $x, y, z$  be points in  $X$ . If  $d(x, z) = 0$ , then  $d(x, z) \leq d(x, y) + d(y, z)$  is trivially satisfied. If  $d(x, z) = 1$ , then  $x \neq z$ , and  $y$  is at least different from one point in  $\{x, z\}$ , so the triangle inequality also holds. This metric  $d$  is called the **discrete metric** on the set  $X$ .

EXAMPLE 9.11. — Let  $X = \mathbb{R}^2$ , and define the **Manhattan metric** on  $X$  by

$$d_{\text{NY}}(x, y) = |x_1 - y_1| + |x_2 - y_2|$$

where we put  $x = (x_1, x_2)$  and  $y = (y_1, y_2)$ . It can be verified (exercise) that  $d_{\text{NY}}$  satisfies all axioms of a metric. The reason why  $d_{\text{NY}}$  is called the Manhattan metric is that in grid-like places such as Manhattan, one can reach from  $(x_1, x_2)$  to  $(y_1, y_2)$  in the following way: first move ‘horizontally’ (i.e., with constant second coordinate) from  $x = (x_1, x_2)$  to  $(y_1, x_2)$  and then ‘vertically’ (with constant first coordinate) from  $(y_1, x_2)$  to  $y = (y_1, y_2)$ , or vice versa:  $x = (x_1, x_2)$  to  $(x_1, y_2)$ , and then to  $y = (y_1, y_2)$ . Since all streets in Manhattan run either from west to east or from north to south,  $d_{\text{NY}}$  measures the relevant distance between two points.

EXERCISE 9.12. — Let  $X$  be the set of all continuous real-valued functions defined on  $[0, 1] \subset \mathbb{R}$ . For  $f, g \in X$  set

$$d_1(f, g) := \max\{|f(x) - g(x)| \mid x \in [0, 1]\} \quad \text{and} \quad d_2(f, g) := \int_0^1 |f(x) - g(x)| dx.$$

Show that  $(X, d_1)$  and  $(X, d_2)$  are metric spaces.

EXAMPLE 9.13. — If  $(X, d)$  is a metric space and  $X_0 \subset X$  is some subset then  $X_0$  inherits an structure of metric space from  $X$ . Indeed, one can easily verify that  $(X_0, d_0)$ , where  $d_0$  the restriction of  $d$  to  $X_0 \times X_0 \subset X \times X$  is a metric space.

For a more concrete instance of this take  $X = \mathbb{R}^3$  with the Euclidean distance  $d$  and let  $X_0$  be the sphere  $\{x \in \mathbb{R}^3 \mid x_1^2 + x_2^2 + x_3^2 = R^2\}$ , for some  $R > 0$ . Then for any pair of points

$x, y$  in the sphere we have

$$\begin{aligned} d_0(x, y) &= \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + (x_3 - y_3)^2} = \sqrt{(x - y) \cdot (x - y)} \\ &= \sqrt{x \cdot x + y \cdot y - 2x \cdot y} = \sqrt{2R^2 - 2x \cdot y} \end{aligned}$$

An arguably more natural metric  $d_1$  on the sphere  $X_0$  can be define measuring the length of the geodesic arc joining  $x$  and  $y$ . One can see that this metric is given by:

$$d_1(x, y) = R \arccos \left( \frac{x \cdot y}{R^2} \right) \in [0, \pi R].$$

### 9.1.3 Sequences, limits, and completeness

The definition of sequence in a set was given in Analysis I. We recall it next:

#### DEFINITION 9.14: SEQUENCES IN A SET

Let  $X$  be a set, a **sequence** in  $X$  is a function  $x : \mathbb{N} \rightarrow X$ . The image  $x(n)$  of  $n \in \mathbb{N}$  is also denoted as  $x_n$  and referred to as the  $n$ -th **term** of  $x$ .

Instead of  $x : \mathbb{N} \rightarrow X$ , one often writes  $(x_n)_{n \in \mathbb{N}}$ ,  $(x_n)_{n \geq 0}$ , or  $(x_n)_{n=0}^\infty$ .

We introduce the following vocabulary, which is useful if used precisely.

#### DEFINITION 9.15: "EVENTUALLY"

Let  $(x_n)_{n \geq 0} \subset X$  be a sequence and let  $\mathcal{P} : X \rightarrow \{\text{true}, \text{false}\}$  be a property that an element in  $X$  might have or not. Then one says that " $x_n$  satisfies  $\mathcal{P}$  eventually" there exist  $N \in \mathbb{N}$  such that  $\mathcal{P}(x_n)$  is true for all  $n \geq N$ . In other words, if  $\mathcal{P}(x_n)$  holds true along the sequence with only finitely many exceptions.

#### DEFINITION 9.16: CONVERGENT SEQUENCE

Let  $(X, d)$  be a metric space,  $x \in X$  and  $(x_n)_{n \in \mathbb{N}}$  be a sequence in  $X$ . We say that  $(x_n)_{n \in \mathbb{N}}$  **converges** to  $x$ , or that  $x$  is the **limit** of the sequence  $(x_n)_{n \in \mathbb{N}}$ , if

$$\lim_n d(x_n, x) = 0.$$

In other words, for any  $\varepsilon > 0$  eventually  $d(x_n, x) < \varepsilon$ .

9.17. — When the metric space  $(X, d)$  is clear from the context, we may write  $\lim_n x_n = x$  or even  $x_n \rightarrow x$  to express that  $(x_n)_{n \in \mathbb{N}}$  converges to  $x$ .

EXERCISE 9.18. — In the setting of Exercise 9.8, show that a sequence converges in  $(X, d)$  if and only if it converges in  $(X, \phi \circ d)$ .

## LEMMA 9.19: UNIQUENESS OF THE LIMIT

*In a metric space, a convergent sequence has only one limit.*

*Proof.* Let  $(X, d)$  be a metric space and let  $A, B \in X$  be limits of some sequence  $(x_n)_{n=0}^\infty$ , we mean to show that  $A = B$ . Take  $\varepsilon > 0$ , then, we can find  $N_A, N_B \in \mathbb{N}$  such that  $d(x_n, A) < \frac{\varepsilon}{2}$  for all  $n \geq N_A$ , and  $d(x_n, B) < \frac{\varepsilon}{2}$  for all  $n \geq N_B$ . Then, for  $N := \max\{N_A, N_B\}$ , we have that

$$d(A, B) \leq d(A, x_N) + d(x_N, B) < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon,$$

where we used the triangular inequality. Since  $\varepsilon > 0$  was arbitrary, it follows that  $d(A, B) = 0$ , and thus  $A = B$  because of the definiteness of  $d$ .  $\square$

We recall the notions of subsequences and accumulation points

## DEFINITION 9.20: SUBSEQUENCE

Let  $(x_n)_{n=0}^\infty$  be a sequence in a set  $X$ .

A **subsequence** of  $(x_n)_{n=0}^\infty$  is a sequence of the form  $(x_{f(k)})_{k=0}^\infty$ , where  $f : \mathbb{N} \rightarrow \mathbb{N}$  is a strictly increasing function. It is standard to denote subsequences by

$$(x_{n_k})_{k \in \mathbb{N}}, \quad (x_{n_k})_{k \geq 0}, \quad (x_{n_k})_{k=0}^\infty \quad (\text{i.e., } n_k := f(k))$$

## DEFINITION 9.21: ACCUMULATION POINTS

Let  $(X, d)$  be a metric space.

- Given  $Y \subset X$ , we say that  $x \in X$  is an **accumulation point** of  $Y$  if there exists a sequence  $(y_n)_{n \geq 0} \subset Y$  converging to  $x$ .
- Given a sequence  $(x_n)_{n \geq 0}$  in  $X$ , we say that  $x$  is an **accumulation point** of  $(x_n)_{n \geq 0}$  if some subsequence converges to  $x$ .

## LEMMA 9.22: ACCUMULATIONS POINTS OF A CONVERGING SEQUENCE

*Let  $(x_n)_{n \in \mathbb{N}}$  be a sequence in a metric space  $X$ , and let  $x \in X$ .  $(x_n)_{n \in \mathbb{N}}$  converges to  $x$  if and only if every subsequence of  $(x_n)_{n \in \mathbb{N}}$  converges to  $x$ .*

*In other words: a sequence in a metric space is convergent if and only if it has a unique accumulation point.*

*Proof.* We first prove the “only if” part. Let  $(x_{f(n)})_{n \in \mathbb{N}}$  be a subsequence, i.e.,  $f : \mathbb{N} \rightarrow \mathbb{N}$  is some strictly increasing map. Given  $\varepsilon > 0$  there is  $N$  such that  $d(x_n, x) < \varepsilon$  for all  $n \geq N$ . Hence,  $d(x_{f(n)}, x) < \varepsilon$  for all  $n \geq N$ , as  $f(n) \geq n$ .

We now prove the “if” part we can simply use that  $(x_n)_{n \geq 0}$  is a subsequence (i.e.,  $f(n) = n$ ) and hence it converges to  $x$ .  $\square$

9.23. — A stronger version of the previous Lemma that is useful in some contexts asserts the following: a sequence  $(x_n)_{n \in \mathbb{N}}$  in a metric space converges to  $x$  if and only if every partial sequence  $(x_{f(n)})_{n \in \mathbb{N}}$  ( $f$  increasing) has a sub-subsequence  $(x_{g(f(n))})_{n \in \mathbb{N}}$  ( $g$  increasing) converging to  $x$ .

While the proof of the “only if” part is similar ( $g(f(n)) \geq n$ ) the “if” part is less trivial than in the previous lemma. One can argue by contraposition: If  $x_n$  does **not** converge to  $x$  then we want to find a subsequence such that we **cannot** extract a sub-subsequence converging to  $x$ .

To do so we start by the negation of “ $x_n$  converges to  $x$ ”. Recall:

$$x_n \rightarrow x \quad \Leftrightarrow \quad \forall \varepsilon > 0 \quad \exists N \in \mathbb{N} \quad \forall n \geq N \quad d(x_n, x) < \varepsilon$$

The negation of this is:

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

In other words, the set of  $\{n \in \mathbb{N} \mid d(x_n, x) \geq \varepsilon\}$  and hence there is an increasing sequence  $(n_k)_{k \geq 0}$  such that  $d(x_{n_k}, x) \geq \varepsilon$ . Notice that any sub-subsequence will still remain at distance  $\geq \varepsilon$  from  $x$  and hence will not converge to  $x$ .

2

This stronger version can be used, for example, to prove that a continuous function  $f : [0, 1] \rightarrow \mathbb{R}$  has a unique minimum point if and only if all sequences  $(x_n)_{n \geq 0} \subset [0, 1]$  such that  $f(x_n) \rightarrow \min_{[0,1]}$  converge to the same limit point.

**LEMMA 9.24: CONVERGENCE: IN  $\mathbb{R}^n$  VERSUS COORDINATE-WISE**

*A sequence in  $\mathbb{R}^n$  converges (in the Euclidean distance) if and only if it converges coordinate-wise.*

*Proof.* Let  $\{x_k\}_{k \in \mathbb{N}} \subset \mathbb{R}^n$  be a sequence. For  $j = 1, \dots, n$ , we denote with  $x_{k,j}$  the  $j$ -th component of the vector  $x_k$ .

Assume that  $x_k \rightarrow x \in \mathbb{R}^n$ . By definition, given  $\varepsilon > 0$  and any  $j$ , it holds

$$|x_{k,j} - z_j| \leq \left( \sum_{i=1}^n (x_{k,i} - z_i)^2 \right)^{1/2} = \|x_k - x\| \leq \varepsilon \quad \text{eventually in } k.$$

This proves that, for each  $j = 1, \dots, n$ ,  $x_{k,j} \rightarrow x_j$  when  $k \rightarrow \infty$  (as sequences of real numbers, with the standard absolute value distance).

Assume now that for each  $j = 1, \dots, d$  it holds

$$x_{k,j} \rightarrow x_j \quad \text{as } k \rightarrow \infty,$$

for some numbers  $z_j \in \mathbb{R}$ . We prove that  $x_k \rightarrow x$  in  $\mathbb{R}^d$  where  $z := (z_1, \dots, z_d)$ . Given  $\varepsilon > 0$ , for each  $j \in \{1, \dots, d\}$ , there exists an  $N_j \in \mathbb{N}$  such that

$$|x_{k,j} - z_j| < \frac{\varepsilon}{n} \text{ for all } k \geq N_j.$$

This means that

$$\left( \sum_{i=1}^n (x_{k,i} - z_i)^2 \right)^{1/2} \leq \sqrt{n} \frac{\varepsilon}{n} < \varepsilon \text{ for all } k \geq \max\{N_1, \dots, N_d\},$$

which proves that  $x_k \rightarrow x$  with respect to the Euclidean distance.  $\square$

We introduce the concept of **completeness** for metric spaces. This concept does not conflict with the notion of completeness that we gave for  $\mathbb{R}$ . We will soon show that  $\mathbb{R}$ , as well as  $\mathbb{C}$ , is complete as a metric space. In contrast, the metric space  $\mathbb{Q}$  is not complete.

#### DEFINITION 9.25: CAUCHY SEQUENCE

A sequence  $(x_n)_{n=0}^{\infty}$  in a metric space  $(X, d)$  is a **Cauchy sequence** if, for every  $\varepsilon > 0$ , there exists  $N \in \mathbb{N}$  such that  $d(x_m, x_n) < \varepsilon$  for all pair of integers  $m, n$  with  $n \geq N$  and  $m \geq N$ .

2

EXERCISE 9.26. — Prove the following elementary facts about Cauchy sequences in a metric space  $(X, d)$ :

- A Cauchy sequence is bounded (meaning that  $\{d(x_n, x_0)\} \subset \mathbb{R}$  is a bounded set)
- Every convergent sequence is a Cauchy sequence.
- A Cauchy sequence converges if and only if it has a converging subsequence.

#### DEFINITION 9.27: COMPLETE METRIC SPACE

A metric space  $(X, d)$  is called **complete** if every Cauchy sequence in  $(X, d)$  converges.

EXAMPLE 9.28. — The interval  $(0, 1) \subset \mathbb{R}$ , endowed with the standard distance  $d(x, y) = |x - y|$ , is *not* complete. However,  $\mathbb{N} \subset \mathbb{R}$  is complete, as well as  $[0, \infty)$ .

EXERCISE 9.29. — Show that  $\mathbb{Q}$ , with the distance inherited from the standard distance on  $\mathbb{R}$ , is **not** a complete metric space.

EXERCISE 9.30. — Show that the space  $X$  of all bounded sequences  $(x_n)_{n \geq 0}$  of real numbers, equipped with the distance

$$d((x_n)_{n \geq 0}, (y_n)_{n \geq 0}) = \sup_{n \geq 0} |x_n - y_n|$$

is complete. Show also that the subspace  $X_0$  of sequences with limit 0 is complete.

**THEOREM 9.31: COMPLETENESS OF  $\mathbb{R}^n$**

For all  $n \geq 1$ ,  $\mathbb{R}^n$  (with the Euclidean distance) is complete. In particular,  $\mathbb{R}$  and  $\mathbb{C}$  are complete.

*Proof.* Similar to the proof Lemma 9.24, a sequence  $(x_k)$  in  $\mathbb{R}^n$  is a Cauchy if and only if  $x_{k,j}$ ,  $j = 1, \dots, n$  are Cauchy (in  $\mathbb{R}$ ). It then follows from Theorem 2.124  $\square$

9.32. — Completion of metric space (extra material) Let  $(X, d)$  be a metric space. We write  $\mathcal{C}_X$  for the set of all Cauchy sequences in  $X$  and define an equivalence relation on  $\mathcal{C}_X$  by

$$(x_n)_{n=0}^\infty \sim (y_n)_{n=0}^\infty \iff \lim_{n \rightarrow \infty} d(x_n, y_n) = 0.$$

The quotient set  $\bar{X} = \mathcal{C}_X / \sim$ , equipped with the metric  $\bar{d}$  given by

$$\bar{d}([(x_n)_{n=0}^\infty], [(y_n)_{n=0}^\infty]) = \lim_{n \rightarrow \infty} d(x_n, y_n),$$

is called the **completion** of  $(X, d)$ . The injection  $\iota : X \rightarrow \bar{X}$ , mapping  $x \in X$  to the class of the constant sequence with value  $x$ , is called the **canonical embedding**. For all  $x, y \in X$ , we have

$$d(x, y) = \bar{d}(\iota(x), \iota(y)),$$

which implies that  $\iota$  is injective.

EXERCISE 9.33. — Show that the objects introduced in 9.32 are well-defined. In particular, verify that  $\bar{d}$  is indeed a metric on  $\bar{X}$ .

EXERCISE 9.34. — As the name suggests, the completion  $(\bar{X}, \bar{d})$  of a metric space is complete, meaning that every Cauchy sequence in  $\bar{X}$  converges. A sequence in  $\bar{X}$  is essentially a sequence of sequences, i.e.,

$$[(x_{m,n})_{n=0}^\infty]_{m=0}^\infty.$$

Show that  $[(x_{n,n})_{n=0}^\infty]$  is a limit of this sequence.

EXERCISE 9.35. — Let  $(X, d)$  be a metric space with completion  $(\overline{X}, \overline{d})$ . Let  $(Y, d_Y)$  be a complete metric space, and let  $f : X \rightarrow Y$  be a function such that

$$d(x, y) = d_Y(f(x), f(y))$$

for all  $x, y \in X$ . Show that there exists a unique function  $\overline{f} : \overline{X} \rightarrow Y$  such that  $f = \overline{f} \circ \iota_X$  and

$$\overline{d}(\overline{x}, \overline{y}) = d_Y(\overline{f}(\overline{x}), \overline{f}(\overline{y}))$$

for all  $\overline{x}, \overline{y} \in \overline{X}$ . This can be interpreted as: “ $\overline{X}$  is the *smallest* complete metric space containing  $X$ .”

### 9.1.4 \*The Reals as the Completion of Rationals (extra material; cf. Grundstrukturen)

In the first semester, we defined  $\mathbb{R}$  as any complete ordered field, postulating its existence (Definitions 2.18 and 2.19). The idea of completion of metric spaces allows one to easily construct a model of  $\mathbb{R}$ . This construction shows, in particular, the existence of a complete ordered field. One can also prove with a bit of patience (although it is not hard to do so) that actually there is only one model of  $\mathbb{R}$ , in the sense that any two complete ordered fields must be isomorphic.

The completion of  $\mathbb{Q}$  serves as a model for a field of real numbers. First, note that the construction of the completion of  $\mathbb{Q}$  does not necessarily require a field of real numbers (as the target space for the standard metric on  $\mathbb{Q}$ ). The set of all Cauchy sequences  $\mathcal{C}$  in  $\mathbb{Q}$  is the set of all sequences of rational numbers  $(q_n)_{n=0}^{\infty}$  such that

$$\forall k \in \mathbb{N} \exists N \in \mathbb{N} : m, n \geq N \implies |q_n - q_m| < 2^{-k}.$$

The set  $\mathcal{C}_{\mathbb{Q}}$  is a vector space over  $\mathbb{Q}$  with component-wise operations, and

$$\mathcal{N} = \{(q_n)_{n=0}^{\infty} \in \mathcal{C} \mid \lim_{n \rightarrow \infty} q_n = 0\}$$

is a linear subspace. The equivalence relation  $(p_n - q_n)_{n=0}^{\infty} \in \mathcal{N}$  in 9.32 translates to  $(p_n - q_n)_{n=0}^{\infty} \in \mathcal{N}$ . We define the set  $\mathbb{R}$  as the quotient

$$\mathbb{R} = \mathcal{C} / \sim = \mathcal{C} / \mathcal{N}$$

in the sense of linear algebra. Thus,  $\mathbb{R}$  is a vector space over  $\mathbb{Q}$ . We denote the injective linear map  $\iota : \mathbb{Q} \rightarrow \mathbb{R}$  by the canonical embedding, which assigns to  $q \in \mathbb{Q}$  the class of the constant sequence with value  $q$ . From now on, elements of  $\mathbb{R}$  are called *real numbers*, and we consider  $\mathbb{Q}$  as a subset of  $\mathbb{R}$  via the canonical embedding  $\iota$ .

We define a product on  $\mathbb{R}$  by component-wise multiplication. That is, for elements  $x = [(p_n)_{n=0}^\infty]$  and  $y = [(q_n)_{n=0}^\infty]$ , we define

$$x \cdot y = [(p_n q_n)_{n=0}^\infty].$$

It can be verified that this gives a well-defined commutative operation on  $\mathbb{R}$ , satisfying the distributive law with respect to addition, and compatible with the multiplication of rational numbers via the canonical embedding. In particular,  $1_{\mathbb{R}} = \iota(1) = [(1)_{n=0}^\infty]$  is the multiplicative identity in  $\mathbb{R}$ . If  $x = [(q_n)_{n=0}^\infty]$  is non-zero, then  $(q_n)_{n=0}^\infty$  is a Cauchy sequence in  $\mathbb{Q}$  that does not converge to zero. Therefore,  $q_n \neq 0$  for all but finitely many  $n \in \mathbb{N}$ . The class of the sequence  $(p_n)_{n=0}^\infty$  given by

$$p_n = \begin{cases} 1 & \text{if } q_n = 0 \\ q_n^{-1} & \text{otherwise} \end{cases}$$

serves as a multiplicative inverse for  $x$ . This shows that  $\mathbb{R}$  is a field with the given operations.

We use the usual order relation on  $\mathbb{Q}$  to construct an order relation on  $\mathbb{R}$ . For elements  $x = [(p_n)_{n=0}^\infty]$  and  $y = [(q_n)_{n=0}^\infty]$  in  $\mathbb{R}$ , we declare

$$x \leq y$$

if there exists a sequence  $(r_n)_{n=0}^\infty \in \mathcal{N}$  such that  $p_n - r_n \leq q_n$  for all  $n \in \mathbb{N}$ . It is left to the diligent reader to verify that this indeed defines a well-defined order relation on  $\mathbb{R}$  that is compatible with the field structure on  $\mathbb{R}$ . Thus,  $\mathbb{R}$  is equipped with the structure of an ordered field.

It remains to show that the ordered field  $\mathbb{R}$  is complete in the sense of Definition 2.198. It is easy to see that  $\mathbb{R}$  satisfies the Archimedean Principle: Let  $x = [(q_n)_{n=0}^\infty] \in \mathbb{R}$  be positive. Then,  $(q_n)_{n=0}^\infty$  is not a null sequence. Thus, there exists a  $k \in \mathbb{N}$  such that  $|q_n| > 2^{-k}$  for infinitely many  $n \in \mathbb{N}$ . However,  $(q_n)_{n=0}^\infty$  is also a Cauchy sequence, so there exists  $N \in \mathbb{N}$  such that  $m, n \geq N \implies |q_n - q_m| < 2^{-k-1}$ . This shows that  $|q_n| > 2^{-k-1}$  and even  $q_n > 2^{-k-1}$  for all but finitely many  $n \in \mathbb{N}$ , since  $x > 0$ . This demonstrates  $\iota(2^{-k-1}) \leq x$ , or simply  $2^{-k-1} \leq x$  as we consider  $\mathbb{Q}$  as a subset of  $\mathbb{R}$ . Thus, the Archimedean Principle holds, as stated in Corollary 2.66. Now, let  $X, Y \subset \mathbb{R}$  be non-empty subsets such that  $x \leq y$  for all  $x \in X$  and  $y \in Y$ . We want to find a real number  $z = [(r_n)_{n=0}^\infty] \in \mathbb{R}$  between  $X$  and  $Y$ . To do this, we first choose arbitrary  $a_0, b_0 \in \mathbb{Q}$  such that  $[a_0, b_0] \cap X \neq \emptyset$  and  $[a_0, b_0] \cap Y \neq \emptyset$ , and set  $r_0 = \frac{1}{2}(a_0 + b_0)$ . If  $x \leq r_0 \leq Y$  for all  $x \in X$  and  $y \in Y$ , we set  $z = r_0$  and we are done. Otherwise, we define  $a_1$  and  $b_1$  as

$$\begin{cases} a_1 = r_0 \text{ and } b_1 = b_0 & \text{if } [a_0, r_0] \cap X \neq \emptyset \text{ and } [a_0, r_0] \cap Y \neq \emptyset \\ a_1 = a_0 \text{ and } b_1 = r_0 & \text{if } [r_0, b_0] \cap X \neq \emptyset \text{ and } [r_0, b_0] \cap Y \neq \emptyset \end{cases}$$

and set  $r_1 = \frac{1}{2}(a_1 + b_1)$ . By continuing this process, we either find an  $r_n$  such that  $x \leq r_n \leq Y$  for all  $x \in X$  and  $y \in Y$ , and we set  $z = r_n$ , or we obtain sequences  $(a_n)_{n=0}^\infty$ ,  $(b_n)_{n=0}^\infty$ , and



$(r_n)_{n=0}^\infty$  with  $|b_n - a_n| \leq 2^{-n}|a_0 - b_0|$  and

$$[a_n, b_n] \cap X \neq \emptyset \quad \text{and} \quad [a_n, b_n] \cap Y \neq \emptyset.$$

As the diligent reader can verify, this implies that  $(a_n)_{n=0}^\infty$ ,  $(b_n)_{n=0}^\infty$ , and  $(r_n)_{n=0}^\infty$  are all Cauchy sequences, and the real number

$$z = [(a_n)_{n=0}^\infty] = [(r_n)_{n=0}^\infty] = [(b_n)_{n=0}^\infty]$$

satisfies the inequalities  $x \leq z \leq y$  for all  $x \in X$  and  $y \in Y$ .

2

## 9.2 Topology of Metric Spaces

### 9.2.1 Open and closed sets

#### DEFINITION 9.36: OPEN BALLS

Let  $(X, d)$  be a metric space,  $x \in X$ , and  $r > 0$  a real number. In this context, we write

$$B(x, r) := \{y \in X \mid d(x, y) < r\}$$

and refer to the set  $B(x, r)$  as the **open ball** with center  $x$  and radius  $r$ .

#### DEFINITION 9.37: OPEN AND CLOSED SETS

Let  $(X, d)$  be a metric space:

- A subset  $E \subset X$  is called **open** if, for every  $x \in E$ , there exists  $r > 0$  such that  $B(x, r) \subset E$ .
- The collection of all open sets,  $\mathcal{T}_d = \{U \subset X \mid U \text{ open}\}$ , is called the **topology** generated by  $d$ .
- A subset  $E \subset X$  is called **closed** if  $X \setminus E$  is open.

3

9.38. — In particular,  $\emptyset$  and  $X$  are always both open and closed. In general, a subset  $U$  needs not to be neither open nor closed.

It is not true in general that the only “clopen” sets in a space are the empty set  $\emptyset$  and the space itself  $X$ . A set is termed “clopen” if it is both open and closed. For illustration, take the space  $X = (0, 1) \cup (2, 3)$ , equipped with the standard metric from  $\mathbb{R}$ . Here, the intervals  $(0, 1)$  and  $(2, 3)$  are clopen: they are open and closed in  $X$ . This example underscores the presence of other clopen sets beyond just  $\emptyset$  and  $X$ . The significance of clopen sets will become more apparent in our discussions on connectedness. As we will see, connected spaces are precisely characterized by the absence of nontrivial (neither empty nor the whole space) clopen sets.

9.39. — Consider the set  $X = [0, 2]$ , equipped with the standard metric inherited from  $\mathbb{R}$ . In this context, the subset  $[0, 1)$  is open within  $X$  (an exercise worth verifying). However, when considered as subset of the whole  $\mathbb{R}$ ,  $[0, 1)$  is neither open nor closed. This example illustrates that statements regarding the openness of a set like  $[0, 1)$  require clarity about the ambient space  $(X, d)$  being considered. In practice, though, such nuances are often glossed over when the context is clear, and delving into these subtleties is usually unnecessary for typical discussions.

EXERCISE 9.40. — Let  $(X, d)$  be a metric space. Show that

- The open ball  $B(x, r)$  is an open set.
- Every finite subset of  $X$  is closed.

**PROPOSITION 9.41: ARBITRARY UNIONS AND FINITE INTERSECTIONS**

*Let  $(U_i)_{i \in I}$  be any family of open subsets of  $X$ , then  $\bigcup_{i \in I} U_i$  is also open. If  $I$  is a finite set, also  $\bigcap_{i \in I} U_i$  is open.*

*Proof.* Set

$$U = \bigcup_{i \in I} U_i$$

and let  $x \in U$ . Then there exists  $i \in I$  with  $x \in U_i$ , and since  $U_i$  is open, there exists an  $\varepsilon > 0$  such that  $B(x, \varepsilon) \subset U_i$ , implying  $B(x, \varepsilon) \subset U$ . Thus,  $U$  is open. Finally, let  $(U_i)_{i \in I}$  be a finite family of open subsets of  $X$ . Set

$$U = \bigcap_{i \in I} U_i$$

and let  $x \in U$ . Then  $x \in U_i$  for all  $i \in I$ , and for each  $i \in I$ , there exists  $\varepsilon_i > 0$  such that  $B(x, \varepsilon_i) \subset U_i$ . For  $\varepsilon := \min\{\varepsilon_i \mid i \in I\}$ , we have  $\varepsilon > 0$  and  $B(x, \varepsilon) \subset U_i$  for all  $i \in I$ . Thus,  $B(x, \varepsilon) \subset U$ , completing the proof.  $\square$

**PROPOSITION 9.42: ARBITRARY INTERSECTIONS AND FINITE UNIONS**

*Let  $(A_i)_{i \in I}$  be any family of closed subsets of  $X$ , then  $\bigcap_{i \in I} A_i$  is also closed. If  $I$  is a finite set, also  $\bigcup_{i \in I} A_i$  is closed.*

*Proof.* Apply Proposition 9.41 to the (open) complements of the closed sets.  $\square$

EXAMPLE 9.43. — The intersection of infinitely many open sets may not be open. Take for example  $\mathbb{R}$  with the standard metric. The intersection of the family of open sets

$$\{(-1/n, 1/n) \mid n \in \mathbb{N}\}$$

gives  $\{0\}$ , which is not open. Taking complements you obtain an example where an infinite union of closed sets is not closed.

**DEFINITION 9.44: INTERIOR, CLOSURE, AND BOUNDARY**

Let  $X$  be a metric space and  $E \subset X$ . We define:

- The **interior**  $E^\circ := \bigcup \{U \subset E \mid U \text{ is open}\}$ , which is the largest open set contained in  $E$ . We will also use the notation  $\text{int}(E)$  to refer to the interior of  $E$ .
- The **closure**  $\bar{E} := \bigcap \{A \supset E \mid A \text{ is closed}\}$ , which is the smallest closed set containing  $E$ .
- The **(topological) boundary**  $\partial E := \bar{E} \setminus E^\circ$ .

**EXERCISE 9.45.** — Using Proposition 9.41, prove that  $E^\circ$  is always open while  $\bar{E}$  and  $\partial E$  are always closed.

**EXERCISE 9.46.** — For balls  $B(x, r)$  in  $\mathbb{R}^n$  (with the Euclidean distance) prove that  $\overline{B(x, r)} = \{y \in \mathbb{R}^n \mid d(x, y) \leq r\}$  and deduce  $\partial B(x, r) = \{y \in \mathbb{R}^n \mid d(x, y) = r\}$ .

**LEMMA 9.47: OPEN AND CLOSED THROUGH SEQUENCES**

Let  $(X, d)$  be a metric space.

- (1) A subset  $U \subset X$  is open if and only if, for every convergent sequence in  $X$  with a limit in  $U$ , the sequence eventually lies in  $U$ .
- (2) A subset  $A \subset X$  is closed if and only if, for every convergent sequence  $(x_n)_{n=0}^\infty$  in  $X$  with  $x_n \in A$  for all  $n \in \mathbb{N}$ , the limit also lies in  $A$ . In other words, if and only if  $A$  coincides with the set of all its accumulation points.

*Proof.* Let  $U \subset X$  be an open subset of  $X$ , and let  $(x_n)_{n=0}^\infty$  be a sequence in  $X$  with a limit  $x$  in  $U$ . Then, there exists  $\varepsilon > 0$  such that  $B(x, \varepsilon) \subset U$ , and since  $(x_n)_{n=0}^\infty$  converges to  $x$ , there exists an  $N \in \mathbb{N}$  such that  $x_n \in B(x, \varepsilon)$  for all  $n \geq N$ . Conversely, let  $V \subset X$  be a non-open subset. Then there exists a point  $x \in V$  such that  $B(x, \varepsilon) \setminus V \neq \emptyset$  for every  $\varepsilon > 0$ . For each  $n \in \mathbb{N}$ , we can find  $x_n \in B(x, 2^{-n}) \setminus V$ . The sequence  $(x_n)_{n=0}^\infty$  in  $X \setminus V$  converges to  $x \in V$ , and satisfies  $x_n \notin V$  for every  $n \in \mathbb{N}$ . This completes the proof of the first statement.

Let  $A \subset X$  be closed, and let  $(x_n)_{n=0}^\infty$  be a convergent sequence in  $X$  with  $x_n \in A$  for all  $n \in \mathbb{N}$ . Let  $x$  be the limit of the sequence  $(x_n)_{n=0}^\infty$ . Then,  $U = X \setminus A$  is open and cannot contain the limit  $x$  of  $(x_n)_{n=0}^\infty$ , as otherwise almost all elements of the sequence  $(x_n)_{n=0}^\infty$  would have to lie in  $U$ . Therefore, the limit  $x$  belongs to  $A$ . Finally, suppose  $A \subset X$  is not closed. Then  $U = X \setminus A$  is not open, and according to the previous argument, there exists a sequence  $(x_n)_{n=0}^\infty$  in  $A = X \setminus U$  with a limit  $x \in U$ .  $\square$

EXERCISE 9.48. — Let  $(X, d)$  be a complete metric space and  $E \subset X$  a closed subset. Show that  $E$  is complete as well.

PROPOSITION 9.49: TOPOLOGICAL NOTION OF CONVERGENCE

Let  $(X, d)$  be a metric space a sequence  $(x_n)_{n \geq 0}$  converges to  $x$  if and only if for all open sets  $U$  containing  $x$ ,  $x_n$  eventually lies in  $U$ .

*Proof.* Notice that we can rewrite the definition of convergence as follows:  $x_n \rightarrow x$  if and only if for all  $\varepsilon > 0$ ,  $x_n$  eventually lies in  $B(x, \varepsilon)$ . Now, if  $U$  is any open set containing  $x$  then by definition of open set there exists  $\varepsilon > 0$  such that  $B(x, \varepsilon) \subset U$  and hence  $x_n \rightarrow x$  implies that  $x_n$  eventually lies in  $U$ , establishing the “only if” direction. For the “if” part, we observe that for any given  $\varepsilon > 0$  we can take  $U = B(x, \varepsilon)$  (open balls are open) and hence  $x_n$  eventually lies in  $B(x, \varepsilon)$ .  $\square$

COROLLARY 9.50: DISTANCES WITH SAME CONVERGENT SEQUENCES

Let  $X$  be a set endowed with two different distances  $d_1$  and  $d_2$ . Then  $(X, d_1)$  and  $(X, d_2)$  have the same convergent sequences if and only if the topologies generated by  $d_1$  and  $d_2$  coincide.

*Proof.* By Corollary 9.49 the notion of convergent sequence only depends on the collection of open sets. That is, it only depends on the distance through the topology it generate. Hence distances generating the same open sets have the same convergent sequences. This proves the “if” part of the statement.

We prove the “only if” part. Suppose that a set  $U \subset X$  is open with respect to  $d_1$ , but not with respect to  $d_2$ . This means that there is  $x \in U$  such that we can find

$$x_k \in B_{d_2}(x, 2^{-k}) \cap (X \setminus U) \neq \emptyset \text{ for all } k \geq 0.$$

By construction, the sequence  $\{x_k\}$  is the convergent with respect to  $d_2$  and then also with respect to  $d_1$  (they have the same convergent sequences by assumption). By Proposition 9.49 ( $U$  is open in  $d_1$ !) we discover that  $x_k \in U$  eventually, which contradicts how we constructed  $\{x_k\}$ .  $\square$

EXERCISE 9.51. — Let  $(X, d)$  be a metric space, and  $A \subset X$  a subset.

- (i) Assume  $X$  is complete and  $A$  is closed. Show that the subspace  $A$  is also complete.
- (ii) Assume  $A$  is complete. Show that  $A \subset X$  is closed.

### 9.2.2 Continuity

We now aim to generalize the concept of continuity to functions defined between metric spaces.

#### DEFINITION 9.52: CONTINUITY

Let  $(X, d_X)$  and  $(Y, d_Y)$  be metric spaces, and let  $f : X \rightarrow Y$  be a function. We say that  $f$  is continuous if one of the following **equivalent** conditions hold:

- (1) We say  $f$  is  **$\varepsilon - \delta$  continuous** if, for all  $x \in X$  and  $\varepsilon > 0$ , there exists a  $\delta > 0$  such that if  $d_X(x, x') < \delta$ ,  $x' \in X \implies d_Y(f(x), f(x')) < \varepsilon$ . In other words,  $f(B(x, \delta)) \subset B(f(x), \varepsilon)$ .
- (2) We say  $f$  is **sequentially continuous** if, for every convergent sequence  $(x_n)_n$  in  $X$  with limit  $x = \lim_{n \rightarrow \infty} x_n$ , the sequence  $(f(x_n))_n$  converges in  $Y$ , with  $f(x) = \lim_{n \rightarrow \infty} f(x_n)$ .
- (3) We say  $f$  is **topologically continuous** if, for every open subset  $U \subset Y$ , the preimage  $f^{-1}(U) = \{x \in X \mid f(x) \in U\}$  is open in  $X$ .

#### PROPOSITION 9.53: THE THREE FACES OF CONTINUITY

Let  $X$  and  $Y$  be metric spaces, and let  $f : X \rightarrow Y$  be a function. The following conditions are equivalent:

- (1) The function  $f$  is  $\varepsilon - \delta$  continuous.
- (2) The function  $f$  is sequentially continuous.
- (3) The function  $f$  is topologically continuous.

*Proof.* (1)  $\implies$  (2): Let  $(x_n)_{n=0}^\infty$  be a convergent sequence in  $X$  with limit  $x \in X$ , and let  $\varepsilon > 0$ . There exists a  $\delta > 0$  such that  $f(x') \in B(f(x), \varepsilon)$  for all  $x' \in B(x, \delta)$ . Since  $(x_n)_{n=0}^\infty$  converges to  $x$ , there exists an  $N \in \mathbb{N}$  such that  $x_n \in B(x, \delta)$  for all  $n \geq N$ . In particular, for  $n \geq N$ ,  $f(x_n) \in B(f(x), \varepsilon)$ . Since  $\varepsilon > 0$  was arbitrary, it follows that  $\lim_{n \rightarrow \infty} f(x_n) = f(x)$ , and thus  $f$  is sequentially continuous as claimed.

$\neg(3) \implies \neg(2)$ : Assume  $f$  is not topologically continuous. Then exists  $U \subset Y$  open such that  $f^{-1}(U)$  is not open. Therefore, there is  $x \in f^{-1}(U)$  and a sequence  $(x_n)_{n \geq 0} \subset X \setminus f^{-1}(U)$  with  $x_n \rightarrow x$ . Then  $f(x) \in U$  and  $f(x_n) \in Y \setminus U$  for all  $n$ , but  $U$  is open this gives that  $f(x_n)$  cannot converge to  $f(x)$ . In other words, we have found a sequence such that  $x_n \rightarrow x$ , but  $f(x_n)$  does not converge to  $f(x)$ .

(3)  $\implies$  (1): Let  $x \in X$  and  $\varepsilon > 0$ . The preimage  $f^{-1}(B(f(x), \varepsilon))$  contains the point  $x$  and is open by assumption, as  $B(f(x), \varepsilon) \subset Y$  is open. Thus, there exists a  $\delta > 0$  such that  $B(x, \delta) \subset f^{-1}(B(f(x), \varepsilon))$ . Therefore,  $f$  is  $\varepsilon$ - $\delta$ -continuous as claimed.  $\square$

**DEFINITION 9.54: UNIFORM AND LIPCHITZ CONTINUITY**

Let  $(X, d_X)$  and  $(Y, d_Y)$  be metric spaces. We say that  $f: X \rightarrow Y$  is

- **Uniformly continuous** if for every  $\varepsilon > 0$  there is  $\delta > 0$  such that  $d_Y(f(x), f(x')) < \varepsilon$  whenever  $d_X(x, x') < \delta$ .
- **Lipschitz continuous** if there is a constant  $L > 0$  such that

$$d_Y(f(x), f(x')) \leq L d_X(x, x') \quad \text{for all } x, x' \in X.$$

The constant  $L$  is called **Lipschitz constant** of  $f$

EXAMPLE 9.55. — In any metric space  $(X, d)$  for any given  $x_0 \in X$  the function  $f(x) = d(x, x_0)$  is Lipschitz (with constant 1). Indeed, the triangle inequality (using also symmetry) yields:

$$-d(x, x') \leq d(x, x_0) - d(x', x_0) \leq d(x, x').$$

EXERCISE 9.56. — Let  $(X, d)$  be a metric space, and let  $E \subset X$  be a non-empty subset. For  $x \in X$ , define

$$f_E(x) = \inf\{d(x, z) \mid z \in E\}.$$

Show that the function  $f_E: X \rightarrow \mathbb{R}$  is 1-Lipschitz continuous, and that  $E \subset X$  is closed if and only if  $E = \{x \in X \mid f_E(x) = 0\}$ .

EXERCISE 9.57. — Let  $(X, d_X)$  and  $(Y, d_Y)$  be metric spaces. Assume that  $Y$  is complete. Show that if  $E \subset X$  and  $f: E \rightarrow Y$  is a uniformly continuous function defined on a subset then there is a unique continuous extension  $\bar{f}: \bar{E} \rightarrow Y$ , which is also uniformly continuous.

### 9.2.3 Banach's Fixed-Point Theorem

Fixed-point theorems investigate when a mapping  $f: X \rightarrow X$  possess a fixed point, i.e., a point  $x \in X$  for which  $f(x) = x$ . They can be rather powerful tools to prove existence theorems.

We prove in this section Banach's Fixed-Point Theorem, which will be used later to prove the Implicit Function Theorem ?? and the existence of solutions to ODEs.

**THEOREM 9.58: BANACH'S FIXED-POINT THEOREM**

Let  $(X, d)$  be a **complete** metric space and let  $T : X \rightarrow X$  be a Lipschitz map with Lipschitz constant  $\lambda < 1$ . In other words, assume that for some  $\lambda \in (0, 1)$  it holds

$$d(T(x), T(x')) \leq \lambda d(x, x') \quad \text{for all } x, x' \in X.$$

Then, there **exists a unique** element  $a \in X$  such that  $T(a) = a$ .

The function  $T$  is a **Lipschitz contraction**. A point  $x \in X$  with  $T(x) = x$  is called a **fixed point** of the mapping  $T$ , and the theorem states that a Lipschitz contraction has a unique fixed point, provided the ambient space is complete (as in Definition 9.27).

*Proof.* First, we show uniqueness of a putative fixed point. Let  $a \in X$  and  $a' \in X$  be fixed points of  $T$ . Then,

$$d(a, a') = d(T(a), T(a')) \leq \lambda d(a, a'),$$

which, since  $\lambda < 1$ , implies  $d(a, a') = 0$  and thus  $a = a'$ .

We turn to prove the existence of a fixed point. Choose any  $x_0 \in X$  and define a sequence  $(x_n)_{n=0}^{\infty}$  recursively by  $x_{n+1} = T(x_n)$ , for  $n \geq 0$ . We claim that the sequence  $(x_n)_{n=0}^{\infty}$  is a Cauchy sequence. Iterating the contractivity assumption we find that, for all integers  $p \geq 0$ ,

$$\begin{aligned} d(x_{n+1}, x_n) &= d(T(x_n), T(x_{n-1})) \leq \lambda d(x_n, x_{n-1}) \\ &\leq \lambda^2 d(x_{n-1}, x_{n-2}) \leq \dots \leq \lambda^n d(x_1, x_0). \end{aligned}$$

Pick now any integers  $m \geq n \geq N$ , then using this observation and the triangular inequality we find

$$d(x_m, x_n) \leq \sum_{p=n}^{m-1} \lambda^p d(x_1, x_0) \leq d(x_1, x_0) \sum_{p \geq N}^{\infty} \lambda^p = \frac{\lambda^N}{1-\lambda} d(x_0, x_1).$$

We crucially used that  $\lambda < 1$  to sum the geometric series. Now given any  $\varepsilon > 0$  we can find some  $N$  so large that  $\frac{\lambda^N}{1-\lambda} d(x_0, x_1) < \varepsilon$ , thus proving that  $(x_n)_{n=0}^{\infty}$  is Cauchy (this estimate is uniform in  $n, m$  as long as they are larger than  $N$ !).

Now we use the completeness assumption to infer that  $x_n \rightarrow a$  for some  $a \in X$ . Since  $T$  is continuous, we have

$$T(a) = \lim_{n \rightarrow \infty} T(x_n) = \lim_{n \rightarrow \infty} x_{n+1} = \lim_{n \rightarrow \infty} x_n = a$$

which shows that  $a$  is a fixed point of  $T$ . □

9.59. — We remark that the proof is constructive and in concrete situation can be implemented in an algorithm to find approximate fixed points.

EXERCISE 9.60. — Find examples for:

- (1) A Lipschitz contraction  $T : X \rightarrow X$  on a non-complete metric space  $X$  that has no fixed point.
- (2) A complete metric space  $(X, d)$  and an isometry (i.e., a mapping  $T : X \rightarrow X$  with  $d(T(x_1), T(x_2)) = d(x_1, x_2)$ ) that has no fixed point

### 9.2.4 Compactness

A closed and bounded interval of the real line is called a **compact** interval, as we saw in Analysis I. We proved some fundamental properties of continuous functions on compact intervals: boundedness, existence of maxima and minima, and uniform continuity. We intend to investigate these and other properties in the broader context of metric spaces. We start giving a general definition of compactness that works in metric spaces.

Let us immediately clarify a possible source of confusion.

4

#### ACHTUNG! 9.61: CLOSED & BOUNDED VS COMPACT

- In a general metric space, it is **not** necessarily true that a closed and bounded set is compact.
- Nevertheless when considering  $\mathbb{R}^n$  with its Euclidean structure, which is the main focus of this course, it will turn out that a closed and bounded set **is** indeed compact, and viceversa.

#### INTERLUDE: “COVER”

Let  $E \subset X$  and let  $\mathcal{U} = \{U_i\}_{i \in I}$  be a family of subsets of  $X$ , where  $I$  is some set of indices. We say that  $\mathcal{U}$  **covers**  $E$  if

$$E \subset \bigcup \mathcal{U} = \bigcup_{i \in I} U_i.$$

Any family  $\mathcal{V} \subset \mathcal{U}$  that still covers  $E$  is called a **subcover** of  $\mathcal{U}$ .



## DEFINITION 9.62: COMPACTNESS

Let  $(X, d)$  be a metric space. A subset  $K \subset X$  is called **compact** if one of the following equivalent conditions hold:

- (1)  $K$  is **sequentially compact**: every sequence  $(x_n)_{n \in \mathbb{N}}$  in  $K$  has a subsequence that is convergent in  $K$ .
- (2)  $K$  is **topologically compact**: every family of open sets  $\{U_i\}_{i \in I}$  that cover  $K$ , has a finite subcover.
- (3)  $K$  is **complete** (i.e., every Cauchy sequence contained in  $K$  has a limit in  $K$ ) and **totally bounded**: for every  $r > 0$ , there exist finitely many  $x_1, \dots, x_n \in K$  such that the balls  $B(x_1, r), \dots, B(x_n, r)$  cover  $K$ .

4

9.63. — The definition of topological compactness does not explicitly use the distance function  $d$ , but it is formulated only in terms of the collection of open sets (i.e., the topology). For this reason it is called “topological”.

9.64. — The Bolzano-Weierstrass Theorem ensures that a closed and bounded interval of  $\mathbb{R}$  is sequentially compact.

EXAMPLE 9.65. —  $\mathbb{Q} \cap [0, 2]$ , endowed with the standard distance, is not topologically compact. Consider the covering

$$\mathbb{Q} \cap [0, 2] = (\mathbb{Q} \cap [0, \sqrt{2})) \cup \bigcup_{p \in \mathbb{Q}, p > \sqrt{2}} (\mathbb{Q} \cap (p, 2]).$$

5

Any finite subcover will miss some rationals  $> \sqrt{2}$ .

EXERCISE 9.66. — Let  $X$  be a metric space. Show that if  $X$  is totally bounded, then it is bounded, i.e.,  $\sup_{x, x' \in X} d(x, x') < +\infty$ .

EXAMPLE 9.67. — The half-open interval  $X = (0, 1] \subset \mathbb{R}$  is not compact. Indeed, the open cover  $\mathcal{U} = \{(2^{-n}, 1] \mid n \in \mathbb{N}\}$  has no finite subcover.

The main result of this section is Theorem 9.69, which shows that the definition of compact metric space is indeed well-posed.

EXERCISE 9.68. — Show that a totally bounded metric space is bounded, and find an example of a bounded metric space that is not totally bounded.

## THEOREM 9.69: THE THREE FACES OF COMPACTNESS

Let  $(X, d)$  be a metric space and  $K \subset X$  a subset, the following statements are equivalent:

- (1)  $K$  is sequentially compact.
- (2)  $K$  is topologically compact.
- (3)  $K$  is totally bounded and complete (in the sense of Cauchy sequences).

We will prove that (1)  $\implies$  (2)  $\implies$  (3)  $\implies$  (1).

*Proof that (1)  $\implies$  (2).* We start with a preliminary observation. Consider the function

$$r(x) := \min \left\{ 1, \sup \{ r > 0 : B(x, r) \subset U \text{ for some } U \in \mathcal{U} \} \right\},$$

defined for all  $x \in K$ .

Notice that  $r(x) > 0$  since  $\mathcal{U}$  is open cover. For each  $x \in K$  we choose — once and for all — some  $U(x) \in \mathcal{U}$  which is almost optimal in the sense that  $B(x, r(x)/2) \subset U(x)$ .

Let us now proceed with the construction of our finite subcover. Pick any  $U_0 \in \mathcal{U}$ . If  $K \subset U_0$  then we are done, otherwise there is some  $x_1 \in K \setminus U_0$ . In this case we set  $U_1 := U(x_1)$ .

Now we check if  $K \subset U_0 \cup U_1$ , in which case we have found our finite subcover. If not, there is some  $x_2 \in K \setminus (U_0 \cup U_1)$  and we set  $U_2 := U(x_2)$ .

Now we check again if  $K \subset U_0 \cup U_1 \cup U_2$ , in which case we have found our finite subcover. If not, there is some  $x_3 \in K \setminus (U_0 \cup U_1 \cup U_2)$  and we set  $U_3 := U(x_3)$ .

If this procedure stops at a certain point, it means that we have found our finite open subcover. So let's assume that it goes on indefinitely, and find a contradiction. We obtain a sequence  $(x_n)_{n \in \mathbb{N}}$  with the property that

$$x_m \notin U_0 \cup \dots \cup U_n \text{ for all } m > n \geq 0.$$

By assumption (1), we have  $x_{n_k} \rightarrow z$  (as  $k \rightarrow \infty$ ) for a suitable subsequence. On the other hand, for each  $k \geq 0$  we have  $z \notin U_{n_k}$  (the complement of an open set is closed, that is sequentially closed), and in particular  $z \notin B(x_{n_k}, r(x_{n_k})/2)$ . Combining this information with  $x_{n_k} \rightarrow z$ , we find  $r(x_{n_k}) \rightarrow 0$  and in particular

$$r(x_{n_k}) = \sup \{ r > 0 : B(x_{n_k}, r) \subset U \text{ for some } U \in \mathcal{U} \} \in (0, 1).$$

This is a contradiction as  $r(z) > 0$ , so for  $k$  large enough we would have

$$B(x_{n_k}, 100r(x_{n_k})) \subset B(z, r(z)/2) \subset U(z) \in \mathcal{U},$$

which, by definition of  $r(x_{n_k})$ , implies  $100r(x_{n_k}) \leq r(x_{n_k})$ , impossible. □

In order to prove (2)  $\implies$  (3) we single out a rephrasing of (2) which is useful to keep in mind.

**LEMMA 9.70: NESTING PRINCIPLE**

Let  $X$  be a metric space, then  $K \subset X$  is topologically compact if and only if has the following property, called the “Nesting property”. For every collection  $\mathcal{A} = \{A_i\}_{i \in I}$  of closed subsets of  $X$ , it holds:

“If every intersection of finitely many sets in  $\mathcal{A}$  has a non-empty intersection with  $K$ , then  $K \cap \bigcap_{i \in I} A_i$  is non-empty.”

*Proof of Lemma 9.70.* Assume  $K$  is compact, and let  $\mathcal{A}$  be a collection of closed subsets of  $X$  with empty intersection. The collection of complements  $\mathcal{U} = \{X \setminus A : A \in \mathcal{A}\}$  is then an open cover of  $K$ , so there exists a finite subcover  $K \subset U_1 \cup \dots \cup U_n$  of it. Set  $A_i := K \setminus U_i$ . Then  $K \cap A_1 \cap \dots \cap A_n = \emptyset$ . Thus,  $K$  satisfies the Nesting property.

Now, assume  $K$  satisfies the Nesting property, and let  $\mathcal{U}$  be an open cover of  $K$ . Then  $\mathcal{A} = \{X \setminus U : U \in \mathcal{U}\}$  is a collection of closed subsets with an empty intersection with  $K$ . According to the Nesting property, there must exist  $A_1, \dots, A_n \in \mathcal{A}$  with an empty intersection. So if we consider  $U_i = X \setminus A_i$ , then  $X = U_1 \cup \dots \cup U_n$  is a finite subcover of  $K$ . Since the cover  $\mathcal{U}$  was arbitrary, this shows that  $K$  is compact.  $\square$

*Proof that (2)  $\implies$  (3).* We first prove that  $X$  is totally bounded. Pick any  $r > 0$ , consider the open covering  $\mathcal{U} := \{B(x, r) : x \in X\}$  and extract a finite subcover  $\{B(x_1, r), \dots, B(x_N, r)\}$ .

Now we prove that  $X$  must be complete, hence we pick a Cauchy sequence  $(x_n)_{n \in \mathbb{N}}$  and show that it has a limit point. For each  $k \geq 0$  there is  $n(k)$  so large that

$$n, m \geq n(k) \implies d(x_n, x_m) < 2^{-k}.$$

For each  $k$ , consider the closed balls  $A_k := \overline{B}(x_{n(k)}, 2^{-k})$ ; any finite intersection of them is nonempty, indeed for every  $k_1, \dots, k_N$  one has that  $x_m \in A_{k_1} \cap \dots \cap A_{k_N}$ , provided  $m \geq \max\{k_1, \dots, k_N\}$ . Hence we can apply the Nesting Principle (see Lemma 9.70) and find some  $z \in \bigcap_{k \geq 0} A_k$ . We claim that  $x_n \rightarrow z$ . Indeed if  $m \geq n(k)$  it holds

$$d(x_m, z) \leq d(x_m, x_{n(k)}) + d(x_{n(k)}, z) \leq 2^{1-k},$$

and  $2^{1-k}$  is arbitrarily small.  $\square$

Before continuing the proof we need an auxiliary Lemma.

**LEMMA 9.71: DIAGONAL ARGUMENT**

Let  $\mathbb{N} \supset N_0 \supset N_1 \supset N_2 \supset \dots$  be a an infinite family of nested sets. Assume further that each  $N_k$  has infinite many elements. Then there exists  $f: \mathbb{N} \rightarrow \mathbb{N}$  strictly increasing such that  $f(k) \in N_k$  for all  $k \geq 0$ .

*Proof of Lemma 9.71.* Set  $f(0)$  equal to an arbitrary element of  $N_0$ . Then set inductively  $f(k) := \min\{m \in N_k : m > f(k-1)\}$ , this set is non-empty because each  $N_k$  is infinite.  $\square$

*Proof that (3)  $\implies$  (1).* We pick any sequence  $(x_n)_{n \in \mathbb{N}}$  and show that admits a Cauchy subsequence, by completeness, this will prove (1).

By assumption, we can cover  $K$  by a finite number of balls of radius 1; it follows that  $(x_n)_{n \in \mathbb{N}}$  will fall infinitely many times (at least) one of these balls. Let this ball be  $B(z_1, 1)$  for some  $z_1 \in X$ . Accordingly, we define the set of indices  $N_0 := \{j \in \mathbb{N} : x_j \in B(z_1, 1)\}$ , which is infinite.

Now we proceed to do same thing to the restricted sequence  $(x_n)_{n \in N_0}$ , but we shorten the radius of from 1 to  $1/2$ . Accordingly, we find a ball  $B(z_2, 1/2)$  such that the set  $N_1 := \{j \in N_0 : x_j \in B(z_2, 1/2)\}$  is infinite.

We proceed inductively, halving the radius each time, and construct a descending family of infinite sets  $\mathbb{N} \supset N_0 \supset N_1 \supset N_2 \supset \dots$  with the property that

$$\forall k \geq 0, \exists z_k \in X, j \in N_k \implies d(z_k, x_j) < 2^{-k}.$$

We apply to these sets Lemma 9.71 and find  $f: \mathbb{N} \rightarrow \mathbb{N}$  strictly increasing such that  $f(k) \in N_k$  for all  $k \geq 0$ . Then, the subsequence  $(x_{f(k)})_{k \in \mathbb{N}}$  is Cauchy: for  $n, m \geq k$  it holds

$$\begin{aligned} f(n) \in N_n, f(m) \in N_m &\implies f(n), f(m) \in N_k \\ &\implies x_{f(n)}, x_{f(m)} \in B(z_k, 2^{-k}) \implies d(x_{f(n)}, x_{f(m)}) < 2^{1-k}. \end{aligned}$$

$\square$

We conclude the chain of implications thus proving Theorem 9.69.

Let us next give some simple consequences of the three faces of compactness result.

#### COROLLARY 9.72: CLOSED IN COMPACT IS COMPACT

*Let  $X$  be a metric space,  $A \subset X$  a closed subset and  $K \subset X$  a compact subset. Then  $A \cap K$  is compact.*

*Proof.* We check that  $A \cap K$  is sequentially compact. Take any sequence  $(x_n)_{n \in \mathbb{N}} \subset A \cap K$ , by compactness of  $K$ , it has a converging subsequence  $x_{f(n)} \rightarrow x$ , for some  $x \in K$ . On the other hand, since  $A$  is closed it contains its accumulation points, so we also have  $x \in A$ .  $\square$

#### COROLLARY 9.73: COMPACT IS ALWAYS CLOSED

*In a metric space, every compact subset is closed.*

*Proof.* Take any sequence in  $E$  which is convergent to some  $x \in X$ . By compactness a suitable subsequence is converging to some  $x' \in E$ , by uniqueness of the limit  $x = x'$ , so  $E$  is closed (see (2) in Lemma 9.47).  $\square$

We also easily get the following version of the Heine Borel theorem in the Euclidean space  $\mathbb{R}^n$ . In its statement, a set  $E \subset \mathbb{R}^n$  is called bounded if there exist  $N \in \mathbb{N}$  such that  $E \subset [-2^N, 2^N]^n$  (prove that this is equivalent to  $\sup \{d(x, x') \mid (x, x') \in E \times E\} < \infty$ ).

**THEOREM 9.74: COMPACT SUBSETS OF  $\mathbb{R}^n$  (HEINE-BOREL)**

*A subset  $K \subset \mathbb{R}^n$  is compact if and only if it is closed and bounded.*

*Proof.* If  $K$  is compact, then it is closed by Corollary 9.73; and bounded, because it is totally bounded.

To show the converse, we show that  $K$  is complete and totally bounded. Since  $\mathbb{R}^n$  is complete and  $K$  is closed, then  $K$  is complete as well (see Exercise 9.51).

On the other hand, by assumption  $K$  is bounded so there exist  $N \in \mathbb{N}$  such that  $K \subset [-2^N, 2^N]^n$ . Given  $r > 0$ , take some large integer  $N' \in \mathbb{N}$  large so that  $2^{-N'} < r/\sqrt{n}$ . Then the union of the balls  $B_r(x)$  with  $x$  running in the finite grid

$$\{2^{-N'}y \mid y = (y_1, \dots, y_n) \in \mathbb{Z}^n \text{ with } -2^{N+N'} \leq y_i \leq 2^{N+N'}\}$$

covers  $K$ , proving that it is totally bounded. □

6

**EXAMPLE 9.75.** — We stress that the Heine-Borel Theorem **fails** for general metric spaces: take  $\mathbb{R}$  with the bounded distance  $d(x, y) := \arctan|x - y|$  (see Exercise 9.8). Then in this metric space the set  $\mathbb{N}$  would be closed and bounded, but it is trivial to construct sequences that do not converge.

**EXERCISE 9.76.** — Show that an open set  $U \subset \mathbb{R}^n$  is complete if and only if  $U = \mathbb{R}^n$ .

### 9.2.5 Compactness and continuity

**THEOREM 9.77: CONTINUOUS IMAGE OF COMPACT IS COMPACT**

*Let  $X$  and  $Y$  be metric spaces, let  $f : X \rightarrow Y$  be a continuous function, and let  $K \subset X$  be a compact subset. Then,  $f(K)$  is a compact subset of  $Y$ .*

*Proof.* We will give two proofs. The first one employs sequential compactness. Suppose that  $(y_n)_{n \geq 0}$  is a sequence of points in  $f(K)$ . Then  $y_n = f(x_n)$  for some  $(x_n)_{n \geq 0}$ . Since  $K$  is compact there is a convergent subsequence  $x_{n_k} \rightarrow x \in K$ . But then since  $f$  is continuous  $y_{n_k} \rightarrow y = f(x) \in f(K)$ .

The second one employs instead topological compactness. Let  $\mathcal{U}$  be an open cover of  $f(A)$ . For each  $U \in \mathcal{U}$ , the set  $f^{-1}(U) \subset X$  is open due to the continuity of  $f$ . The collection  $\{f^{-1}(U) \mid U \in \mathcal{U}\}$  is an open cover of  $A$ . Since  $A$  is compact, there exist  $U_1, \dots, U_n \in \mathcal{U}$  such

that  $\{f^{-1}(U_i) \mid 1 \leq i \leq n\}$  is a cover of  $K$ . This implies that  $\{U_i \mid 1 \leq i \leq n\}$  is a cover of  $f(K)$ , and since  $\mathcal{U}$  was arbitrary, it shows that  $f(K)$  is compact.  $\square$

**PROPOSITION 9.78: CONTINUOUS IN A COMPACT IS UNIFORMLY CONTINUOUS**

*Let  $(X, d_X)$  and  $(Y, d_Y)$  be metric spaces, and  $f : X \rightarrow Y$  a continuous function. If  $X$  is compact, then  $f$  is uniformly continuous.*

*Proof.* Let  $\varepsilon > 0$ . Due to the continuity of  $f$ , for each  $x \in X$ , there exists  $\delta_x > 0$  such that  $f(B(x, \delta_x)) \subset B(f(x), \frac{\varepsilon}{2})$ . The collection  $\{B(x, \frac{1}{2}\delta_x) \mid x \in X\}$  forms an open cover of  $X$ . Since  $X$  is compact by assumption, there exists a finite subcover of this collection. This implies the existence of  $x_1, \dots, x_n \in X$  such that

$$X = B(x_1, \frac{1}{2}\delta_{x_1}) \cup \dots \cup B(x_n, \frac{1}{2}\delta_{x_n}).$$

Let  $\delta = \frac{1}{2} \min\{\delta_{x_1}, \dots, \delta_{x_n}\}$ . For  $x, x' \in X$  with  $d_X(x, x') < \delta$ , there exists  $k$  such that  $x \in B(x_k, \frac{1}{2}\delta_{x_k})$ . This implies  $x' \in B(x_k, \delta_{x_k})$ , leading to

$$d_Y(f(x), f(x')) \leq d_Y(f(x), f(x_k)) + d_Y(f(x_k), f(x')) \leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon,$$

which completes the proof.  $\square$

**COROLLARY 9.79: WEIERSTRASS**

*Let  $X$  a metric space,  $f : X \rightarrow \mathbb{R}$  be a continuous function, and  $K \subset X$  a compact subset. Then,  $f$  admits a maximum point, i.e., there exist  $\bar{x} \in X$  such that  $f(\bar{x}) = \sup_K f = \sup \{f(x) \mid x \in K\}$ . An analogous statement holds for the minimum. In particular,  $f$  must be bounded.*

*Proof.*  $f(K) \subset \mathbb{R}$  is compact by Theorem 9.77 and nonempty, so  $\sup f(K) \in f(K)$ , any element  $\bar{x} \in f^{-1}(\sup f(K))$  works.

The fact that  $\sup f(K) \in f(K)$  is readily proved: by definition of supremum there is a sequence  $(s_n) \subset f(K)$  such that  $s_n \rightarrow \sup f(K)$ , but then  $\sup f(K) \in f(K)$ , since  $f(K)$  is closed and so it contains its accumulation points.  $\square$

## 9.2.6 Connectedness

## DEFINITION 9.80: CONNECTEDNESS

Let  $X$  be a metric space. A nonempty subset  $E \subseteq X$  is called **disconnected** if there exist two disjoint open sets  $U_1$  and  $U_2$  such that  $E \cap U_i \neq \emptyset$  for  $i = 1, 2$  and  $E \subset U_1 \cup U_2$ . Conversely,  $E$  is called **connected** if, for any pair open sets  $U_1$  and  $U_2$  such that  $E \cap U_i \neq \emptyset$  for  $i = 1, 2$  and  $E \subset U_1 \cup U_2$  it is required that  $U_1 \cap U_2$  is not empty. A subset  $F \subset E$  is called a **connected component of  $E$**  if  $F$  is non-empty, connected, and every  $G \subset E$  such that  $F \subsetneq G$ ,  $G$  is disconnected.

EXERCISE 9.81. — Let  $X$  be a metric space, and let  $E_1$  and  $E_2$  be connected subsets. If the intersection  $E_1 \cap E_2$  is non-empty, then the union  $E_1 \cup E_2$  is connected. Can you generalize this property to arbitrary unions? Use this property to define the connected component of a point as the union of all connected sets that contain the point.

PROPOSITION 9.82: CONNECTED SUBSETS OF  $\mathbb{R}$ 

A non-empty subset  $E \subset \mathbb{R}$  is connected in  $\mathbb{R}$  if and only if  $E$  is an interval.

*Proof.* Assume  $E \subset \mathbb{R}$  is not an interval. Then there exist real numbers  $x_1 < y < x_2$  with  $x_1, x_2 \in E$  and  $y \notin E$ . Then the two disjoint open subset

$$U_1 = (-\infty, y) \quad \text{and} \quad U_2 = (y, \infty)$$

cover  $E$ , so  $E$  is not connected.

To prove the opposite implication we also argue by contraposition. Suppose that  $E$  is disconnected and let  $U_1, U_2$  disjoint open sets, both intersecting  $E$ , such that  $E \subset U_1 \cup U_2$ . Choose  $x_1 \in E \cap U_1$  and  $x_2 \in E \cap U_2$ . Without loss of generality, assume  $x_1 < x_2$ .

Now let us define

$$t_* := \sup\{t \geq x_1 : [x_1, t] \subset U_1\}, \quad (9.2)$$

since  $x_2 \in U_2$  we have  $t_* \in [x_1, x_2)$ . On the one hand  $t_*$  is the supremum (and hence the limit) of points in  $\mathbb{R} \setminus U_2$ , a closed set, and hence  $t_* \in \mathbb{R} \setminus U_2$ .

On the other hand let us show that  $t_* \notin U_1$ . If  $t_* \in U_1$  then  $t_* < x_2$  and, since  $U_1$  is open, we could enlarge a bit  $t_*$  while still satisfying (9.2). This violates the very definition of  $t_*$  in (9.2).

Therefore we have found two points  $x_1, x_2$  in  $E$  and third point  $t_* \in (x_1, x_2)$  with  $t_* \in \mathbb{R} \setminus (U_1 \cup U_2) \subset \mathbb{R} \setminus E$ . Hence,  $E$  cannot be an interval.  $\square$

## PROPOSITION 9.83: CONTINUOUS IMAGE OF CONNECTED

Let  $X$  and  $Y$  be metric spaces, and let  $f : X \rightarrow Y$  be continuous. If  $E \subset X$  is connected, then the image  $f(E)$  is a connected subspace of  $Y$ .

*Proof.* Suppose  $f(E)$  is disconnected. Then there exists two nonempty, disjoint, open sets  $U_1, U_2$  of  $Y$  such that  $f(E)$  is covered by  $U_1 \cup U_2$ . But then  $f^{-1}(U_1)$  and  $f^{-1}(U_2)$  are two nonempty disjoint open sets covering  $E$ , contradicting its connectedness.  $\square$

**COROLLARY 9.84: INTERMEDIATE VALUE THEOREM**

Let  $I \subset \mathbb{R}$  be an interval,  $f : I \rightarrow \mathbb{R}$  a continuous function, and  $a, b \in I$ . For every  $y \in \mathbb{R}$  between  $f(a)$  and  $f(b)$ , there exists an  $x \in I$  between  $a$  and  $b$  such that  $f(x) = y$ .

*Proof.* Without loss of generality, we can assume  $a < b$ . Applying Propositions 9.82 and 9.83  $f([a, b])$  is connected. But then, using again Proposition 9.82,  $f([a, b]) \subset \mathbb{R}$  must be an interval. As  $f(a), f(b) \in f([a, b])$ , all values between  $f(a)$  and  $f(b)$  lie in  $f([a, b])$ .  $\square$

**EXERCISE 9.85.** — Show the following generalization of the Intermediate Value Theorem: Let  $X$  be a connected topological space, and  $f : X \rightarrow \mathbb{R}$  be a continuous function. Let  $a, b \in X$ . Then, for every  $y \in \mathbb{R}$  between  $f(a)$  and  $f(b)$ , there exists  $x \in X$  such that  $f(x) = y$ .

**INTERLUDE: PATHS AND CURVES**

Let  $X$  be a metric space. A **path** or **curve** in  $X$  is a continuous function  $\gamma : [0, 1] \rightarrow X$ . We call  $\gamma(0)$  the **starting point** and  $\gamma(1)$  the **ending point**. We also say that  $\gamma$  is a path from  $\gamma(0)$  to  $\gamma(1)$ . A path  $\gamma$  with  $\gamma(0) = \gamma(1)$  is called **closed path** or a **loop**. If  $s : [a, b] \rightarrow [0, 1]$  is a bijective continuous function with continuous inverse we say that  $\gamma \circ s$  is a **re-parametrization** of  $\gamma$ . Furthermore, exactly one of the following happens

- either,  $s(a) = 0, s(b) = 1$ , then we say that  $s$  is **orientation-preserving**,
- or,  $s(a) = 1, s(b) = 0$ , then we say that  $s$  is **orientation-reversing**.

**DEFINITION 9.86:**

Let  $X$  be a metric space. We call  $E \subset X$  **path-connected** if, for every two points  $x, y \in E$ , there exists a path  $\gamma : [0, 1] \rightarrow E$  from  $x = \gamma(0)$  to  $y = \gamma(1)$ .

**LEMMA 9.87:**

Let  $X$  be a metric space. If  $E \subset X$  is path-connected then it is connected.

*Proof.* Suppose that  $E$  is disconnected in the topological sense, then there exist non-empty, disjoint, open sets  $U_1$  and  $U_2$  such that  $E \subset U_1 \cup U_2$ . Let  $x_1 \in U_1$  and  $x_2 \in U_2$ . If  $E$  were path-connected, there would exist a path  $\gamma : [0, 1] \rightarrow E$  from  $x_1$  to  $x_2$ . However, this implies that  $V_1 = \gamma^{-1}(U_1)$  and  $V_2 = \gamma^{-1}(U_2)$  are non-empty, disjoint open subsets of  $[0, 1]$  with  $V_1 \cup V_2 \supset [0, 1]$ ; a contradiction since  $[0, 1]$  is connected.  $\square$



EXERCISE 9.88. — Sketch the subset  $E \subset \mathbb{R}^2$  given by

$$E = \{0\} \times [-1, 1] \cup \{(t, \sin(\frac{1}{t})) \mid t > 0\}$$

and show that  $E$  is connected but not path-connected.

PROPOSITION 9.89:

*Let  $U \subset \mathbb{R}^n$  be an open subset. Then  $U$  is path-connected if and only if  $U$  is connected.*

*Proof.* If  $U$  is path-connected, then  $U$  is also connected according to Lemma 9.87. Now, assume  $U$  is connected, non-empty, and  $x_0 \in U$  is a fixed point. We define the set

$$G = \{x \in U \mid \text{there exists a path in } U \text{ from } x_0 \text{ to } x\}$$

and want to show that  $G = U$ . Since  $U$  is connected and  $G$  is non-empty, it suffices to show that both  $G$  and  $U \setminus G$  are open.

Let  $x \in G$  and  $\gamma : [0, 1] \rightarrow U$  be a path from  $x_0$  to  $x$ . Since  $U$  is open, there exists  $r > 0$  such that  $B(x, r) \subset U$ . For any  $y \in B(x, r)$ , the straight path  $t \mapsto (1-t)x + ty$ , connecting  $x$  and  $y$ , lies in  $U$ . Concatenating these paths yields the path

$$t \mapsto \begin{cases} \gamma(2t) & \text{if } 0 \leq t \leq \frac{1}{2} \\ (2-2t)x + (2t-1)y & \text{if } \frac{1}{2} < t \leq 1 \end{cases}$$

from  $x_0$  to  $y$ . Thus,  $y \in G$ , and since  $y$  was arbitrary, we have  $B(x, r) \subset G$ . This shows that  $G$  is open. Using a similar argument, we can show that  $U \setminus G$  is open. If  $x \in U \setminus G$  and  $r > 0$  with  $B(x, r) \subset U$ , then all points in  $B(x, r)$  are not in  $G$ . If  $y \in G \cap B(x, r)$ , a concatenation of paths as above would connect  $x$  to  $x_0$ . Therefore,  $B(x, r) \subset U \setminus G$ , and  $U \setminus G$  is open. Thus,  $G$  is closed. □

COROLLARY 9.90:  $\mathbb{R}^n$  AND ITS BALLS ARE CONNECTED

*For all  $n \geq 1$  and  $r > 0$ , the metric space  $\mathbb{R}^n$  (with the standard distance) and the subsets  $B(x, r)$  and  $\overline{B(x, r)}$  of  $\mathbb{R}^n$  are connected.*

## 9.3 Normed vector spaces

### 9.3.1 Definition of Normed Vector spaces

A **norm** on a real vector space  $V$  is a function from  $V$  to  $\mathbb{R}$  that assigns to each vector a non-negative number, informally its **length**. In general there are many different norms on a vector space, and we can use any of them to construct a distance and turn  $V$  into a metric space. A particularly interesting class of norms is obtained from **scalar products**. Many notions of this sections would carry out to vector spaces over  $\mathbb{C}$  with little modifications, but we will stick to real vector spaces.

#### DEFINITION 9.91: NORMED VECTOR SPACE

Let  $V$  be a vector space over  $\mathbb{R}$ . A **norm** on  $V$  is a mapping  $\|\cdot\| : V \rightarrow [0, \infty)$  that satisfies the following three properties.

- (1) (Definiteness) For all  $v \in V$ ,  $\|v\| = 0 \iff v = 0$ .
- (2) (Homogeneity) For all  $v \in V$  and all  $\alpha \in \mathbb{R}$ ,  $\|\alpha v\| = |\alpha| \|v\|$ .
- (3) (Triangle Inequality) For all  $v, w \in V$ ,  $\|v + w\| \leq \|v\| + \|w\|$ .

The pair  $(V, \|\cdot\|)$  is called a **normed vector space**.

7

EXAMPLE 9.92. — Let  $n \in \mathbb{N}$ . The **maximum norm** or **infinity norm**  $\|\cdot\|_\infty$ , and the **1-norm**  $\|\cdot\|_1$  on  $\mathbb{R}^n$  are defined by

$$\|v\|_\infty = \max\{|v_1|, |v_2|, \dots, |v_n|\} \quad \text{and} \quad \|v\|_1 = \sum_{j=1}^n |v_j|$$

for  $v = (v_1, \dots, v_n) \in \mathbb{R}^n$ . The properties of definiteness and homogeneity, as well as the triangle inequality, can be verified by exercise.

EXAMPLE 9.93. — If  $V$  is the vector space of continuous  $\mathbb{R}$ -valued functions on  $[0, 1]$ , we define analogously the 1-norm and the infinity norm as

$$\|f\|_1 = \int_0^1 |f| dx \quad \text{and} \quad \|f\|_\infty = \sup\{|f(x)| \mid x \in [0, 1]\}.$$

9.94. — From now on, in order to keep the notation, simple, we will denote the Euclidean norm of a vector  $x$  in  $\mathbb{R}^n$  by  $|x|$ , instead of  $\|x\|$ , that we will reserve for (less standard) norms. Notice that this notation does not create any ambiguity or collision with previously introduced notations. Indeed:

- If  $n = 1$  the Euclidean norm coincides with the absolute value.
- If  $n = 2$  and we are identifying the  $\mathbb{R}^2$  and  $\mathbb{C}$  via the usual map  $(x_1, x_2) \mapsto x_1 + ix_2$ , then Euclidean norm coincides with the complex absolute value.

9.95. — The Euclidean norm on  $\mathbb{R}^n$  holds a special position among all norms on  $\mathbb{R}^n$ . On  $\mathbb{R}^2$  or  $\mathbb{R}^3$ , it measures the “physical” length of vectors. However, many other norms  $\|\cdot\|$  confer  $\mathbb{R}^n$  the structure of normed vector space. A standard family of norms is given by

$$\|x\|_p := \left( \sum_{i=1}^n |x_i|^p \right)^{1/p},$$

where  $p \in [1, +\infty)$  is a given number. Notice that the Euclidean norm  $|x|$  corresponds to the  $p = 2$  case.

One can check (exercise) that for any given  $x \in \mathbb{R}^n$

$$\lim_{p \rightarrow +\infty} \|x\|_p = \max_{1 \leq i \leq n} |x_i|$$

is also a norm. That is why the maximum norm is commonly called infinity norm and denoted  $\|\cdot\|_\infty$ .

7

We immediately observe that Normed Vector Spaces are “automatically” Metric Spaces.

**LEMMA 9.96: NORMED VECTOR SPACES ARE METRIC SPACES**

Let  $V$  be a vector space over  $\mathbb{R}$  and  $\|\cdot\|$  be a norm on  $V$ . Define the function

$$d: V \times V \rightarrow [0, \infty), \quad d(v, w) := \|v - w\|,$$

then  $(V, d)$  is a metric space.

*Proof.* We check definiteness, symmetry, and the triangle inequality in the definition of a metric 9.3. For  $v, w \in V$ , we have  $d(v, w) = \|v - w\| \geq 0$ , and

$$d(v, w) = 0 \iff \|v - w\| = 0 \iff v - w = 0 \iff v = w$$

by the definiteness of the norm. Using homogeneity of the norm for  $\alpha = -1$ , we have for  $v, w \in V$ ,

$$d(v, w) = \|v - w\| = \|(-1)(v - w)\| = \|w - v\| = d(w, v)$$

thus establishing the symmetry of  $d$ . Finally, using the triangle inequality of the norm, we obtain

$$d(u, w) = \|u - w\| = \|(u - v) + (v - w)\| \leq \|u - v\| + \|v - w\| = d(u, v) + d(v, w)$$

for all  $u, v, w \in V$ . This shows the triangle inequality for  $d$ , so  $d$  is indeed a metric on  $V$ .  $\square$

We have seen that a normed space is naturally a metric space, now we check that the norm is indeed continuous. We show sequential continuity.

**LEMMA 9.97: THE NORM IS CONTINUOUS WITH RESPECT TO ITS OWN DISTANCE**

Let  $V$  be a  $\mathbb{R}$ -vector space, and let  $\|\cdot\|$  be a norm on  $V$ . Let  $(v_n)_{n=0}^{\infty}$  be a sequence in  $V$  converging with respect to the norm  $\|\cdot\|$  to a limit  $v \in V$ . Then,

$$\lim_{n \rightarrow \infty} \|v_n\| = \|v\|.$$

*Proof.* By definition,  $(v_n)_{n=0}^{\infty}$  converges to  $v$ , if and only if  $d(v_n, v) = \|v_n - v\|$  converges to 0. But using the triangle inequality:

$$-\|v_n - v\| \leq \|v_n\| - \|v\| \leq \|v_n - v\|.$$

Thus, if  $\|v_n - v\| \rightarrow 0$ , then necessarily  $\|v_n\| \rightarrow \|v\|$ .  $\square$

### 9.3.2 Inner Product Spaces are Normed Vector Spaces

**DEFINITION 9.98: INNER PRODUCT SPACE**

Let  $V$  be a vector space over  $\mathbb{R}$ . An **inner product** on  $V$  is a map

$$\langle -, - \rangle : V \times V \rightarrow \mathbb{R}$$

that satisfies the following properties for all  $u, v, w \in V$  and  $\alpha, \beta \in \mathbb{R}$ :

- (1) (Bilinearity)  $\langle \alpha u + \beta v, w \rangle = \alpha \langle u, w \rangle + \beta \langle v, w \rangle$ .
- (2) (Symmetry)  $\langle v, w \rangle = \langle w, v \rangle$ .
- (3) (Definiteness)  $\langle v, v \rangle \geq 0$  and  $\langle v, v \rangle = 0 \iff v = 0$ .

9.99. — An important example of an inner product is the **Euclidean inner product** or **standard inner product** on  $\mathbb{R}^n$ . It is given by

$$\langle -, - \rangle : V \times V \rightarrow \mathbb{R} \quad \langle v, w \rangle = \sum_{k=1}^n v_k w_k$$

for  $v = (v_1, \dots, v_n)$  and  $w = (w_1, \dots, w_n)$ . The proof of bilinearity and symmetry is left as an exercise. We verify definiteness. Let  $v = (v_1, \dots, v_n) \in \mathbb{R}^n$ . Then,

$$\langle v, v \rangle = \sum_{k=1}^n v_k v_k = \sum_{k=1}^n v_k^2 \geq 0$$

is a non-negative real number. If  $v = 0$ , then  $\langle v, v \rangle = 0$ . If  $\langle v, v \rangle = 0$ , then each term  $|v_k|^2$  must be zero, and thus  $v_k = 0$  for all  $k$ , implying  $v = 0$ .

**PROPOSITION 9.100: CAUCHY-SCHWARZ INEQUALITY**

Let  $V$  be a vector space over  $\mathbb{R}$ , let  $\langle \cdot, \cdot \rangle$  be an inner product on  $V$ , and let  $\|\cdot\| : V \rightarrow \mathbb{R}$  be given by  $\|v\| = \sqrt{\langle v, v \rangle}$ . Then the inequality holds

$$|\langle v, w \rangle| \leq \|v\| \|w\| \quad (9.3)$$

for all  $v, w \in V$ . Furthermore, equality in (9.3) holds if and only if  $v$  and  $w$  are linearly dependent.

*Proof.* Notice first that for all  $\alpha, \beta$  positive real numbers (9.3) holds iff and only if

$$|\langle \alpha v, \beta w \rangle| \leq \|\alpha v\| \|\beta w\| \quad (9.4)$$

Now  $v = 0$  or  $w = 0$ , then both sides of (9.3) are zero. So, putting  $\alpha = 1/\|v\|$  and  $\beta = 1/\|w\|$  we may assume without loss of generality that  $\|v\| = \|w\| = 1$ . Then:

$$\begin{aligned} 0 &\leq \|v - w\|^2 = \langle v - w, v - w \rangle = \langle v, v - w \rangle - \langle w, v - w \rangle \\ &= \langle v, v \rangle - \langle v, w \rangle - \langle w, v \rangle + \langle w, w \rangle = \|v\|^2 - 2\langle v, w \rangle + \|w\|^2 \\ &= 2 - 2|\langle v, w \rangle| = 2(\|v\| \|w\| - \langle v, w \rangle). \end{aligned}$$

□

We prove that the norm induced by an inner product is indeed a norm on  $V$ .

**COROLLARY 9.101:**

Let  $V$  be a vector space over  $\mathbb{R}$ , let  $\langle -, - \rangle$  be an inner product on  $V$ . The map defined by (9.5)

$$\|\cdot\| : V \rightarrow \mathbb{R}, \quad \|v\| = \sqrt{\langle v, v \rangle}$$

satisfies the triangular inequality and is a norm.

*Proof.* Definiteness and homogeneity follow directly from the definiteness and bilinearity of the inner product. We only need to prove the triangle inequality. Let  $v, w \in V$ . Using the Cauchy-Schwarz inequality, we have the estimate

$$\begin{aligned} \|v + w\|^2 &= \langle v + w, v + w \rangle = \|v\|^2 + \langle v, w \rangle + \langle w, v \rangle + \|w\|^2 \\ &= \|v\|^2 + 2\langle v, w \rangle + \|w\|^2 \leq \|v\|^2 + 2|\langle v, w \rangle| + \|w\|^2 \\ &\leq \|v\|^2 + 2\|v\| \|w\| + \|w\|^2 = (\|v\| + \|w\|)^2, \end{aligned}$$

which implies the desired result after taking the square root. □

9.102. — Let  $V$  be a vector space over  $\mathbb{R}$ . If  $\langle \cdot, \cdot \rangle$  is an inner product on  $V$ , we call the norm treated in Corollary 9.101

$$\| \cdot \| : V \rightarrow \mathbb{R}, \quad \|v\| = \sqrt{\langle v, v \rangle} \quad (9.5)$$

the **induced norm** by  $\langle \cdot, \cdot \rangle$ . In particular, from the Euclidean inner product on  $V = \mathbb{R}^n$ , we can define a norm on  $V = \mathbb{R}^n$ . The **Euclidean norm** on  $V = \mathbb{R}^n$  is given by

$$\|x\| = \sqrt{\langle x, x \rangle} = \sqrt{\sum_{k=1}^n |v_k|^2}$$

for all  $v = (v_1, \dots, v_n) \in \mathbb{R}^n$ .

### 9.3.3 Equivalence of norms in finite dimensional normed spaces

#### INTERLUDE: FINITE AND INFINITE DIMENSIONAL VECTOR SPACES

A **basis** for a vector space  $V$  is a subset  $\mathcal{B} = \{e_i\}_{i \in I} \subset V$  such that

- every  $v \in V$  can be written as  $v = \sum_{i \in I} v_i e_i$  for some coefficients  $\{v_i\}_{i \in I} \subset \mathbb{R}$ , only finitely many of which are nonzero (so that the previous sum always makes sense, it is not a series!)
- the coefficients  $\{v_i\}_{i \in I}$  are uniquely determined, in other words the following implication holds:

$$\sum_{i \in I} v_i e_i = 0, \text{ with finitely many non-zero } v_i \in \mathbb{R} \implies v_i = 0, \quad \forall i \in I.$$

For every vector space we can obtain a sequence of linearly independent vectors  $e_1, e_2, e_3, \dots$ . If this sequence necessarily stops, we obtain a finite basis and we say that the vector space is finite dimensional. If the sequence can be continued indefinitely we say that the vector space is infinite dimensional.

All bases of a finite dimensional vector space have the same number of vectors. This number is called the **dimension** of  $V$ .

EXERCISE 9.103. — Show that the space of polynomials with real coefficients  $\mathbb{R}[x]$  is an infinite dimensional vector space and find a basis.

9.104. — If  $V$  has finite dimension  $n \in \mathbb{N}$  and we fix a basis  $\mathcal{B} = \{e_i\}_{1 \leq i \leq n}$  then map

$$v_{\mathcal{B}}(x_1, \dots, x_n) \mapsto x_1 e_1 + \dots + x_n e_n$$

is a (vector space) isomorphism and allows to treat  $V$  as  $\mathbb{R}^n$  for most practical tasks. In particular, if  $(V, \|\cdot\|)$  is a normed space then  $\iota_{\mathcal{B}}$  induces a norm in  $\mathbb{R}^n$  defined as

$$\|x\| = \|\iota_{\mathcal{B}}(x)\|.$$

This is a motivation to prove results for  $\mathbb{R}^n$  equipped with norms different from the Euclidean one. Indeed, all the concepts and results that can be stated for general norms will automatically hold in “abstract” finite dimensional normed vector space. We see next an important instance of this.

**DEFINITION 9.105: EQUIVALENT (I.E., COMPARABLE) NORMS**

Let  $V$  be a vector space over  $\mathbb{R}$  and  $\|\cdot\|_1$  and  $\|\cdot\|_2$  be two norms on  $V$ . We call  $\|\cdot\|_1$  and  $\|\cdot\|_2$  **equivalent** if there are constants  $A > 0$  and  $B > 0$  such that

$$\|v\|_1 \leq A\|v\|_2 \quad \text{and} \quad \|v\|_2 \leq B\|v\|_1 \quad \text{for all } v \in V.$$

**EXAMPLE 9.106.** — Let  $n \in \mathbb{N}$ . The 1-norm  $\|\cdot\|_1$  and the maximum norm  $\|\cdot\|_{\infty}$  given in Example 9.92 are equivalent, as the inequalities

$$\|v\|_{\infty} \leq \|v\|_1 \quad \text{and} \quad \|v\|_1 \leq n\|v\|_{\infty}$$

hold for all  $v \in \mathbb{R}^n$ . As we will show in Theorem 9.107, all norms on a finite-dimensional vector space over  $\mathbb{R}$  are equivalent to each other. This is not the case for infinite-dimensional vector spaces. For example, the norms given in 9.92 on the space of continuous functions on  $[0, 1]$  are not equivalent.

**THEOREM 9.107: ALL NORMS ARE EQUIVALENT**

*All norms on  $\mathbb{R}^n$  are equivalent (i.e., comparable).*

*Proof.* Let  $\|\cdot\|$  be the Euclidean norm on  $\mathbb{R}^n$ , and let  $\|\cdot\|'$  denote another norm on  $\mathbb{R}^n$ . We show that  $\|\cdot\|$  and  $\|\cdot\|'$  are equivalent, which proves the Theorem.

Let  $e_1, \dots, e_n$  denote the standard basis of  $\mathbb{R}^n$ , and let  $A = \max\{\|e_1\|', \|e_2\|', \dots, \|e_n\|'\}$ . For any vector  $v = x_1e_1 + \dots + x_n e_n \in V$ , we have

$$\|v\|' \leq \sum_{k=1}^n |x_k| \cdot \|e_k\|' \leq A \cdot \sum_{k=1}^n |x_k| \leq A\sqrt{n} \sqrt{\sum_{k=1}^n |x_k|^2} = A\sqrt{n}\|v\|$$

which already shows one of the two required inequalities. For the second one, consider unit sphere (with respect to the Euclidean norm)

$$S := \{v \in \mathbb{R}^n : \|v\| = 1\}$$

and the real number  $B = \inf\{\|v\|' \mid v \in S\}$ . Let us show that  $B > 0$ . Indeed, by definition of “inf” there exists a sequence  $(v_n)_{n=0}^\infty$  in  $S$  such that the sequence  $\|v_n\| \rightarrow B$ . Since  $(v_n)_{n=0}^\infty$  belongs to a closed and bounded set, it contains a convergent subsequence by the Heine-Borel theorem, so by replacing  $(v_n)_{n=0}^\infty$  with such a subsequence, we can ensure that  $(v_n)_{n=0}^\infty$  converges to some  $v \in S$  with respect to the Euclidean norm. But since

$$\|w - v_n\|' \leq A\sqrt{n}\|w - v_n\|$$

we deduce that the sequence  $(v_n)_{n=0}^\infty$  also converges to  $v$  with respect to the norm  $\|\cdot\|'$ . Thus, by Lemma 9.97,

$$\|v\| = 1 \quad \text{and} \quad \|v\|' = B$$

and, in particular,  $v \neq 0$  and  $B > 0$ , otherwise  $\|\cdot\|'$  would not be a norm, giving zero length to a nonzero vector.

Finally any vector  $x \neq 0$  in  $\mathbb{R}^n$ ,  $\frac{x}{\|x\|}$  is an element of  $S$ , and it satisfies

$$\frac{\|x\|'}{\|x\|} = \left\| \frac{1}{\|x\|} x \right\|' \geq B$$

Thus,  $\|x\| \leq B^{-1}\|x\|'$  for all  $x \neq 0$  in  $\mathbb{R}^n$ . Since this inequality is also trivially true for  $x = 0$ , it is shown that the norms  $\|\cdot\|$  and  $\|\cdot\|'$  are equivalent.  $\square$

9.108. — Notice that as a simple consequence of Theorem 9.107, in every finite dimensional vector space  $V$  all norms are equivalent. Indeed, we can fix a basis  $\mathcal{B} = \{e_1, \dots, e_n\}$  of  $V$  use the inclusion map  $\iota_{\mathcal{B}}$  defined in 9.104 to export the result for  $\mathbb{R}^n$  to  $V$ .

**COROLLARY 9.109: ALL NORMS LEAD TO SAME TOPOLOGICAL PROPERTIES**

*Topological properties, such as compactness or connectedness, of a subset  $E \subseteq \mathbb{R}^n$ , remain unchanged regardless of the norm chosen to define the metric in  $\mathbb{R}^n$ .*

*Proof.* Since any two given norms in  $\mathbb{R}^n$  are equivalent the associated distances  $d_1, d_2$  satisfy, for some constant  $C$ ,  $C^{-1}d_1(x, y) \leq d_2(x, y) \leq Cd_1(x, y)$  for all  $x, y \in \mathbb{R}^n$ . Hence, it is readily shown that  $U$  is open with respect to  $d_1$  if and only if it is open with respect to  $d_2$ .

As a consequence,  $(\mathbb{R}^n, d_1)$  and  $(\mathbb{R}^n, d_2)$  have the same open sets and thus identical topological properties (all properties that we could define only in terms of the topologies, i.e., the collections of open sets)  $\square$



---

## Chapter 10

# Multidimensional Differentiation

In this chapter, we extend the concept of derivatives to functions defined on open subsets  $U \subset \mathbb{R}^n$  and taking values in  $\mathbb{R}^m$ , for general positive integers  $n, m$ .

From now on, we consider always  $\mathbb{R}^n$  to be equipped with the Euclidean norm (as in Definition 9.1). To keep the notation lighter we will write  $|x|$  instead of  $\|x\|$ , to refer to the Euclidean norm of  $x \in \mathbb{R}^n$ . Notice that this does not create any ambiguity with the complex absolute value in  $\mathbb{R}$  or the complex absolute value in  $\mathbb{C}$  because they both coincide with the Euclidean norm. Therefore the notion of convergence for sequences  $(x_k)_{k \geq 0}$  in  $\mathbb{R}^n$  is the one given by the Euclidean distance

$$x_k \rightarrow y \quad \iff \quad \lim_{k \rightarrow \infty} |x_k - y| = 0$$

In this section, vectors will be regarded as column vectors to adhere to the standard linear algebra conventions on matrix multiplication. That is vectors  $x$  in  $\mathbb{R}^n$  and  $y$  in  $\mathbb{R}^m$  will be regarded as:

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \in \mathbb{R}^n, \quad y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix} \in \mathbb{R}^m.$$

Similarly, if  $U \subset \mathbb{R}^n$  is an open set and  $f : U \rightarrow \mathbb{R}^m$  a function, its components will be denoted always in column form

$$x \mapsto f(x) = \begin{bmatrix} f_1(x) \\ f_2(x) \\ \vdots \\ f_m(x) \end{bmatrix}.$$

Notice that for  $i = 1, \dots, m$  we have  $f_i : U \rightarrow \mathbb{R}$ .

**DEFINITION 10.1: LIMITS OF FUNCTIONS**

If  $f : U \rightarrow \mathbb{R}^m$  is some function and  $x_o \in U$ , we say that the limit of  $f(x)$ , as  $x$  converges to  $x_o$ , is  $y_o \in \mathbb{R}^m$  and write

$$\lim_{x \rightarrow x_o} f(x) = y_o$$

if for every sequence  $(x_k)_{k \geq 0} \subset U$  converging to  $x_o$  we have

$$f(x_k) \rightarrow y_o, \quad \text{or, in other words,} \quad |f(x_k) - y_o| \rightarrow 0$$

as  $k \rightarrow \infty$ .

We will also make use the “little o and big O notations”, updated to the context of  $\mathbb{R}^n$ . If  $U \subset \mathbb{R}^n$  is an open set,  $x_o \in U$ , and  $f : U \rightarrow \mathbb{R}^m$  and  $g : U \rightarrow \mathbb{R}$  are two functions, we write  $f(x) = o(|g(x)|)$  as  $x \rightarrow x_o$  if

$$\lim_{x \rightarrow x_o} \frac{|f(x)|}{|g(x)|} = 0.$$

Also, we write  $f(x) = O(|g(x)|)$

$$\limsup_{x \rightarrow x_o} \frac{|f(x)|}{|g(x)|} < \infty.$$

8

We remark that this useful notation has to be used with care, since **it is not symmetric**:

$$f(x) = O(|x_1|) \Rightarrow f(x) = O(|x|), \quad \text{but} \quad f(x) = O(|x|) \not\Rightarrow f(x) = O(|x_1|)!$$

In other words

$$f(x) = O(|x_1|) \text{ and } f(x) = O(|x|), \quad \text{but} \quad O(|x_1|) \neq O(|x|).$$

## 10.1 The Differential

### 10.1.1 The derivative for real valued functions

The derivative of a real-variable function  $f : \mathbb{R} \rightarrow \mathbb{R}$  at some point  $x_o \in \mathbb{R}$  has various equivalent interpretations. Of course each of these interpretations provides the same number  $f'(x_o)$ , but the “meaning” we attach to this number is slightly different in each case:

- **Slope of tangent line to the graph.** We look at the graph of  $f$ , i.e. the curve  $\{y = f(x)\} \subset \mathbb{R}^2$  and write the tangent line to the graph at  $(x_o, f(x_o))$  in the form  $y = ax + b$ . Then we have  $a = f'(x_o)$ .
- **Zoom-in limit.** When we zoom-in more and more to inspect the graph of  $f$  around a point  $(x_o, y_o = f(x_o)) \in \mathbb{R}^2$  we see in the limit the line  $y = f'(x_o)x$ . More precisely, in

each compact interval  $[-C, C]$ , the functions

$$f_r : x \mapsto \frac{f(x_0 + rx) - f(x_0)}{r},$$

converge uniformly, as  $r \downarrow 0$ , to the linear function  $x \mapsto f'(x_0)x$ .

- **Coefficient in infinitesimal linear approximation.** The linear function that best approximates  $f$  in tiny intervals  $x_0$ . More precisely, assume we want to approximate  $f(x) \sim a + bx$  around  $x \sim x_0$ , for some real numbers  $a, b$ . Then, choosing  $b = f'(x_0)$ , we get approximation error of size  $o(|x - x_0|)$  as  $x \rightarrow x_0$ ;
- **Stretching Factor.** Look at a short interval  $I$  around  $x_0$  and the corresponding interval  $f(I)$  around  $f(x_0)$ . These two intervals are related through a “stretching factor” which tends to be  $f'(x_0)$  as  $I$  is taken shorter. Look [here](#) for an animation.

To generalize derivatives to functions  $f: \mathbb{R}^n \rightarrow \mathbb{R}^m$  we will start from the third point of view. As it will be clear later on, all these viewpoints (conveniently reinterpreted) are valid also in the several variables case.

### 10.1.2 Definition of differential

#### DEFINITION 10.2: DIFFERENTIAL OF A FUNCTION

Let  $U \subset \mathbb{R}^n$  be open and  $f: U \rightarrow \mathbb{R}^m$  be a function. Then  $f$  is called **differentiable** at  $x_0 \in U$  if there exists a *linear map*  $L: \mathbb{R}^n \rightarrow \mathbb{R}^m$  such that

$$\lim_{x \rightarrow 0} \frac{|f(x_0 + x) - f(x_0) - L(x)|}{|x|} = 0$$

holds. If such  $L$  exists it is unique (exercise) and it is called the **differential** of  $f$  at the point  $x_0$ , and we denote it as

$$L = Df_{x_0}.$$

The function  $f$  is called **differentiable in  $U$**  if it is differentiable at every point in  $U$ .

EXERCISE 10.3 (Uniqueness of the differential). — Suppose that  $L$  and  $L'$  are two differentials of  $f$  at  $x_0$ . Show that  $\lim_{x \rightarrow 0} L(x)/|x| = 0$  for all  $x$  and deduce that  $(L - L')v = 0$  for all  $v$ , and hence  $(L - L') = 0$ .

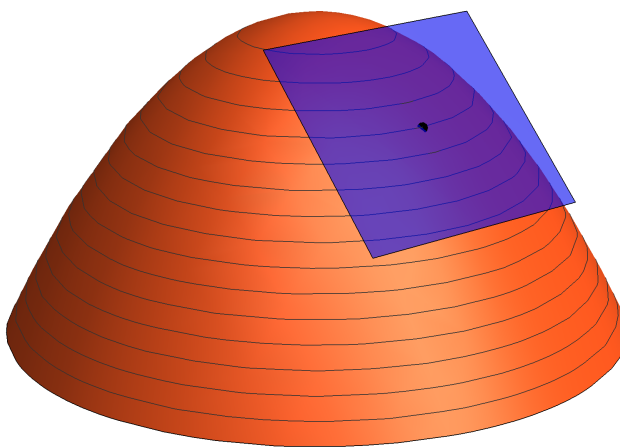
10.4. — If  $f: U \rightarrow \mathbb{R}^m$  is differentiable at the point  $x_0 \in U$ , with differential  $L = Df_{x_0}$ , we can express  $f$  as

$$f(x_0 + x) = f(x_0) + L(x) + R(x)$$

Here we recognize the affine-linear approximation  $x \mapsto f(x_0) + L(x)$  to  $f$ , and a remainder term  $R(x)$ , for which, according to the definition of the total derivative,  $R(x) = o(|x|)$  holds

for  $x \rightarrow 0$ . Another notation often found in the literature for differential of  $f$  at the point  $x_0$  is  $df_{x_0}$ .

10.5. — For functions  $f : \mathbb{R} \rightarrow \mathbb{R}$  the derivative  $f'(x_0)$  is a real number. Notice that in this case  $L(y) = Df_{x_0}(y) = f'(x_0)y$ .



8

Figure 10.1: For a function  $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ , the best infinitesimal linear approximation corresponds to the tangent plane of the graph in  $\mathbb{R}^3$ .

**Applet 10.6 (Tangent Plane).** *As shown in the above image, we depict the tangent planes for the graphs of two functions  $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ . Additionally, we visualize the partial derivatives and directional derivatives in Definition 10.8. Is there a directional derivative that vanishes at every point?*

**LEMMA 10.7: DIFFERENTIAL COMPONENT-WISE**

Let  $U \subset \mathbb{R}^n$  be open, and let  $f : U \rightarrow \mathbb{R}^m$  be a function and let  $f_j : U \rightarrow \mathbb{R}$  denote  $j$ -th component of  $f$ . Then,  $f$  is differentiable at  $x_0 \in U$  if and only if, for all  $1 \leq j \leq m$ , the component  $f_j$  is differentiable at  $x_0$ . In this case, we have, for all  $v \in \mathbb{R}^n$ ,

$$(Df_{x_0}(v))_j = (Df_j)_{x_0}(v).$$

*Proof.* Assume that  $f_j$  is differentiable at  $x_0$  for every  $j$ . There exists a linear function  $L_j : \mathbb{R}^n \rightarrow \mathbb{R}$  and a remainder term  $R_j : \mathbb{R}^n \rightarrow \mathbb{R}$  such that

$$f_j(x_0 + x) = f_j(x_0) + L_j(x) + R_j(x)$$

with  $R_j(x) = o(|x|)$  for  $x \rightarrow 0$ . We can summarize

$$f(x_o + x) = \begin{pmatrix} f_1(x_o + x) \\ \vdots \\ f_m(x_o + x) \end{pmatrix} = \begin{pmatrix} f_1(x_o) \\ \vdots \\ f_m(x_o) \end{pmatrix} + \begin{pmatrix} L_1(x) \\ \vdots \\ L_m(x) \end{pmatrix} + \begin{pmatrix} R_1(x) \\ \vdots \\ R_m(x) \end{pmatrix}$$

In summary, it can be written as  $f(x_o + x) = f(x_o) + L(x) + R(x)$ . In this expression,  $L$  is linear, and it holds that  $R_j(x) = o(|x|)$  for  $x \rightarrow 0$  and we conclude using Lemma 9.24. Thus,  $f$  is differentiable, and the claimed formula for  $Df(x_o)$  holds. Similarly, if  $f$  is differentiable at the point  $x_o$ , then each component is differentiable and the claimed formula follows.  $\square$

#### DEFINITION 10.8: DIRECTIONAL DERIVATIVE

Let  $U \subset \mathbb{R}^n$  be an open subset,  $x_o \in U$ ,  $v \in \mathbb{R}^n$  and  $f : U \rightarrow \mathbb{R}^m$ . The **directional derivative of  $f$  in the direction  $v$  at  $x_o$**  is

$$\partial_v f(x_o) := \left. \frac{d}{ds} \right|_{s=0} f(x_o + sv) = \lim_{s \rightarrow 0} \frac{f(x_o + sv) - f(x_o)}{s} \in \mathbb{R}^m,$$

provided that the limit exists. If  $v = e_j$ , for some  $j \in \{1, \dots, n\}$ , we denote:

$$\partial_{e_j} f(x_o) = D_j f(x_o) = \partial_j f(x_o) = \frac{\partial f(x_o)}{\partial x_j} = \begin{pmatrix} \partial_j f_1(x_o) \\ \partial_j f_2(x_o) \\ \vdots \\ \partial_j f_n(x_o) \end{pmatrix}$$

Of course, if the partial derivative in the  $j$ -th coordinate exists at every point in  $U$ , we obtain a function  $\partial_j f : U \rightarrow \mathbb{R}^m$ , which we call the  **$j^{\text{th}}$  directional derivative of  $f$** .

10.9. — The partial derivative and the directional derivative along any vector are the derivatives with respect to one of the independent variables, considering all other variables as constants (they are ‘frozen’). For example, for the function  $f : \mathbb{R}^3 \rightarrow \mathbb{R}$  given by  $f(x, y, z) = x(y^2 + \sin(z))$ , the partial derivatives with respect to all coordinate directions are given by

$$\begin{aligned} \partial_x f(x, y, z) &= y^2 + \sin(z) \\ \partial_y f(x, y, z) &= 2xy \\ \partial_z f(x, y, z) &= x \cos(z) \end{aligned}$$

for all  $(x, y, z) \in \mathbb{R}^3$ , as we can apply all known rules from Analysis I. If the total derivative exists, we can connect it with partial derivatives and derivatives along arbitrary vectors using the following proposition.

**PROPOSITION 10.10: DIFFERENTIABLE IMPLIES LINEAR DIRECTIONAL DERIVATIVES**

Let  $U \subset \mathbb{R}^n$  be open and let  $f: U \rightarrow \mathbb{R}^m$  be differentiable at  $x_o \in U$ . Then, for each  $v \in \mathbb{R}^n$ , the derivative of  $f$  in the direction  $v$  exists, and we have

$$\partial_v f(x_o) = Df_{x_o}(v) \in \mathbb{R}^m.$$

In particular,  $\partial_{\alpha v + \beta w} f(x_o) = \alpha \partial_v f(x_o) + \beta \partial_w f(x_o)$ , for all  $\alpha, \beta \in \mathbb{R}$  and  $v, w \in \mathbb{R}^n$ .

*Proof.* Assuming the total derivative  $Df(x_o)$  exists, according to the definition of the derivative,  $f(x_o + h) = f(x_o) + Df(x_o)(h) + o(|h|)$  holds for  $h \rightarrow 0$ . Choosing  $h = sv$  for  $s \rightarrow 0$  and  $v \in \mathbb{R}^n$ , we get

$$\partial_v f(x_o) = \lim_{s \rightarrow 0} \frac{f(x_o + sv) - f(x_o)}{s} = \lim_{s \rightarrow 0} (Df(x_o)(v) + o(1)) = Df(x_o)(v)$$

which concludes the proof. □

**EXERCISE 10.11.** — Let  $U \subset \mathbb{R}^n$  be open, and let  $f_1, f_2: U \rightarrow \mathbb{R}$  be functions. Assume that  $f_1$  and  $f_2$  are differentiable at  $x_o \in U$ . Show that the functions  $f_1 + f_2$  and  $f_1 f_2$  are differentiable at  $x_o$  and that

$$\begin{aligned} D(f_1 + f_2)(x_o) &= Df_1(x_o) + Df_2(x_o) \\ D(f_1 f_2)(x_o) &= f_1(x_o) Df_2(x_o) + f_2(x_o) Df_1(x_o) \end{aligned}$$

Formulate and prove analogous statements for directional derivatives in the direction of a fixed vector  $v \in \mathbb{R}^n$ .

**THEOREM 10.12: SUFFICIENT CONDITION FOR DIFFERENTIABILITY**

Let  $U \subset \mathbb{R}^n$  be open, and  $f: U \rightarrow \mathbb{R}^m$  be a function. If for every  $j \in \{1, \dots, m\}$  the partial derivative  $\partial_j f$  exists on the entire  $U$  and defines a continuous function, then  $f$  is differentiable on the entire  $U$ .

*Proof.* Due to Lemma 10.7, we can assume  $m = 1$ . Let's fix  $x_o \in U$ , and we need to show that  $f$  is differentiable at  $x_o$ . By replacing  $f$  with  $x \mapsto f(x + x_o) - f(x_o)$ , we can also assume that  $x_o = 0$  and  $f(0) = 0$ . For  $x = (x_1, \dots, x_n) \in U$ , we then have

$$\begin{aligned} f(x) &= f(x_1, x_2, x_3, \dots, x_n) - f(0, x_2, x_3, \dots, x_n) \\ &\quad + f(0, x_2, x_3, \dots, x_n) - f(0, 0, x_3, \dots, x_n) \\ &\quad + f(0, 0, x_3, \dots, x_n) - \dots \\ &\quad + \dots - f(0, 0, \dots, 0, x_n) \\ &\quad + f(0, 0, \dots, 0, x_n) - f(0, 0, \dots, 0, 0). \end{aligned}$$

The function  $[0, x_j] \rightarrow \mathbb{R}$  defined by  $t \mapsto f(0, 0, \dots, 0, t, x_{j+1}, \dots, x_n)$  is continuously differentiable by hypothesis. Its derivative is given by the  $j$ -th partial derivative of  $f$ . Therefore, by the Mean Value Theorem, there exists an intermediate point  $\xi_j \in [0, x_j]$  such that

$$\partial_j f(0, \dots, 0, \xi_j, x_{j+1}, \dots, x_n) x_j = f(0, \dots, 0, x_j, x_{j+1}, \dots, x_n) - f(0, \dots, 0, 0, x_{j+1}, \dots, x_n)$$

holds. For any choice of such intermediate points  $\xi_j \in [0, x_j]$ , we obtain

$$\begin{aligned} f(x) &= \partial_1 f(\xi_1, x_2, x_3, \dots, x_n) x_1 \\ &\quad + \partial_2 f(0, \xi_2, x_3, \dots, x_n) x_2 \\ &\quad + \dots \\ &\quad + \partial_n f(0, 0, \dots, 0, \xi_n) x_n. \end{aligned}$$

8

To show that the linear function  $L : (v_1, \dots, v_n) \mapsto \partial_1 f(0) v_1 + \dots + \partial_n f(0) v_n$  is the derivative  $Df(0)$ , we need to estimate the difference  $R(x) := f(x) - L(x) = f(0 + x) - f(0) - L(x)$ .

$$\begin{aligned} R(x) &= (\partial_1 f(\xi_1, x_2, x_3, \dots, x_n) - \partial_1 f(0)) x_1 \\ &\quad + (\partial_2 f(0, \xi_2, x_3, \dots, x_n) - \partial_2 f(0)) x_2 \\ &\quad + \dots \\ &\quad + (\partial_n f(0, 0, \dots, 0, \xi_n) - \partial_n f(0)) x_n \end{aligned}$$

According to the assumptions of the theorem and because  $\frac{|x_j|}{|x|} \leq 1$  for all  $x \in \mathbb{R}^n$ , the asymptotics

$$\lim_{x \rightarrow 0} \frac{R(x)}{|x|} = 0$$

holds, demonstrating that  $f$  is differentiable at  $x_0 = 0$  with the derivative  $Df(0) = L$ .  $\square$

The following exercise shows that the mere existence of partial derivatives of a function  $f$ , without the continuity assumption, does not necessarily imply that the function is differentiable.

EXERCISE 10.13. — Consider the function  $f : \mathbb{R}^2 \rightarrow \mathbb{R}$  defined by

9

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

for  $(x, y) \in \mathbb{R}^2$ . Show that the partial derivatives  $\partial_x f$  and  $\partial_y f$  exist everywhere in  $\mathbb{R}^2$ , but  $f$  is not differentiable at  $(0, 0)$ .

DEFINITION 10.14:  $C^1$  FUNCTIONS

We call a function  $f : U \rightarrow \mathbb{R}^m$  on an open subset  $U \subset \mathbb{R}^n$  **continuously differentiable** if all the partial derivatives  $\partial_i f$ ,  $i = 1, \dots, n$  exist and continuous in  $U$ . We will write in this case  $f \in C^1(U, \mathbb{R}^m)$  and often  $f \in C^1(U)$  when  $m = 1$ .

10.15. — Notice that  $f \in C^1(U, \mathbb{R}^m)$  if and only if the map  $Df : U \rightarrow \text{Hom}(\mathbb{R}^n, \mathbb{R}^m)$  mapping  $x \mapsto Df_x$  is continuous.

10.16. — From Proposition 10.10, it follows in particular that the total derivative (when it exists)  $Df(x_o)$  is uniquely determined by the partial derivatives. Specifically, for  $v = a_1 e_1 + \dots + a_n e_n \in \mathbb{R}^n$ ,

$$Df_{x_o}(v) = \sum_{i=1}^n a_i Df_{x_o}(e_i) = \sum_{i=1}^n a_i \partial_i f(x_o).$$

The  $m \times n$  matrix of the linear map  $Df_{x_o} : \mathbb{R}^n \rightarrow \mathbb{R}^m$  is thus given with respect to the canonical bases by the matrix  $n \times m$  matrix

$$(\partial_1 f(x_o), \partial_2 f(x_o), \dots, \partial_n f(x_o)) = \begin{pmatrix} \frac{\partial f_1}{\partial x_1}(x_o) & \frac{\partial f_1}{\partial x_2}(x_o) & \cdots & \frac{\partial f_1}{\partial x_n}(x_o) \\ \frac{\partial f_2}{\partial x_1}(x_o) & \frac{\partial f_2}{\partial x_2}(x_o) & \cdots & \frac{\partial f_2}{\partial x_n}(x_o) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1}(x_o) & \frac{\partial f_m}{\partial x_2}(x_o) & \cdots & \frac{\partial f_m}{\partial x_n}(x_o) \end{pmatrix}$$

This matrix is referred to as the **Jacobian matrix** of  $f$  evaluated at the point  $x_o$ , commonly denoted by  $Jf(x_o)$ . It also is standard to denote the Jacobian matrix as  $Df(x_o)$ ; however, to prevent confusion with  $Df_{x_o}$  —linear map versus its matrix representation—, by now we will use the notation  $Jf(x_o)$ . While the difference between linear map and a matrix in canonical basis might appear to be a minor detail, it will become more significant in future applications such as in Differential Geometry or Physics.

Notice that, given  $U \subset \mathbb{R}^n$  open,  $f \in C^1(U, \mathbb{R}^m)$  if and only if the maps

$$Jf : U \rightarrow \text{Mat}_{m,n}(\mathbb{R}) \cong \mathbb{R}^{m \times n}$$

defined by  $x \mapsto Jf(x)$  is a continuous .

EXAMPLE 10.17. — The Jacobian matrix  $Jf(x, y, z)$  of the function  $f : \mathbb{R}^3 \rightarrow \mathbb{R}^4$  given by  $f(x, y, z) = (x^2 + y, y^2 + z, x + z^2, xyz)$  is computed as follows:



$$Jf(x, y, z) = \begin{pmatrix} \frac{\partial f_1}{\partial x} & \frac{\partial f_1}{\partial y} & \frac{\partial f_1}{\partial z} \\ \frac{\partial f_2}{\partial x} & \frac{\partial f_2}{\partial y} & \frac{\partial f_2}{\partial z} \\ \frac{\partial f_3}{\partial x} & \frac{\partial f_3}{\partial y} & \frac{\partial f_3}{\partial z} \\ \frac{\partial f_4}{\partial x} & \frac{\partial f_4}{\partial y} & \frac{\partial f_4}{\partial z} \end{pmatrix} = \begin{pmatrix} 2x & 1 & 0 \\ 0 & 2y & 1 \\ 1 & 0 & 2z \\ yz & xz & xy \end{pmatrix}$$

Where each entry of the matrix is the partial derivative of the function's output component with respect to an input variable.

EXAMPLE 10.18. — Let  $f : \mathbb{R}^2 \rightarrow \mathbb{R}^2$  be defined by  $f(x, y) = (x^2 - \cos(xy), y^4 - \exp(x))$ . The Jacobian matrix of  $f$  at  $(x, y)$  is then

$$\begin{pmatrix} 2x + \sin(xy)y & \sin(xy)x \\ -\exp(x) & 4y^3 \end{pmatrix},$$

which is continuous as a function of  $(x, y) \in \mathbb{R}^2$ .

9

### 10.1.3 The Chain Rule

#### INTERLUDE: THE HILBERT-SCHMIDT NORM OF A LINEAR MAP (OR MATRIX)

Let  $M : \mathbb{R}^n \rightarrow \mathbb{R}^m$  is be linear map. We define the Hilbert-Schmidt norm of  $M$  (or of the matrix  $(M_j^i)$  as

$$\|M\|_2 = \sqrt{\text{trace}(M^T M)} = \sqrt{\sum_{i=1}^n |Me_i|^2}.$$

Notice that

$$|My| \leq \|M\|_2 |y| \quad \text{for all } y \in \mathbb{R}^n$$

Writing  $y = \sum_{i=1}^n y_i e_i$  we obtain:

$$\begin{aligned} |My|^2 &= |M \sum_{i=1}^n y_i e_i|^2 = |\sum_{i=1}^n y_i Me_i|^2 \\ &\leq \left( \sum_{i=1}^n |y_i| |Me_i| \right)^2 \leq |y|^2 \sum_{i=1}^n |Me_i|^2 = \|M\|_2^2 |y|^2. \end{aligned}$$

Similarly, the Hilbert-Schmidt norm of the matrix in  $\text{Mat}_{m,n}(\mathbb{R})$  is defined as the norm of the associated linear map  $\mathbb{R}^n \rightarrow \mathbb{R}^m$ .

**THEOREM 10.19: CHAIN RULE (POINTWISE VERSION)**

Let  $k, m, n \geq 1$ , and let  $U \subset \mathbb{R}^n$  and  $V \subset \mathbb{R}^m$  be open. If  $f : U \rightarrow V$  is differentiable at  $x_o$  and  $g : V \rightarrow \mathbb{R}^k$  is differentiable at  $f(x_o)$ , then  $g \circ f$  is differentiable at  $x_o$ , and the differential of  $(g \circ f)$  at  $x_o$  is given by

$$D(g \circ f)_{x_o} = Dg_{f(x_o)} \circ Df_{x_o} \quad (10.1)$$

In other words, at the level of matrices

$$J(g \circ f)(x_o) = Jg(f(x_o)) \cdot Jf(x_o)$$

or equivalently (components):

$$\frac{\partial (g \circ f)_j}{\partial x_i}(x_o) = \sum_{\ell=1}^m \frac{\partial g_j}{\partial y_\ell}(f(x_o)) \frac{\partial f_\ell}{\partial x_i}(x_o)$$

9

*Intuitive idea of proof.* Since  $f$  is differentiable at  $x_o$  we have

$$f(x) - f(x_o) \approx L(x - x_o),$$

for some  $L$  linear. Also, if  $g$  is differentiable at  $y_o = f(x_o)$  we have

$$g(y) - g(y_o) \approx M(y - y_o),$$

for some  $M$  linear. Therefore, putting  $y = f(x)$  we obtain

$$g(f(x)) - g(f(x_o)) = g(y) - g(y_o) \approx M(y - y_o) = M(f(x) - f(x_o)) \approx M(L(x - x_o)).$$

Since the composition of linear maps is linear this suggests that  $g \circ f$  is differentiable at  $x_o$  and that the differential of the composition is the composition of differentials. In the next proof we will make such argument work rigorously.  $\square$

*Proof.* By the definition of differentiability of  $f$  at  $x_o$  and  $g$  at  $y_o = f(x_o)$ , we have

$$f(x_o + x) = f(x_o) + L(x) + R(x) \quad \text{and} \quad g(y_o + y) = g(y_o) + M(y) + S(y)$$

with  $L = Df_{x_o}$ ,  $R(x) = o(|x|)$  as  $x \rightarrow 0$ ,  $M = Dg_{y_o}$ , and  $S(y) = o(|y|)$  as  $y \rightarrow 0$ .

Now, for  $x \in \mathbb{R}^n$  with  $x$  sufficiently close to 0 (i.e., with  $|x|$  small enough), put  $y = f(x_\circ + x) - f(x_\circ) = Df_{x_\circ}(x) + R(x)$ . We obtain the equation

$$\begin{aligned} g(f(x_\circ + x)) &= g(y_\circ + y) = g(y_\circ) + M(y) + S(y) \\ &= g(f(x_\circ)) + M(L(x)) + \underbrace{M(R(x)) + S(L(x) + R(x))}_{T(x)} \end{aligned}$$

and we want to show that  $T(x) = o(|x|)$  as  $x \rightarrow 0$ . Since  $R(x) = o(|x|)$  as  $x \rightarrow 0$ , and  $M$  is linear  $|M(R(x))| \leq \|M\|_2 |R(x)| = o(|x|)$  as  $x \rightarrow 0$ .

It remains to show that  $S(L(x) + R(x)) = o(|x|)$  as  $x \rightarrow 0$ . For this, notice that  $|L(x) + R(x)| \leq \|L\|_2 |x| + |R(x)| \leq C|x|$  for  $C = \|L\|_2 + 1$  for all  $x$  sufficiently close to zero (since  $|R(x)| = o(|x|) \leq |x|$  as  $x \rightarrow 0$ ). Hence, using  $|S(y)| = o(|y|)$  as  $y \rightarrow 0$ ,

$$\lim_{x \rightarrow 0} \frac{S(L(x) + R(x))}{|x|} = \lim_{x \rightarrow 0} \frac{S(L(x) + R(x))}{|L(x) + R(x)|} \frac{|L(x) + R(x)|}{|x|} \leq C \lim_{x \rightarrow 0} \frac{S(x)}{|x|} = 0.$$

The previous computation implicitly assumes that, along the sequence  $x_k \rightarrow 0$  chosen to test the limit as  $x \rightarrow 0$ ,  $S(L(x_k) + R(x_k))$  does not vanish. However, the argument is easily adapted to the case that it could possibly vanish. The details are left as an exercise to the reader.

Thus, we conclude the differentiability of  $g \circ f$  at  $x_\circ$  and the equation (10.1).  $\square$

9

#### COROLLARY 10.20: CHAIN RULE

Let  $k, m, n \geq 1$ , and let  $U \subset \mathbb{R}^n$  and  $V \subset \mathbb{R}^m$  be open. If  $f \in C^1(U, \mathbb{R}^m)$ ,  $g \in C^1(V)$  and  $f(U) \subset V$ , then  $g \circ f \in C^1(U)$  and

$$\partial_i(g \circ f) = \sum_{j=1}^m (\partial_j g) \circ f \partial_i f_j. \quad (10.2)$$

EXAMPLE 10.21. — (How to use in practice the chain rule). Consider the functions

$$g(x, y) := e^{x+2y}, \quad f(x, y) = (\sin(x), \ln(1+y))^T.$$

We aim to compute  $\frac{\partial(g \circ f)}{\partial x}$  and  $\frac{\partial(g \circ f)}{\partial y}$ . In practice, we can operate in this way:

1. We call the two real variables of the map  $g$  with different names, so  $g(u, v) = e^{u+2v}$  and we set

$$u = f_1(x, y) = \sin(x), \quad v = f_2(x, y) = \ln(y+1).$$

2. Compute the necessary partial derivatives:

$$\frac{\partial g}{\partial u} = e^{u+2v}, \quad \frac{\partial g}{\partial v} = 2e^{u+2v}, \quad \frac{\partial u}{\partial x} = \cos(x), \quad \frac{\partial u}{\partial y} = 0, \quad \frac{\partial v}{\partial x} = 0, \quad \frac{\partial v}{\partial y} = \frac{1}{y+1}.$$

3. Apply the chain rule, and replace  $u, v$  with their expression in terms of  $x, y$ :

$$\begin{aligned}\frac{\partial g}{\partial x} &= \frac{\partial g}{\partial u} \cdot \frac{\partial u}{\partial x} + \frac{\partial g}{\partial v} \cdot \frac{\partial v}{\partial x} = e^{u+2v} \cos(x) + 2e^{u+v} \cdot 0 = e^{\sin(x)} \cdot \cos(x)(1+y)^2, \\ \frac{\partial g}{\partial y} &= \frac{\partial g}{\partial u} \cdot \frac{\partial u}{\partial y} + \frac{\partial g}{\partial v} \cdot \frac{\partial v}{\partial y} = e^{u+2v} \cdot 0 + 2 \frac{e^{u+2v}}{1+y} = 2(1+y)e^{\sin(x)}.\end{aligned}$$

Notice that in this particular example the chain rule is not clearly more “economic” than differentiate directly

$$(g \circ f)(x, y) = (1+y)^2 e^{\sin(x)},$$

double check that the result is the same.

EXERCISE 10.22 (Euler’s identity for homogeneous functions). — Assume  $f \in C^1(\mathbb{R}^n \setminus \{0\})$  is positively homogeneous of degree  $\lambda \in \mathbb{R}$ , that is to say

$$f(rx) = r^\lambda f(x) \text{ for all } r > 0, x \neq 0.$$

Show that  $\sum_{j=1}^n x_j \partial_j f(x) = \lambda f(x)$ .

EXAMPLE 10.23. — Let’s consider the special case  $n = 1$  for the chain rule. Suppose  $I \subset \mathbb{R}$  is an open interval, and  $\gamma : I \rightarrow V \subset \mathbb{R}^m$  is a differentiable function with values in an open subset  $V \subset \mathbb{R}^m$ . Further, let  $f : V \rightarrow \mathbb{R}^k$  be differentiable. Then, the chain rule implies that  $f \circ \gamma$  is differentiable, and the formula

$$(f \circ \gamma)'(t) = Df_{\gamma(t)}(\gamma'(t))$$

holds for all  $t \in I$ .

### 10.1.4 The Mean Value Theorem

We formulate a generalization of the Mean Value Theorem for real-valued differentiable functions on an open set  $U \subset \mathbb{R}^n$ . To do this, we consider a given function  $f$  along a straight segment in the open set.

#### THEOREM 10.24: MEAN VALUE THEOREM

Let  $U \subset \mathbb{R}^n$  be open, and  $f : U \rightarrow \mathbb{R}$  be differentiable. Let  $x_o \in U$  and  $h \in \mathbb{R}^n$  such that  $x_o + th \in U$  for all  $t \in [0, 1]$ . Then, there exists  $t \in (0, 1)$  such that for  $\xi = x_o + th$ , the equation

$$f(x_o + h) - f(x_o) = Df_\xi(h) = \partial_h f(\xi)$$

is satisfied.

*Proof.* The derivative of the straight path  $\gamma : t \mapsto x_o + th$  for fixed  $x_o, h \in \mathbb{R}^n$  is given by  $\gamma'(t) = h$ . Therefore, the function  $g = f \circ \gamma : [0, 1] \rightarrow \mathbb{R}$  satisfies all the conditions of the one-dimensional Mean Value Theorem due to the chain rule in Theorem 10.19. Hence, there exists  $t \in (0, 1)$  with  $g(1) - g(0) = g'(t) = Df(x_o + th)(h)$  according to the chain rule, and thus

$$f(x_o + h) - f(x_o) = g(1) - g(0) = g'(t) = Df(\xi)(h)$$

for  $\xi = x_o + th$ . □

#### DEFINITION 10.25: LOCAL LIPSCHITZ CONTINUITY

A function  $f : X \rightarrow Y$  between metric spaces  $X, Y$  is called **locally Lipschitz continuous** if, for every  $x_o \in X$ , there exists  $\epsilon > 0$  such that  $f|_{B(x_o, \epsilon)}$  is Lipschitz continuous.

#### COROLLARY 10.26: DIFFERENTIABILITY VS LIPSCHITZ CONTINUITY

Let  $U \subset \mathbb{R}^n$  be open and let  $f \in C^1(U, \mathbb{R}^m)$ . Then,  $f$  is locally Lipschitz continuous. If  $U$  is additionally **convex** (i.e., for every two points  $x, y \in U$  the segment  $\{tx + (1-t)y \mid t \in [0, 1]\}$  joining them is contained in  $U$ ), and the Jacobi matrix  $Jf$  is bounded in the whole  $U$ , then  $f$  is Lipschitz continuous.

*Proof.* It suffices to consider the case  $m = 1$ . First, assume first that  $U$  is convex and the Jacobi matrix is bounded in  $U$ . That is there exists  $M \geq 0$  such that  $\|Df_\xi\|_2 = \|Jf(\xi)\|_2 \leq M$  for all  $\xi \in U$ . From the mean value Theorem 10.24, it follows for  $x, y \in U$

$$|f(x) - f(y)| = |Df(\xi)(x - y)| \leq M|x - y|$$

for some  $\xi \in U$ , since  $U$  is convex and thus contains the straight segment between  $x$  and  $y$ . This proves the second statement in the corollary.

The first statement follows from the second applied to the ball  $U_0 = B(x_o, \epsilon)$  where  $\epsilon > 0$  is chosen such that  $\overline{B(x_o, \epsilon)} \subset U$ . Indeed,  $U_0$  is convex, and the mapping  $\xi \mapsto Df(\xi)$  is a continuous function on the compact set  $\overline{U_0} \subset U$ , implying the boundedness of the differential on  $U_0$ . □

#### COROLLARY 10.27: SOLUTIONS OF $Df = 0$

Let  $U \subset \mathbb{R}^n$  be open and let  $f : U \rightarrow \mathbb{R}^m$  be differentiable with  $Df(x) = 0$  for all  $x \in U$ . Then,  $f$  is constant on each connected component of  $U$ . In particular, if  $U$  is connected then  $f$  is constant.

*Proof.* It suffices to consider the case  $m = 1$ . Assuming that  $U$  is non-empty, we choose  $x_o \in U$  and consider the subset

$$U' = \{x \in U \mid f(x) = f(x_o)\}$$

of  $U$ .

On the one hand, since  $f$  is continuous,

$$U \setminus U' = f^{-1}((-\infty, f(x_0)) \cup (f(x_0), \infty))$$

is open (it is the pre-image of an open set by a continuous function).

9

On other hand by the mean value theorem, it follows that  $U'$  is open: Indeed, for  $x \in U'$ , there exists  $\epsilon > 0$  such that  $B(x, \epsilon) \subset U$ , and since every point  $y \in B(x, \epsilon)$  can be connected by a straight path to  $x$ , it follows from Theorem 10.24 that  $f(y) = f(x) = f(x_0)$ . Thus,  $y \in U'$ , and since  $y \in B(x, \epsilon)$  was arbitrary,  $B(x, \epsilon) \subset U'$ .

Finally, since  $U$  is connected, and  $U'$  and  $U \setminus U'$  are two disjoint open sets covering  $U$  we have  $U' = U$  (as  $x_0 \in U'$  so it can not be empty). Hence, the corollary follows.  $\square$

## 10.2 Higher Derivatives

### 10.2.1 Definition and basic properties

Recall that for functions  $f : \mathbb{R} \rightarrow \mathbb{R}$  we defined second and higher order derivatives recursively as follows  $f^{(k+1)} = (f^{(k)})'$ ,  $k \geq 0$ , where  $f^{(0)} = f$ . Next we will introduce higher order derivatives for function  $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ .

#### DEFINITION 10.28: $C^k$ FUNCTIONS

Let  $U \subset \mathbb{R}^n$  be open,  $f : U \rightarrow \mathbb{R}^m$  be a function, and  $k > 1$ . We say that  $f$  belongs to  $C^k(U, \mathbb{R}^m)$  if for all  $i = 1, \dots, n$  the partial derivative  $\partial_i f$  exists and belongs to  $C^{k-1}(U, \mathbb{R}^m)$ .

Notice that this recursive definition makes sense because  $C^1(U, \mathbb{R}^m)$  have been previously defined.

When  $f \in C^k(U, \mathbb{R}^m)$  we say that  $f$  is ' $k$ -times continuously differentiable' or that ' $f$  is of class  $C^k$ '.

Given  $i_1, \dots, i_k$  in  $\{1, \dots, n\}$ , the  $k$ -th partial derivative  $\partial_{i_1} \partial_{i_2} \cdots \partial_{i_k} f$  is recursively defined as  $\partial_{i_1}(\partial_{i_2} \cdots \partial_{i_k} f)$  and it is a continuous function on  $U$ .

Consistently with the previous definition, it is convenient to define  $C^0(U, \mathbb{R}^m)$  as the class of continuous functions  $U \rightarrow \mathbb{R}^m$ .

We define  $C^\infty(U, \mathbb{R}^m)$  as the class of functions that belongs to  $C^k(U, \mathbb{R}^m)$  for all  $k \geq 1$ . Similarly to the  $k = 1$  case,  $C^k(U, \mathbb{R})$  is denoted  $C^k(U)$  (also when  $k$  is replaced by  $\infty$ )

#### PROPOSITION 10.29: HIGHER REGULARITY OF SUMS, PRODUCTS AND COMPOSITIONS

Let  $U \subset \mathbb{R}^n$  open,  $f, g \in C^k(U, \mathbb{R})$ ,  $h \in C^k(V, \mathbb{R}^n)$  with  $V \subset \mathbb{R}^m$  open with  $h(V) \subset U$ , then

(1)  $f + g, f \cdot g$  are of class  $C^k$

(2)  $f \circ h$  is of class  $C^k$

*Proof.* Property (1) easily follows by the recursive (i.e., inductive) definition of  $C^k$  functions. Indeed, if  $f, g$  are of class  $C^k$ , for some  $k \geq 1$  then, for all  $i$

$$\partial_i(f + g) = \underbrace{\partial_i f}_{C^{k-1}} + \underbrace{\partial_i g}_{C^{k-1}} \quad \text{and} \quad \partial_i(fg) = \underbrace{\partial_i f}_{C^{k-1}} \cdot \underbrace{g}_{C^k} + \underbrace{f}_{C^k} \cdot \underbrace{\partial_i g}_{C^{k-1}} \quad (10.3)$$

This establishes first the case  $k = 1$  and then  $k > 1$  by induction over  $k$ .

To establish (2) we proceed by induction over  $k$ . On the one hand, since the composition of continuous is continuous the case  $k = 0$  follows. On the other hand, if  $k \geq 1$  by the chain

rule (Corollary 10.20) we have that  $f \circ h$  is differentiable and

$$\partial_j(f \circ h) = \sum_{\ell=1}^m \underbrace{\partial_\ell f \circ h}_{C^{k-1}} \underbrace{\partial_j h_\ell}_{C^{k-1}}.$$

Thus, using (1) we obtain that  $\partial_i(h \circ f)$  is of class  $C^{k-1}$ , for  $1 \leq j \leq n$ , so  $f \circ h \in C^k$ .  $\square$

## 10.2.2 Schwartz's Theorem and Multi-indexes notation

### THEOREM 10.30: SCHWARZ'S THEOREM

Let  $U \subset \mathbb{R}^n$  be open and let  $f \in C^2(U, \mathbb{R}^m)$ , then

$$\partial_j \partial_i f = \partial_i \partial_j f, \quad \text{for all } i, j \in \{1, \dots, n\}.$$

*Intuitive idea of proof.* For a smooth enough function  $f : \mathbb{R}^2 \rightarrow \mathbb{R}$  we should have, for  $h \in \mathbb{R}$  with very small modulus,

$$\partial_1 f(x_1, x_2) \approx \frac{f(x_1 + h, x_2) - f(x_1, x_2)}{h}$$

and then

$$\begin{aligned} \partial_2(\partial_1 f(x_1, x_2)) &\approx \frac{\partial_1 f(x_1, x_2 + h) - \partial_1 f(x_1, x_2)}{h} \\ &\approx \frac{1}{h} \left( \frac{f(x_1 + h, x_2 + h) - f(x_1, x_2 + h)}{h} - \frac{f(x_1 + h, x_2) - f(x_1, x_2)}{h} \right) \\ &= \frac{f(x_1 + h, x_2 + h) - f(x_1, x_2 + h) - f(x_1 + h, x_2) + f(x_1, x_2)}{h^2} \end{aligned}$$

Similarly,

$$\partial_2 f(x_1, x_2) \approx \frac{f(x_1, x_2 + h) - f(x_1, x_2)}{h}$$

and then

$$\begin{aligned} \partial_1(\partial_2 f(x_1, x_2)) &\approx \frac{\partial_2 f(x_1, x_2 + h) - \partial_2 f(x_1, x_2)}{h} \\ &\approx \frac{1}{h} \left( \frac{f(x_1 + h, x_2 + h) - f(x_1 + h, x_2)}{h} - \frac{f(x_1, x_2 + h) - f(x_1, x_2)}{h} \right) \\ &= \frac{f(x_1 + h, x_2 + h) - f(x_1, x_2 + h) - f(x_1 + h, x_2) + f(x_1, x_2)}{h^2} \end{aligned}$$

Therefore,

$$\partial_2 \partial_1 f(x_1, x_2) \approx \partial_1 \partial_2 f(x_1, x_2)$$

and we need to show that the approximation errors in the previous infinitesimally small as  $h$  goes to zero. Notice that this is nontrivial as we are dividing by  $h^2$ , which is 'doubly small'. The proof below will make this heuristic rigorous.  $\square$



*Proof.* If  $i = j$  there is nothing to prove, so by symmetry of the statement with respect to  $i, j$  we may assume  $i < j$ . We want to show that for all  $y = (y_1, y_2, \dots, y_n) \in U$ ,  $\partial_j \partial_i f(y) = \partial_i \partial_j f(y)$ .

It suffices to consider the case  $n = 2$ ,  $m = 1$  and  $i = 1$ ,  $j = 2$ . Indeed, for general  $n \geq 2$  and  $m$  and fixed  $i_1, i_2 \in \{1, \dots, n\}$  with  $i_1 \neq i_2$  we may apply the considered special case to

$$\tilde{f}(x_1, x_2) = f_k(y_1, \dots, \overbrace{x_1}^i, \dots, \overbrace{x_2}^j, \dots, y_n) \quad \text{for fixed } k \in \{1, \dots, m\}.$$

Thus, we assume without loss of generality  $n = 2$  and  $m = 1$ . For  $x = (x_1, x_2) \in U$  and a sufficiently small  $h > 0$ , we define a function  $F$  by

$$F(h) = f(x_1 + h, x_2 + h) - f(x_1 + h, x_2) - f(x_1, x_2 + h) + f(x_1, x_2).$$

Furthermore, for a sufficiently small but fixed  $h > 0$ , we consider the differentiable function  $\varphi : [0, 1] \rightarrow \mathbb{R}$  given by  $\varphi(t) = f(x_1 + th, x_2 + h) - f(x_1 + th, x_2)$  and obtain

$$F(h) = \varphi(1) - \varphi(0) = \varphi'(\xi_1) = (\partial_1 f(x_1 + \xi_1 h, x_2 + h) - \partial_1 f(x_1 + \xi_1 h, x_2))h$$

for some  $\xi_1 \in (0, 1)$  by the one-dimensional Mean Value Theorem ??.

10

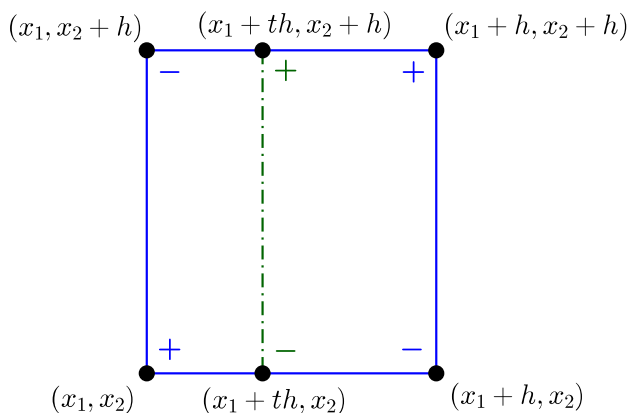


Figure 10.2: The function  $h \mapsto F(h)$  is a signed sum of function values of  $f$  at the corners of a square (here marked by a solid line). The function  $t \mapsto \varphi(t)$  corresponds to the difference of function values on a vertical segment through the square.

Applying the one-dimensional Mean Value Theorem again to  $\psi : [0, 1] \rightarrow \mathbb{R}$  given by  $\psi(t) = \partial_1 f(x_1 + \xi_1 h, x_2 + th)$  along with the chain rule, we obtain

$$F(h) = (\partial_1 f(x_1 + \xi_1 h, x_2 + h) - \partial_1 f(x_1 + \xi_1 h, x_2))h = \partial_2 \partial_1 f(x_1 + \xi_1 h, x_2 + \xi_2 h)h^2$$

for some intermediate point  $\xi_2 \in (0, 1)$ .

Similarly, now defining  $\tilde{\varphi}(t) = f(x_1 + h, x_2 + th) - f(x_1, x_2 + th)$  and  $\tilde{\psi}(t) = \partial_1 f(x_1 + th, x_2 + \tilde{\xi}_2 h)$  we obtain

$$F(h) = \partial_1 \partial_2 f(x_1 + \tilde{\xi}_1 h, x_2 + \tilde{\xi}_2 h) h^2.$$

for suitable  $\tilde{\xi}_1, \tilde{\xi}_2 \in (0, 1)$ . Dividing by  $h^2 > 0$ , we obtain

$$\partial_2 \partial_1 f(x_1 + \xi_1 h, x_2 + \xi_2 h) = \partial_1 \partial_2 f(x_1 + \tilde{\xi}_1 h, x_2 + \tilde{\xi}_2 h).$$

Since  $\xi_1, \xi_2, \tilde{\xi}_1, \tilde{\xi}_2 \in (0, 1)$ , the points  $(\xi_1 h, \xi_2 h)$  and  $(\tilde{\xi}_1 h, \tilde{\xi}_2 h)$  tend to  $(0, 0)$  as  $h$  tends to 0. Therefore, due to the continuity of both partial derivatives, we conclude  $\partial_2 \partial_1 f(x) = \partial_1 \partial_2 f(x)$ .  $\square$

As a consequence of Schwartz's Theorem higher order derivatives are also independent of the order in which we take the partial derivatives

10

**COROLLARY 10.31: SCHWARZ IN  $C^k$**

Let  $U \subset \mathbb{R}^n$  open and  $f \in C^k(U)$ . Then for every  $i_1, i_2, \dots, i_k \in \{1, \dots, n\}$  and for every permutation  $\sigma : \{1, 2, \dots, k\} \rightarrow \{1, 2, \dots, k\}$  the derivative we have

$$\partial_{i_1} \partial_{i_2} \cdots \partial_{i_k} f = \partial_{i_{\sigma(1)}} \partial_{i_{\sigma(2)}} \cdots \partial_{i_{\sigma(k)}} f,$$

*Proof.* It is a straightforward consequence from Schwartz's Theorem, e.g. using that any permutation of  $\{1, 2, \dots, k\}$  can be obtained as a suitable composition the  $k - 1$  transpositions  $\{(1, 2), (2, 3), \dots, (k - 1, k)\}$ . It is left as an exercise to the reader to give a proof by induction over  $k$ .  $\square$

Thanks to Schwartz's Theorem, multi-indexes are useful to express higher order derivatives.

## INTERLUDE: MULTI-INDECES AND POLYNOMIALS OF SEVERAL VARIABLES

Any  $\alpha \in \mathbb{N}^n$  is called a **multi-index**, the length of a multi-index  $\alpha$  is defined as

$$|\alpha| := \alpha_1 + \dots + \alpha_n.$$

We say that  $\beta \leq \alpha$  if  $\beta_i \leq \alpha_i$  for all  $i = 1, \dots, n$ , and define the factorial

$$\alpha! := \alpha_1! \dots \alpha_n! \quad (\text{recall that } 0! = 1).$$

A polynomial of  $n$  variables  $X_1, \dots, X_n$  of degree  $k$  with real coefficients can be uniquely (and compactly) expressed as

$$\sum_{\alpha \in \mathbb{N}^n, |\alpha| \leq k} c_\alpha X^\alpha := \sum_{|\alpha| \leq k} c_{(\alpha_1, \dots, \alpha_n)} X_1^{\alpha_1} \dots X_n^{\alpha_n},$$

where  $c_{(\alpha_1, \dots, \alpha_n)} \in \mathbb{R}$ .

Many combinatorial formulas are simple when expressed in multi-index notation, such as the multinomial formula:

$$(X_1 + X_2 + \dots + X_n)^k = \sum_{|\alpha|=k} \frac{k!}{\alpha!} X^\alpha.$$

Thanks to Corollary 10.31 multiindex notation is also very useful to express partial derivatives. Indeed, if  $f \in C^k$  and  $|\alpha| \leq k$  we define

$$\partial^\alpha f := \partial_1^{\alpha_1} \partial_2^{\alpha_2} \dots \partial_n^{\alpha_n} f.$$

EXERCISE 10.32. — Prove the identity

$$\frac{n^k}{k!} = \sum_{\alpha \in \mathbb{N}^n, |\alpha|=k} \frac{1}{\alpha!}.$$

## 10.2.3 Multidimensional Taylor Approximation

## THEOREM 10.33: TAYLOR'S THEOREM

Let  $U \subset \mathbb{R}^n$  be open and let  $f \in C^{k+1}(U)$ ,  $k \geq 0$ . Let  $x_0 \in U$  and  $h \in \mathbb{R}^n$  such that  $x_0 + th \in U$  for all  $t \in [0, 1]$ . Then, we have

$$f(x_0 + h) = \sum_{\alpha \in \mathbb{N}^n, |\alpha| \leq k} \partial^\alpha f(x_0) \frac{h^\alpha}{\alpha!} + R_{k+1}f(x_0, h)$$

where the reminder is given by

$$R_{k+1}f(x_0, h) := \int_0^1 (k+1)(1-t)^k \sum_{\alpha \in \mathbb{N}^n, |\alpha|=k+1} \partial^\alpha f(x_0 + th) \frac{h^\alpha}{\alpha!} dt = O(|h|^{k+1}).$$

*Proof.* Since  $U$  is open, there exists  $\varepsilon > 0$  such that  $x + th \in U$  for all  $t \in (-\varepsilon, 1 + \varepsilon)$ . We apply the one-dimensional Taylor approximation to  $\varphi : (-\varepsilon, 1 + \varepsilon) \rightarrow \mathbb{R}$  given by  $\varphi(t) = f(x + th)$ . According to Taylor's Theorem, we obtain the Taylor approximation around 0 at 1

$$\varphi(1) = \sum_{m=0}^k \frac{\varphi^{(m)}(0)}{m!} + \int_0^1 \varphi^{(k+1)}(t) \frac{(1-t)^k}{k!} dt. \quad (10.4)$$

Applying the chain rule in Theorem 10.19 to  $\varphi$ , we get for  $t \in (-\varepsilon, 1 + \varepsilon)$  the derivatives

$$\varphi'(t) = \sum_{i=1}^n \partial_i f(x_0 + th) h_i = \sum_{|\alpha|=1} \partial^\alpha f(x_0 + th) h^\alpha.$$

Let us show that, for all  $m \leq k + 1$ :

$$\varphi^{(m)}(t) = m! \sum_{|\alpha|=m} \partial^\alpha f(x_0 + th) \frac{h^\alpha}{\alpha!},$$

Indeed, using the chain rule and induction over  $m \geq 1$ , we have

$$\begin{aligned} \varphi^{(m+1)}(t) &= \frac{d}{dt} \varphi^{(m+1)}(t) = \frac{d}{dt} \left( m! \sum_{|\alpha|=m} \partial^\alpha f(x_0 + th) \frac{h^\alpha}{\alpha!} \right) \\ &= m! \left( \sum_{|\alpha|=m} \sum_{i=1}^n \partial_i \partial^\alpha f(x_0 + th) \frac{h^\alpha h_i}{\alpha!} \right) \\ &= m! \left( \sum_{|\beta|=m+1} \partial^\beta f(x_0 + th) h^\beta \sum_{1 \leq i \leq n, \beta_i \geq 1} \frac{\beta_i}{\beta!} \right). \end{aligned}$$

For the last equality we have used the following counting argument: If  $\alpha$  is a multiindex with  $|\alpha| = m$  then  $\partial_i \partial^\alpha f = \partial^\beta f$  and  $h_i h^\alpha = h^\beta$ , where  $\beta$  is a multiindex with  $|\beta| = m + 1$ .

10

11

Namely:

$$(\beta_1, \dots, \beta_i, \dots, \beta_n) = (\alpha_1, \dots, \alpha_i + 1, \dots, \alpha_n). \quad (10.5)$$

In particular we have

$$\frac{1}{\alpha!} = \frac{\alpha_i + 1}{\alpha_1! \dots (\alpha_i + 1)! \dots \alpha_n!} = \frac{\beta_i}{\beta!} \quad (10.6)$$

Notice that different  $\alpha$ 's will give the same  $\beta$  for different  $i$ 's. More precisely, for all  $\beta$  with  $|\beta| = m + 1$  and  $i \in \{1, 2, \dots, n\}$  such that  $\beta_i \geq 1$  we have  $\partial_i \partial^\alpha f = \partial^\beta f$  and  $h_i h^\alpha = h^\beta$  for exactly one  $i$ : with  $\alpha$  and  $\beta$  satisfying (10.5) and (10.6). Hence,

$$\sum_{|\alpha|=m} \sum_{i=1}^n \frac{\partial_i \partial^\alpha f h_i h^\alpha}{\alpha!} = \sum_{|\beta|=m+1} \sum_{1 \leq i \leq n, \beta_i \geq 1} \frac{\beta_i \partial^\beta f h^\beta}{\beta!}.$$

Substituting this into (10.4), we obtain the theorem.  $\square$

**COROLLARY 10.34: POLYNOMIAL EXPANSIONS DETERMINE HIGHER DERIVATIVES**

Let  $x_0 \in U$ ,  $f \in C^{k+1}(U)$  and  $P(x)$  a polynomial of degree  $k \geq 0$ . Assume that

$$|f(x_0 + h) - P(h)| = o(|h|^k) \quad \text{as } h \rightarrow 0.$$

Then for all  $|\alpha| \leq k$ ,  $\partial^\alpha f(x_0) = \partial^\alpha P(0)$ .

11

*Proof.* By Taylor's theorem we immediately get

$$\left| P(h) - \sum_{\alpha \in \mathbb{N}^n, |\alpha| \leq k} \partial^\alpha f(x_0) \frac{h^\alpha}{\alpha!} \right| = o(|h|^k),$$

but two polynomials of degree  $k$  whose difference is  $o(|h|^k)$  must have exactly the same coefficients (exercise).  $\square$

This Corollary can be useful to compute Taylor polynomials for explicit functions, without having to care about factorials etc.

**EXAMPLE 10.35.** — Let us compute the Taylor polynomial of degree 2 of  $\sqrt{1+x-y^2}$  around the origin. From Analysis I, you know that

$$\sqrt{1+t} = 1 + \frac{t}{2} - \frac{t^2}{8} + O(t^3), \quad t \rightarrow 0.$$

Plugging in  $t = x - y^2$  we find

$$\begin{aligned} \sqrt{1+x-y^2} &= 1 + \frac{x-y^2}{2} - \frac{(x-y^2)^2}{8} + O((x-y^2)^3) \\ &= 1 + \frac{1}{2}x - \frac{1}{8}x^2 - \frac{1}{2}y^2 + \frac{1}{4}xy^2 - \frac{1}{8}y^4 + O((x-y^2)^3). \end{aligned}$$

Now we want  $(x, y) \rightarrow (0, 0)$ , and a reminder which is  $o(r^2)$  where  $r := \sqrt{x^2 + y^2}$ . Observing that

$$\frac{1}{4}xy^2 = O(r^3), \quad \frac{1}{8}y^4 = O(r^4), \quad O((x - y^2)^3) = O(r^3), \quad \text{as } r \downarrow 0,$$

we find out expansion

$$\sqrt{1 + x - y^2} = 1 + \frac{1}{2}x - \frac{1}{8}x^2 - \frac{1}{2}y^2 + O(r^3), \quad \text{as } r \downarrow 0.$$

EXERCISE 10.36. — Compute the Taylor polynomials up to the quadratic order at  $(0, 0)$  of the following functions in two variables

$$\sin(xy), \quad \sqrt{1 + x + y^2}, \quad \exp \arctan(x - y), \quad \frac{1}{1 - x^2 - y^2}, \dots$$

Don't use the general formula!

EXERCISE 10.37. — Prove the Taylor expansion of the determinant close to the identity is

$$\det(I + tX) = 1 + t \operatorname{tr}(X) + \frac{t^2}{2} (\operatorname{tr}(X)^2 - \operatorname{tr}(X^2)) + O(t^3),$$

and that the one of the inverse matrix function is

$$(I + tX)^{-1} = 1 - tX + t^2 X^2 + O(t^3).$$

**Applet 10.38** ([Taylor Approximation](#)). We observe how the first, second, or third-order Taylor approximations approximate the function  $f(x, y) = \sin(x) \cos(y) + 2$ .

### 10.2.4 Real analytic functions of several variables

#### DEFINITION 10.39: REAL ANALYTIC FUNCTIONS

Let  $U \subset \mathbb{R}^n$  be an open set. We say that  $f \in C^\infty(U)$  is analytic if for every  $x_0 \in U$  there exist  $\varrho > 0$  and  $C$  such that

$$\sup_{B_\varrho(x_0)} |\partial^\alpha f| \leq C |\alpha|! (n\varrho)^{-|\alpha|} \quad (10.7)$$

holds for every multiindex  $\alpha$  (of arbitrarily large order).

Thanks to Taylor's theorem, a function is analytic if and only if it can be written as power series around any point in its domain of definition.

**THEOREM 10.40: ANALYTIC EXPANSION**

Let  $U \subset \mathbb{R}^n$  be an open set and  $f : U \rightarrow \mathbb{R}$  be analytic.

Given  $x_0 \in U$ , let  $\varrho$  and  $C$  be as in Definition 10.39. Consider the power series:

$$f_{x_0,k}(h) := \sum_{\alpha \in \mathbb{N}^n, |\alpha| \leq k} \partial^\alpha f(x_0) \frac{h^\alpha}{\alpha!}.$$

with  $k \rightarrow \infty$ .

Then, for all  $r \in (0, \varrho)$  the series is absolutely convergent the following sense: for all  $k < \ell$

$$\sup_{h \in B_r(0)} \sum_{k \leq |\alpha| \leq \ell} \left| \partial^\alpha f(x_0) \frac{h^\alpha}{\alpha!} \right| \leq \frac{C(r/\varrho)^k}{1 - (r/\varrho)}. \quad (10.8)$$

In particular the series converges for  $h \in B_r(0)$ :  $f_{x_0}(h) := \lim_{k \rightarrow \infty} f_{x_0,k}(h)$  is well-defined for all  $h \in B_r(0)$  and

$$\sup_{h \in B_r(0)} |f_{x_0,k}(h) - f_{x_0}(h)| \leq \frac{C(r/\varrho)^k}{1 - (r/\varrho)}.$$

Moreover the series represents  $f$  around  $x_0$ :

$$f(x_0 + h) = f_{x_0}(h) \quad \text{for all } h \in B_r(0).$$

*Proof.* Noticing  $|h^\alpha| = h_1^{\alpha_1} \cdots h_n^{\alpha_n} \leq |h|^{|\alpha|}$  and using (10.7) we obtain, for all  $k \geq 0$ :

$$\sum_{|\alpha|=k} \left| \partial^\alpha f(x_0) \frac{h^\alpha}{\alpha!} \right| \leq \sum_{|\alpha|=k} C |h|^k (n\varrho)^{-k} \frac{k!}{|\alpha|}.$$

Using the multinomial type identity

$$n^k = (1 + 1 + \cdots + 1)^k = \sum_{|\alpha|=k} \frac{k!}{|\alpha|}$$

we obtain then that for  $|h| < r$

$$\sum_{|\alpha|=k} \left| \partial^\alpha f(x_0) \frac{h^\alpha}{\alpha!} \right| \leq C(r/\varrho)^k.$$

From here (10.8) follows summing the geometric series.

To prove the representation formula we estimate similarly the remainder in Taylor's theorem. For  $|h| < r$  we have:

$$\begin{aligned}
|R_{k+1}f(x_0, h)| &= \left| \int_0^1 (k+1)(1-t)^k \sum_{|\alpha|=k+1} \partial^\alpha f(x_0 + th) \frac{h^\alpha}{\alpha!} dt \right| \\
&\leq \int_0^1 (k+1)(1-t)^k \sum_{|\alpha|=k+1} \left| \partial^\alpha f(x_0 + th) \frac{h^\alpha}{\alpha!} \right| dt \\
&\leq \int_0^1 (k+1)(1-t)^k C(r/\varrho)^k dt \\
&= C(r/\varrho)^k.
\end{aligned}$$

Therefore, for  $h \in B_r(0)$

$$|f(x_0 + h) - f_{x_0}(h)| \leq |f(x_0 + h) - f_{x_0, k}(h)| + |f_{x_0, k}(h) - f_{x_0}(h)| \leq C(r/\varrho)^k + \frac{C(r/\varrho)^k}{1 - (r/\varrho)}.$$

Sending  $k \rightarrow \infty$  we conclude  $f(x_0 + h) = f_{x_0}(h)$  in  $B_r(0)$ .  $\square$

#### COROLLARY 10.41: UNIQUE CONTINUATION PRINCIPLE

Assume that  $U \subset \mathbb{R}^n$  is a connected open set and  $f$  and  $g$  are two real analytic functions in  $U$ . If at some point  $x_0 \in U$  we have  $\partial^\alpha f(x_0) = \partial^\alpha g(x_0)$  for every multi-index  $\alpha$  then  $f = g$  in the whole  $U$ .

In particular, if  $f$  and  $g$  coincide in a nonempty open subset  $V \subset U$  then they must coincide in the whole  $U$ .

*Proof.* Let

$$U' := \{x \in U \mid \partial^\alpha f(x) = \partial^\alpha g(x) \text{ for all } \alpha \in \mathbb{N}^n\}$$

The complement of  $U'$  in  $U$  can be written as

$$U \setminus U' = \bigcup_{\alpha \in \mathbb{N}^n} (\partial^\alpha (f - g))^{-1}((-\infty, 0) \cup (0, +\infty)).$$

Since each  $\partial^\alpha (f - g)$  is a continuous function the previous set is a union of open sets, hence an open set.

Let us now show that  $U'$  is open. Indeed if  $x \in U$  then since both  $f$  and  $g$  are analytic there exist  $\rho > 0$  (the minimum of the two  $\rho$ 's for  $f$  and  $g$  at  $x$ ) and  $C > 1$  (the maximum of the two  $\rho$ 's for  $f$  and  $g$  at  $x$ ) such that  $f(x_0 + h) = f_x(h)$  for  $|h| < \varrho/2$  and  $g(x_0 + h) = g_x(h)$  for  $|h| < \varrho/2$ . But since by assumption  $x \in U'$  we have  $f_x(h) = g_x(h)$ . It follows that  $f - g = 0$  coincide in  $B_{\varrho/2}(x_0)$  and hence  $B_{\varrho/2}(x_0) \subset U'$ .

Since  $U$  is connected and  $U'$  nonempty (it contains  $x_0$ ) we obtain that  $U \setminus U'$  must be empty. Thus,  $f - g = 0$  in all of  $U$ .  $\square$



---

## Chapter 11

# Optimization and applications, Convexity

Differential calculus is crucial for solving optimization problems encountered in fields such as engineering, economics, and physics. These problems typically involve identifying the maximum or minimum values of a function subject to certain constraints. The Weierstrass theorem assures the existence of these optimal values for continuous functions defined over closed and bounded subsets of  $\mathbb{R}^n$ , but it does not provide a method for locating them. Thus, finding these optimal points necessitates additional techniques.

For a function like  $f(t) = t^3 - 2t + 1$  defined on the compact interval  $[0, 1]$ , the search for minima (or maxima) involves checking critical points where  $f'(t_0) = 0$ . For this function, solving  $f'(t) = 3t^2 - 2 = 0$  yields a single solution  $t_0 = \sqrt{2/3}$  within  $(0, 1)$ . At this point,  $f(\sqrt{2/3}) \approx -0.089$ . However, the extremum could also be at the boundary points  $t = 0$  and  $t = 1$ , with  $f(0) = 1$  and  $f(1) = 0$ . The minimum of these three values is approximately  $-0.089$ , indicating that the minimum occurs at the interior point  $\sqrt{2/3}$ .

12

What about multi-variable optimization? Consider, for example, minimizing  $f(x, y, z) = xy - y^2z + 6z^3$  within the closed unit ball  $\overline{B_1} \subset \mathbb{R}^3$ . To find potential interior points for minimal values, we extend the condition  $f'(t) = 0$  to higher dimensions. We must also consider potential minimum points on the domain's boundary, in this case, the sphere  $\partial B_1(0)$ , defined by the equation:

$$\partial B_1(0) = \{(x, y, z) \in \mathbb{R}^3 : g(x, y, z) = x_1^2 + x_2^2 + x_3^2 - 1 = 0\}.$$

This naturally leads to a 'constrained minimization problem': finding  $(x, y, z)$  that minimize  $f(x, y, z)$  among all points satisfying  $g(x, y, z) = 0$ .

This section will delve into techniques for tackling both unconstrained and constrained optimization problems in multiple variables.

## 11.1 First order optimality condition for interior points

### DEFINITION 11.1: GRADIENT

Let  $U \subset \mathbb{R}^n$  be an open set and  $f : U \rightarrow \mathbb{R}$  a  $C^1$  function. We define the **gradient** of  $f$  at  $x \in U$ , denoted  $\nabla f(x)$  as the (column) vector  $(\partial_1 f(x), \partial_2 f(x), \dots, \partial_n f(x))^T$ . In other words,  $\nabla f(x) = Jf(x)^T$ . Hence  $\nabla f : U \rightarrow \mathbb{R}^n$  is continuous map associating a vector to each point in  $U$  (i.e., a so-called ‘vector field’).

11.2. — For  $f \in C^1(U)$ ,  $x_0 \in U$  is called a **critical point** of  $f$  if the gradient of  $f$  vanishes at  $x_0$  (equivalently  $Df_{x_0}$  is the zero map).

Let  $U \subseteq \mathbb{R}^n$  be open and non-empty. We discuss the relationship between derivatives and extrema of real-valued functions  $f : U \rightarrow \mathbb{R}$ . As with functions in one variable, the vanishing of the derivative is a necessary, but not sufficient, condition for the presence of an extremum.

11.3. — Recall that an element  $x_0 \in U$  is called a **local maximum** of  $f$  if there exists  $r > 0$  such that  $f(x) \leq f(x_0)$  for all  $x \in B(x_0, r)$ . We say  $x_0$  is an **isolated local maximum** or **strict local maximum** if there exists  $r > 0$  such that  $f(x) < f(x_0)$  for all  $x \in B(x_0, r)$  with  $x \neq x_0$ . The definition of a **local minimum** is analogous, and collectively, we refer to them as **local extrema**.

### PROPOSITION 11.4: THE GRADIENT VANISHES AT A LOCAL EXTREMUM

Let  $U \subset \mathbb{R}^n$  be open,  $f : U \rightarrow \mathbb{R}$  be a  $C^1$  function, and let  $x_0 \in U$  be a point where  $f$  is differentiable and assumes a local extremum. Then  $\partial_i f(x_0) = 0$  for all  $i = 1, \dots, n$ . In other words,  $\nabla f(x_0) = 0$  and  $x_0$  is a critical point of  $f$ .

*Proof.* Without loss of generality, assume that  $f$  attains a local minimum at  $x_0$ . For all  $i \in \{1, \dots, n\}$  and sufficiently small  $h > 0$ , we have, by assumption,

$$f(x_0 + he_i) - f(x_0) \geq 0 \quad \text{and} \quad f(x_0 - he_i) - f(x_0) \geq 0$$

and thus  $\partial_i f(x_0) = 0$ , due to

$$\partial_i f(x_0) = \lim_{h \rightarrow 0^+} \frac{f(x_0 + he_i) - f(x_0)}{h} \geq 0, \quad \partial_i f(x_0) = \lim_{h \rightarrow 0^+} \frac{f(x_0 - he_i) - f(x_0)}{-h} \leq 0.$$

□

## 11.2 First order optimality condition with constraints: Lagrange multipliers

In this section we discuss the method of Lagrange multipliers, a useful method to tackle constrained minimization problems of the type

$$\min \{f(x) \mid g_1(x) = \dots = g_k(x) = 0\}. \quad (11.1)$$

### PROPOSITION 11.5: LAGRANGE MULTIPLIERS

Let  $U \subset \mathbb{R}^n$  be open,  $B_r(x_0) \subset U$ ,  $k \leq n$  and  $f, g_1, \dots, g_k$  functions in  $C^1(U, \mathbb{R})$ . Let

$$M := \{x \in U : g_1(x) = \dots = g_k(x) = 0\} \neq \emptyset,$$

and suppose that  $f|_M$  has a local minimum at  $x_0 \in M$ :

$$f(x) \geq f(x_0) \text{ for all } x \in M \cap B_r(x_0).$$

Then there are real numbers  $\lambda_0, \dots, \lambda_k$  such that

$$\lambda_0 \nabla f(x_0) + \lambda_1 \nabla g_1(x_0) + \dots + \lambda_k \nabla g_k(x_0) = 0, \quad \lambda_0^2 + \dots + \lambda_k^2 = 1. \quad (11.2)$$

12

*Proof “à la De Giorgi”.* Possibly replacing  $f(x)$  with  $\tilde{f}(x) := f(x) + |x - x_0|^2$  (notice that  $\nabla \tilde{f}(x_0) = \nabla f(x_0)$ ) and taking a smaller  $r$  (e.g.,  $\tilde{r} := r/2$ ), we may assume without loss of generality that  $x_0$  is the only local minimum in  $M \cap \overline{B_r(x_0)}$  (in other words, it is a strict local minimum).

**Step 1.** For given  $\varepsilon > 0$ , consider the penalized function

$$f_\varepsilon(x) := f(x) + \frac{1}{2\varepsilon}(g_1^2(x) + \dots + g_k(x)^2),$$

defined for  $x \in \overline{B_r(x_0)}$ . Take some sequence  $\varepsilon_\ell \rightarrow 0$  as  $\ell \rightarrow \infty$  and let  $x_\ell$  be a point of minimum of  $f_{\varepsilon_\ell}$  in the compact set  $\overline{B_r(x_0)}$ .

**Step 2.** We claim that  $f_{\varepsilon_\ell}(x_\ell) \rightarrow f(x_0)$  and  $x_\ell \rightarrow x_0$ , as  $\varepsilon \rightarrow 0$ . Indeed, by minimality

$$f_{\varepsilon_\ell}(x_\ell) \leq f_{\varepsilon_\ell}(x_0) = f(x_0). \quad (11.3)$$

Hence

$$g_1^2(x_\ell) + \dots + g_k(x_\ell)^2 \leq 2\varepsilon_\ell f_{\varepsilon_\ell}(x_\ell) \leq 2\varepsilon_\ell f(x_0) \rightarrow 0.$$

So, whenever a subsequence  $(x_{\ell_m})_{m \geq 1}$  converges to  $\bar{x} \in B_r(x_0)$  then  $\bar{x} \in M \cap \overline{B_r(x_0)}$ . Hence, we have

$$f(x_0) \leq f(\bar{x}) = \lim_m f(x_{\ell_m}) \leq \limsup_\ell f_{\varepsilon_\ell}(x_\ell) \leq f(x_0),$$

where we have used again (11.3). This implies  $\bar{x} = x_0$  (recall that  $x_0$  is a strict local minimum). This proves that  $f_{\varepsilon_\ell}(x_\ell) \rightarrow f(x_0)$  and that  $(x_k)_{k \geq 0}$  can only accumulate at  $x_0$ . On the other hand  $\{x_k\} \subset \bar{B}_r(x_0)$  is bounded so it must have accumulation points. Hence we must have  $x_k \rightarrow x_0$  as  $\varepsilon_\ell \downarrow 0$  (there is not even need to pass a sub-sequence).

**Step 3.** In particular  $x_\ell \in B_r(x_0)$  eventually, will be an interior critical point of  $f_{\varepsilon_\ell}$ , so by Proposition 11.4,

$$0 = \varepsilon_\ell \nabla f_{\varepsilon_\ell}(x_\ell) = \varepsilon_\ell \nabla f(x_\ell) + g_1(x_\ell) \nabla g_1(x_\ell) + \dots + g_k(x_\ell) \nabla g_k(x_\ell),$$

12 where we used that  $\partial_i(g_j)^2 = g_j \partial_i g_j$  for  $i = 1, \dots, n$  and so  $\nabla(g_j)^2 = g_j \nabla g_j$ .

This means that the  $k+1$  vectors  $\{\nabla f(x_\ell), \nabla g_1(x_\ell), \dots, \nabla g_k(x_\ell)\}$  are linearly dependent, hence there is a unit vector  $\lambda^\ell \in \mathbb{R}^{k+1}$  such that

$$0 = \lambda_0^\ell \nabla f(x_\ell) + \lambda_1^\ell \nabla g_1(x_\ell) + \dots + \lambda_k^\ell \nabla g_k(x_\ell). \quad (11.4)$$

**Step 4.** Since the unit  $k$ -sphere  $\{y \in \mathbb{R}^{k+1} \mid |y|^2 = 1\}$  is compact (it is a closed bounded subset of  $\mathbb{R}^{k+1}$ ) the sequence  $\lambda^\ell$  has a converging subsequence to a limit point  $\lambda$  in the unit  $k$ -sphere. But then passing (11.4) to the limit along this convergent subsequence we obtain (11.2).  $\square$

11.6. — Often Proposition 11.5 is used in practice in the following way, under the extra assumption that

For all  $x \in U$ , the vectors  $\nabla g_j(x)$ ,  $1 \leq j \leq k$ , are linearly independent.

In this situation we consider the so-called **Lagrangian function**

$$L : U \times \mathbb{R}^k \rightarrow \mathbb{R} \quad L(x, \lambda) = f(x) - \sum_{j=1}^k \lambda_j g_j(x).$$

13 The components of  $\lambda \in \mathbb{R}^k$  are called **Lagrange multipliers**, then Proposition 11.5 says that if  $x_0$  is a local minimum for the constrained problem there exists  $\lambda \in \mathbb{R}^k$  such that the equations

$$\partial_{x_i} L(x_0, \lambda) = 0 \quad \text{and} \quad \partial_{\lambda_j} L(x_0, \lambda) = 0$$

are satisfied for all  $i \in \{1, \dots, n\}$  and  $j \in \{1, \dots, k\}$ .

EXAMPLE 11.7. — Consider the function  $F : \mathbb{R}^2 \rightarrow \mathbb{R}$  given by  $F(x, y) = y^3 - x^2$ , and the compact set

$$K = \{(x, y) \in [-1, 1]^2 \mid F(x, y) = 0\}$$

and aim to find the minimum of the function  $f(x, y) = 4y - 3x$  on  $K$ .

Put  $M = K \cap (-1, 1)^2 = K \setminus \{(0, 0), (1, 1), (-1, 1)\}$ . On the one hand,

$$f(0, 0) = 0, \quad f(1, 1) = 1, \quad f(-1, 1) = 7.$$

On the other hand, we can use the method of Lagrange multipliers to seek the local minima of  $f|_M$ . The Lagrange function associated with  $f$  and  $F$  is given by

$$L(x, y, \lambda) = 4y - 3x - \lambda(y^3 - x^2).$$

The partial derivatives are calculated as

$$\begin{aligned} \partial_x L(x, y, \lambda) &= -3 + 2\lambda x \\ \partial_y L(x, y, \lambda) &= 4 - 3\lambda y^2 \\ \partial_\lambda L(x, y, \lambda) &= -(y^3 - x^2). \end{aligned}$$

From  $-3 + 2\lambda x = 0$ , we deduce  $\lambda \neq 0$  and  $x \neq 0$  with  $\lambda = \frac{3}{2x}$ . Similarly, from  $4 - 3\lambda y^2 = 0$ , we conclude that  $y \neq 0$  and  $\lambda = \frac{4}{3y^2}$ . Thus,  $\frac{3}{2x} = \frac{4}{3y^2}$  or equivalently  $x = \frac{9}{8}y^2$ . Furthermore,  $\partial_\lambda L(x, y, \lambda) = -(y^3 - x^2) = 0$ . Substituting  $x = \frac{9}{8}y^2$ , we obtain

$$0 = y^3 - \left(\frac{9}{8}y^2\right)^2 = y^3 - \frac{9^2}{8^2}y^4 = y^3 \left(1 - \frac{9^2}{8^2}y\right).$$

Since  $y \neq 0$ , this yields  $y = \frac{8^2}{9^2}$  and  $x = \frac{9}{8}y^2 = \frac{8^3}{9^3}$ . Therefore, using the Lagrange multipliers method, we find a single additional candidate for extremal values:

$$f\left(\frac{8^3}{9^3}, \frac{8^2}{9^2}\right) = 4\frac{8^2}{9^2} - 3\frac{8^3}{9^3} = 1.053\dots$$

The set of all points in  $K$  where  $f$  attains a local extremum on  $K$  is thus contained in

$$\{(0, 0), (1, 1), (1, -1), \left(\frac{8^3}{9^3}, \frac{8^2}{9^2}\right)\}.$$

The global maximum of  $f$  is at the point  $(1, 1)$  with a value of 7, and the global minimum is at the point  $(0, 0)$  with a value of 0.

**Applet 11.8 (Lagrange Multipliers and Normal Vectors).** *In this applet, we illustrate Proposition 11.5 with one constrain ( $k = 1$ ). We see that gradient vectors  $\nabla f$  and  $\nabla g$  must be parallel at minima and maxima of the constrained problem.*

### 11.3 Application: Spectral theorem for symmetric matrices

With the method of Lagrange multipliers, we can relatively easily prove the following important theorem from linear algebra.

**THEOREM 11.9: SPECTRAL THEOREM**

*Every symmetric matrix  $A \in \text{Mat}_{n,n}(\mathbb{R})$  is diagonalizable, and there exists an orthonormal basis of  $\mathbb{R}^n$  consisting of real eigenvectors of  $A$ .*

11.10. — Theorem 11.9 can be reformulated in the context of orthogonal matrices. To recall, an  $n \times n$  matrix  $O$  is termed orthogonal if it satisfies  $O^T O = \text{Id}$ , implying that  $O^{-1} = O^T$ . Orthogonal matrices are significant as they represent rotations in  $\mathbb{R}^n$  and preserve the length of vectors. This preservation can be demonstrated by the equation

$$|Ov|^2 = (Ov)^T Ov = v^T O^T Ov = v^T v = |v|^2,$$

valid for any vector  $v \in \mathbb{R}^n$ . It highlights that applying an orthogonal matrix to a vector does not change the vector's magnitude.

A set of vectors  $\{v_j\}$  forms an orthonormal basis if, and only if, the matrix composed of their components  $v_{j,i}$  is orthogonal. Consequently, Theorem 11.9 can be equivalently stated as: For any symmetric matrix  $A$ , there exists an orthogonal matrix  $O$  such that  $O^T A O$  yields a diagonal matrix.

Theorem 11.9 will be an immediate consequence of the following:

**LEMMA 11.11: FINDING EIGENVECTORS THROUGH MINIMIZATION**

*Let  $n \geq 1$  and  $A \in \text{Mat}_{n,n}(\mathbb{R})$  be a symmetric matrix. Let  $k \in \{0, 1, 2, \dots, n-1\}$  and assume that  $v_1, \dots, v_k$  are linearly independent real eigenvectors of  $A$ . Then,  $A$  has a real eigenvector  $v$  satisfying  $\langle v, v_j \rangle = 0$  for all  $j = 1, \dots, k$ .*

*Proof.* Define  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  as:

$$f(x) = \langle x, Ax \rangle = \sum_{k,\ell=1}^n a_{k\ell} x_k x_\ell.$$

Consider the linear subspace

$$H = \{x \in \mathbb{R}^n \mid \langle x, v_j \rangle = 0 \text{ for all } 0 \leq j \leq k\},$$

and the unit sphere

$$\mathbb{S}^{n-1} = \{x \in \mathbb{R}^n \mid |x| = 1\},$$

and define the compact set

$$K = \mathbb{S}^{n-1} \cap H.$$

Notice that

$$K = \{x \in \mathbb{R}^n \mid g_1(x) = g_2(x) = \dots = g_k(x) = g_*(x) = 0\}$$

where

$$g_j(x) = \langle x, v_j \rangle = \sum_{k=1}^n x_k v_{j,k}, \quad g_*(x) = |x|^2 - 1 = -1 + \sum_{k=1}^n (x_k)^2.$$

Since  $K$  is compact  $f|_K$  attains its extremal values. Let  $v \in K$  be an extremum point.

According Proposition 11.5, there exist  $\lambda_0, \lambda_1, \dots, \lambda_k, \lambda_* \in \mathbb{R}$  with  $\lambda_0^2 + \lambda_1^2 + \dots + \lambda_k^2 + \lambda_*^2 = 1$  such that

$$\lambda_0 \nabla f(v) + \sum_{j=1}^k \lambda_j \nabla g_j(v) + \lambda_* \nabla g_*(v) = 0. \quad (11.5)$$

Now we compute

$$\begin{aligned} \partial_i f(x) &= \partial_i \left( \sum_{k,\ell=1}^n a_{k\ell} x_k x_\ell \right) = \sum_{\ell=1}^n a_{i\ell} x_\ell + \sum_{k=1}^n a_{ki} x_k, \\ \partial_i g_j(x) &= \partial_i \left( \sum_{k=1}^n x_k v_{j,k} \right) = v_{j,i} \\ \partial_i g_*(x) &= \partial_i \left( \sum_{k=1}^n x_k^2 \right) = 2x_i \end{aligned}$$

where we used  $\partial_i x_k$  is zero unless  $k = i$ . Moreover, that  $A$  is symmetric, we obtain

13

$$\partial_i f(x) = 2 \sum_{\ell=1}^n a_{i\ell} x_\ell = 2(Ax)_i.$$

Hence we have shown

$$\nabla f(x) = 2Ax, \quad \nabla g_j(x) = v_j, \quad \nabla g_*(x) = 2x.$$

and thus (11.5) becomes

$$\lambda_0 Av + \sum_{j=1}^k \lambda_j v_j + \lambda_* v = 0.$$

Since  $v_j$  are linearly independent and  $v \in K$  is nonzero and perpendicular to all of them we obtain that  $\lambda_0 \neq 0$ . Also, using that  $v_j$  are eigenvectors of  $A$  we have

$$\langle Av, v_j \rangle = (Av)^T v_j = v^T A^T v_j = v^T A v_j = \langle v, A v_j \rangle = \mu_j \langle v, v_j \rangle = 0$$

for all  $j$ , where  $\mu_j$  denotes the eigenvalue associated to  $v_j$ . Hence,  $w_1 := \lambda_0 Av + \lambda_* v$  and  $w_2 := \sum_{j=1}^k \lambda_j v_j$  are perpendicular. Since their sum is zero it must be  $w_1 = w_2 = 0$  from which it follows that  $v$  is an eigenvector of  $A$  with real eigenvalue  $-\lambda_*/\lambda_0$ .  $\square$

*Proof of Theorem 11.9.* By applying Lemma 11.11 with  $k = 0$ , we obtain a first real eigenvector  $v_1$ , which can be assumed to be normalized, i.e.,  $|v_1| = 1$ . Subsequently, applying Lemma 11.11 again with  $k = 1$ , we identify a second real eigenvector  $v_2$  that is orthogonal to  $v_1$  and also normalized. By repeating this procedure for  $k$  times, starting from  $k = 0$  and progressing to  $k = n - 1$ , we ultimately find an orthonormal basis consisting of real eigenvectors.  $\square$

## 11.4 Second order optimality conditions

### DEFINITION 11.12: GRADIENT, HESSIAN, AND LAPLACIAN

Let  $U \subset \mathbb{R}^n$  be an open set and  $f : U \rightarrow \mathbb{R}$  a  $C^2$  function. We define the **Hessian matrix** of  $f \in C^2(U)$  at  $x \in U$  is the  $n \times n$  matrix

$$H_{ij}f(x) = \partial_i \partial_j f(x)$$

for  $i, j \in \{1, \dots, n\}$ . Schwarz's theorem 10.30 entails that  $Hf(x)$  is a symmetric matrix. An alternative standard notation for the Hessian matrix is  $D^2f(x)$ .

The **Laplacian** of  $f$  is the trace of the Hessian

$$\Delta f(x) := \operatorname{tr} Hf(x) = \sum_{i=1}^n \partial_{ii} f(x)$$

EXAMPLE 11.13. — Let  $u : \mathbb{R}^n \rightarrow \mathbb{R}$  of class  $C^2$ , check that

$$\begin{aligned} \partial_i(\arctan(u)) &= \frac{\partial_i u}{1+u^2}, & \partial_i(1/u) &= -\frac{\partial_i u}{u^2}, & \partial_i(|Du|^2) &= 2 \sum_j \partial_j u \partial_{ij} u, \\ \Delta(|Du|^2) &= 2 \sum_{i,j} (\partial_{ij} u)^2 + 2 \sum_i \partial_i u \partial_i(\Delta u). \end{aligned}$$

If  $O = (O_{ij}) \in O(n)$  and  $v(x) := u(Ax)$  then

$$\partial_i v(x) = \sum_j O_{ij} \partial_j u(Ox) \text{ and } \Delta v(x) = \sum_i \partial_{ii} v(x) = \sum_j (\partial_{jj} u)(Ox) = \Delta u(Ox).$$

EXERCISE 11.14 (Polarisation formula). — Let  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  be a smooth function. Prove that the map

$$\mathbb{R}^n \ni e \mapsto \left. \frac{d^2}{dt^2} \right|_{t=0} f(te) = e^T Hf(0)e,$$

determines all the second derivatives  $\partial_{ij} f$ .



## INTERLUDE: THE SIGN OF A SQUARE SYMMETRIC MATRIX

A symmetric matrix  $A \in \mathbb{R}^{n \times n}$  is called:

- (1) **Positive definite** if all its eigenvalues are positive.
- (2) **Negative definite** if all eigenvalues are negative.
- (3) **Indefinite** if at least one eigenvalue is positive and at least one is negative.
- (4) **Degenerate** if zero is an eigenvalue.

## PROPOSITION 11.15: HESSIAN TEST

Let  $U \subset \mathbb{R}^n$  be open and  $f \in C^3(U)$ , and let  $x_0 \in U$  with  $\nabla f(x_0) = 0$ . Let  $Hf(x_0)$  be the Hessian matrix of  $f$  at point  $x_0$ .

- (1) If  $Hf(x_0)$  is positive definite, then  $f$  has a strict local minimum at  $x_0$ .
- (2) If  $Hf(x_0)$  is negative definite, then  $f$  has a strict local maximum at  $x_0$ .
- (3) If  $Hf(x_0)$  is indefinite and non-degenerate, then  $f$  has no local extremum at  $x_0$ .  
In this case  $x_0$  is called a **saddle point**.

*Proof.* We will prove the first of the three statements. The proof of the other two are similar and are left as an exercise to the reader.

We notice that, since by assumption the first derivatives of  $f$  vanish at  $x_0$ , the quadratic Taylor expansion (provided by Theorem 10.33) can be written as

$$f(x_0 + h) - f(x_0) = \frac{1}{2} \sum_{i,j=1}^n (Hf)_{ij}(x_0) h_i h_j + O(|h|^3). \quad (11.6)$$

Now, if the Hessian of  $f$  at  $x_0$ ,  $H := Hf(x_0)$  is positive definite, and we put  $\lambda_1$  be its smallest (positive) eigenvalue we have

$$\sum_{i,j=1}^n H_{ij} v_i v_j \geq \lambda_1 |v|^2$$

Indeed, using that  $O^T H O$  is diagonal for some  $O$  orthogonal, we obtain, for any vector  $v$  (putting  $w = O^T v$ )

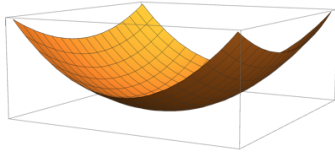
$$\sum_{i,j=1}^n (Hf)_{ij} v_i v_j = v^T H v = w^T (O^T H O) w = \sum_{\ell=1}^n \lambda_\ell |w_\ell|^2 \geq \lambda_1 |w|^2 = \lambda_1 |v|^2$$

Therefore (11.6) yields the existence of  $C$  large such that

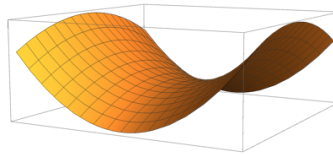
$$f(x_0 + h) - f(x_0) \geq \frac{\lambda_1}{2} |h|^2 - C|h|^3 = |h|^2(1 - C|h|)$$

which is strictly positive for all  $0 < |h| < \frac{1}{C}$ . This means that  $f$  has a strict local minimum at  $x_0$ .  $\square$

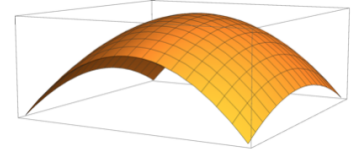
EXAMPLE 11.16. — The behavior of the following functions  $f : \mathbb{R}^2 \rightarrow \mathbb{R}$  at the point  $0 \in \mathbb{R}^2$  illustrates the three cases in the proposition.



$$f(x,y) = x^2 + y^2$$



$$f(x,y) = x^2 - y^2$$



$$f(x,y) = -x^2 - y^2$$

The corresponding Hessian matrices are  $\begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$ ,  $\begin{pmatrix} 2 & 0 \\ 0 & -2 \end{pmatrix}$ , and  $\begin{pmatrix} -2 & 0 \\ 0 & -2 \end{pmatrix}$ . When the Hessian matrix is degenerate, i.e., if 0 is an eigenvalue of  $H(x_0)$ , then the Hessian test is inconclusive and nothing can be said: for example the function  $f(x,y) = ax^4 + by^4$  has a local maximum, a local minimum, or neither at 0 depending on the choice of  $a$  and  $b$ . But the Hessian matrix at 0 is the zero matrix regardless of the choices of  $a$  and  $b$ .

13

EXAMPLE 11.17. — Let  $a, b \in \mathbb{R}$  be fixed parameters. We define  $f : \mathbb{R}^2 \rightarrow \mathbb{R}$  by  $f(x,y) = x \sin(y) + ax^2 + by^2$  for  $(x,y) \in \mathbb{R}^2$  and consider the point  $0 \in \mathbb{R}^2$ . We have  $Df(0) = 0$ , and the Hessian matrix of  $f$  at 0 is given by

$$H = \begin{pmatrix} 2a & 1 \\ 1 & 2b \end{pmatrix}$$

with  $\det H = 4ab - 1$ . We obtain the following cases.

- If  $a > 0$  and  $4ab - 1 > 0$ , then  $H$  is positive definite, and  $f$  has a local minimum at 0.
- If  $a < 0$  and  $4ab - 1 > 0$ , then  $H$  is negative definite, and  $f$  has a local maximum at 0.
- If  $\det H = 4ab - 1 = 0$ , then the Hessian matrix is degenerate.
- If  $4ab - 1 < 0$ , then  $H$  is indefinite, and 0 is a saddle point.

EXERCISE 11.18. — Let  $\alpha \in \mathbb{R}$ . Find all points  $(x,y) \in \mathbb{R}^2$  where the derivative of the function given by  $f(x,y) = x^3 - y^3 + 3\alpha xy$  vanishes. Determine whether each point is an extremum and, if so, whether it is a local minimum or maximum.

## 11.5 The Fundamental Theorem of Algebra via minimization

In this section, we present a straightforward proof demonstrating the existence of complex roots for polynomials through a minimization approach. It's noteworthy to emphasize that this minimization method requires examination beyond merely the first or second-order expansions at potential extremum points: it necessitates consideration of higher-order expansions.

### THEOREM 11.19: FUNDAMENTAL THEOREM OF ALGEBRA

*Every non-constant polynomial  $f \in \mathbb{C}[z]$  has a root in  $\mathbb{C}$ . Hence, applying the division algorithm,  $f$  has  $n$  complex roots (counted with multiplicity).*

*Proof.* Let  $f \in \mathbb{C}[z]$  be a polynomial of degree  $n > 0$ . Dividing  $f$  by  $a_n$  we may assume without loss of generality  $a_n = 1$ .

Consider the non-negative number

$$\mu := \inf\{|f(z)| : z \in \mathbb{C}\}.$$

Notice that, by the triangle inequality, in the region  $|z| \geq 1$  we have

$$|f(z)| = \left| \sum_{k=0}^n a_k z^k \right| \geq |z^n| - \sum_{k=0}^{n-1} |a_k| |z|^k \geq |z|^{n-1} (|z| - C)$$

for  $C := \sum_{k=0}^n |a_k|$ . So, if we fix any  $R \geq C + 1 + 10\mu$  we have

$$|f(z)| \geq 1 + 10\mu \text{ as soon as } |z| \geq R.$$

This shows that the ‘‘battle for the infimum’’ is fought inside the compact ball  $K := \{z \in \mathbb{C} : |z| \leq R\}$ , that is to say

$$\min\{|f(z)| : z \in K\} = \inf\{|f(z)| : z \in K\} = \mu,$$

since in  $K^c$  we have  $|f(z)| \geq 1 + 10\mu$ . So let  $z_0 \in K$  with  $|f(z_0)| = \mu$ . Since  $|f(z)| \geq 1 + 10\mu > \mu$  on  $\partial K$ , we obtain that  $z_0$  is an interior minimum point in  $K$ .

We claim that  $z_0$  is a root of  $f$ , we set for brevity

$$g(z) := f(z_0 + z) = \sum_{k=0}^n a_k (z_0 + z)^k = \sum_{k=0}^n b_k z^k,$$

for some coefficients  $b_k \in \mathbb{C}$ . Assume by contradiction that  $f(z_0) \neq 0$ , that is to say  $b_0 = g(0) \neq 0$ . Also, let  $\ell \geq 1$  be the smallest index  $\geq 1$  with  $b_\ell \neq 0$ . Now writing  $z$  in exponential

form we have

$$g(re^{i\varphi}) = b_0 + b_\ell r^\ell e^{i\ell\varphi} + O(r^{\ell+1}) = b_0 \left(1 + \frac{b_\ell}{b_0} r^\ell e^{i\ell\varphi}\right) + O(r^{\ell+1}) \quad \text{as } r \downarrow 0,$$

Write  $\frac{b_\ell}{b_0} = se^{i\psi}$  for some  $s > 0$  and choose  $\varphi = \frac{-\psi + \pi}{\ell}$ , so that  $e^{i(\ell\varphi + \psi)} = -1$ , and

$$|b_0| = |g(0)| \leq |g(re^{i\varphi})| = |b_0(1 - sr^\ell) + O(r^{\ell+1})| \leq |b_0|(1 - sr^\ell) + Mr^{\ell+1}$$

for all sufficiently small  $r > 0$  and some fixed constant  $M > 0$ . Reshuffling this expression we find

$$0 < s|b_0| \leq Mr$$

which is impossible if we let  $r \downarrow 0$ . Thus, it was a contradiction to assume  $b_0 \neq 0$ , and this means that  $f(z_0) = 0$ .  $\square$

EXERCISE 11.20. — Let  $U \subset \mathbb{C}$  be open, and  $f : U \rightarrow \mathbb{C}$  be a complex-valued function that can be locally represented by power series (an analytic function). More precisely, for every  $x_0 \in U$ , there exists an  $r > 0$  such that  $B(x_0, r) \subset U$ , and  $f$  on  $B(x_0, r)$  is equal to a power series around  $x_0$  with a convergence radius greater than or equal to  $r$ . Mimicking the proof of Theorem 11.19, show that the function  $z \mapsto |f(z)|$  does not assume a minimum value on  $U$ .

14

EXERCISE 11.21. — For the sake of completeness, we present an elementary argument for the proof of Lemma 11.11 using the Fundamental Theorem of Algebra. Let  $n \geq 1$  and  $A \in \text{Mat}_{n,n}(\mathbb{R})$  be a symmetric matrix.

- (i) Show that all complex eigenvalues of  $A$  are real.
- (ii) Prove Lemma 11.11 by showing that  $A$  has a complex eigenvector if and only if  $A$  has a real eigenvector.

The geometric understanding of the eigenvalues of  $A$  gained in our proof can also be utilized differently. As an example, one can prove a special case of the Courant-Fischer theorem.

- (iii) Show that the values

$$\min_{x \in \mathbb{S}^{n-1}} x^t A x, \quad \max_{x \in \mathbb{S}^{n-1}} x^t A x$$

represent the smallest and largest eigenvalue of  $A$ , respectively.

EXERCISE 11.22. — For  $n \geq 2$ , prove that two points  $x, y \in \mathbb{S}^{n-1}$  have maximum distance if and only if  $x = -y$ . Consider the function  $(x, y) \mapsto \|x - y\|^2$  on  $\mathbb{S}^{n-1} \times \mathbb{S}^{n-1} \subset \mathbb{R}^{2n}$ .

## 11.6 Convexity

### DEFINITION 11.23: CONVEX SETS AND FUNCTIONS

A nonempty subset  $A \subset \mathbb{R}^n$  is called **convex** if for any two points  $x, y \in A$  and any  $t \in [0, 1]$ , the point  $(1 - t)x + ty$  also lies in  $A$ . In other words, the line segment connecting any two points in  $A$  lies entirely within  $A$ .

A function  $f : A \rightarrow \mathbb{R}$  is called **convex** if for every  $x, y \in A$  and any  $t \in [0, 1]$ , the following inequality holds:

$$f((1 - t)x + ty) \leq (1 - t)f(x) + tf(y).$$

This means that the line segment connecting any two points on the graph of  $f$  lies above or on the graph. In other words, the graph ‘dips in the middle’ (i.e, it is ‘U-shaped’).

### PROPOSITION 11.24: NONNEGATIVE HESSIAN AND CONVEXITY

Suppose that  $U \subset \mathbb{R}^n$  is open and convex and let  $f : U \rightarrow \mathbb{R}$ . Then  $f$  is convex if and only if one of the following conditions holds

(a) For all  $x \in U$  the Hessian  $Hf(x)$  is nonnegative definite (i.e., all of its eigenvalues are nonnegative)

(b) For all  $x$  and  $y$  in  $U$  we have

$$f(y) - f(x) \geq Df_x(y - x).$$

*Proof.* Assume  $f$  is convex. For any given  $x \in U$  and  $v \in \mathbb{R}^n$  with  $|v|=1$ , let  $g(s) := f(x + sv)$ ,  $s \in (-\varepsilon, \varepsilon)$ . Using the convexity of  $g$  we obtain  $g(a(1 - t) + bt) \leq (1 - t)g(a) + tg(b)$ , for all  $a, b \in (-\varepsilon, \varepsilon)$  and  $t \in [0, 1]$ . But then by Corollary ?? we have  $g''(s) \geq 0$  for all  $s$  close to 0.

Now the chain rule gives

$$g'(s) = \sum_i \partial_i f(x + se)v_i \quad g''(s) = \sum_{i,j} \partial_{ij} f(x + se)v_i v_j$$

So we have shown

$$g''(0) = H_{ij} f(x)v_i v_j \geq 0.$$

Since  $v \in \mathbb{R}^n$  with  $|v| = 1$  this implies that all the eigenvalues of  $Hf(x)$  have are nonnegative, proving (a).

Now assume (a) and let us prove (b). Reciprocally assume and fix  $x, y$  in  $U$  with  $x \neq y$ . Put  $v := y - x$  and define, as above,  $g(s) := f(x + tv)$ . Now  $g$  is defined for  $s \in (-\varepsilon, 1 + \varepsilon)$ . Now, by the same computation as above we obtain  $g''(s) \geq 0$  and hence, by Taylor’s theorem

$$f(y) - f(x) = g(1) - g(0) = g'(0) + \int_0^1 (1 - t)g''(t)dt \geq g'(0) = Df_x(y - x).$$

Finally if (b) holds then for any  $x, y \in U$  and  $t \in [0, 1]$ , putting  $z = (1 - t)x + ty$  we have

$$f(x) - f(z) \geq Df_z(x - z) \quad \text{and} \quad f(y) - f(z) \geq Df_z(y - z)$$

Multiplying the first inequality by  $(1 - t)$  and the second by  $t$  and summing we obtain

$$(1 - t)f(x) + tf(y) - f(z) \geq Df_z((1 - t)x + ty - z) = Df_z(0) = 0.$$

Hence, the convexity inequality follows.  $\square$

EXERCISE 11.25. — Show that condition (b) in Proposition 11.24 is equivalent to the convexity of when  $f$  is  $C^1$ . Deduce that for a  $C^1$  function on an open convex set, every critical point is a point of minimum.

14

**PROPOSITION 11.26: JENSEN INEQUALITY**

Let  $U \subset \mathbb{R}^n$  be open and convex and  $f : U \rightarrow \mathbb{R}$  convex. Then for every collection of points  $x_i \in U$  and positive ‘weights’  $w_i \in (0, 1)$ ,  $i = 1 \dots N$ , with  $\sum_i w_i = 1$  we have

$$f\left(\sum_i w_i x_i\right) \leq \sum_i w_i f(x_i).$$

*Proof.* We will assume for simplicity that  $f$  is  $C^2$  we will give a proof using Proposition 11.24 (b). It is left as an exercise to give proof for any convex function by induction over  $N$ .

Let  $z = \sum_i w_i x_i$ . Then

$$f(x_i) - f(z) \geq Df_z(x_i - z).$$

Multiplying by  $w_i$  and summing over  $i$  we obtain the desired conclusion.  $\square$

---

## Chapter 12

# Inverse and implicit function theorems and submanifolds

### 12.1 The inverse function Theorem

#### 12.1.1 Small Lipschitz perturbations of the identity

The following lemma is the essential preliminary step towards the inverse function theorem.

**LEMMA 12.1: SMALL LIPSCHITZ PERTURBATIONS OF THE IDENTITY**

Let  $U \subset \mathbb{R}^n$  be an open set and assume that  $F : U \rightarrow \mathbb{R}^n$  is a function of the form  $F(x) = x + \phi(x)$ , where  $\phi$  is Lipschitz with constant  $\lambda < 1$ . Then

(1) Whenever  $B_r(x) \subset U$  we have

$$B_{(1-\lambda)r}(F(x)) \subset F(B_r(x)) \subset B_{(1+\lambda)r}(F(x)).$$

In particular,  $F(U)$  is open.

(2)  $F$  is injective and  $F^{-1} : F(U) \rightarrow U$  is Lipschitz with constant  $\frac{1}{1-\lambda}$ .

15

*Proof.* We start with the first inclusion in (1), which is the core of this proof. Given a point  $y \in B_{(1-\lambda)r}(F(p_0))$ , for some  $p_0 \in U$  such that  $B_r(x_0) \subset U$  we want to find  $x \in B_r(x_0)$  that such that

$$F(x) = y.$$

It is convenient to look at this equation as

$$x = y - \phi(x)$$

because then we want to find a fixed point of the map  $x \mapsto y - \phi(x)$ , and we can use an argument almost identical the one in the Banach fixed point theorem.

Indeed, consider the sequence of points defined by recurrence, starting at  $x_0$ ,

$$x_{k+1} = y - \phi(x_k), \quad k \geq 0$$

Our goal is to show that  $x_k$  always belongs to  $U$  so that  $x_{k+1}$  is always well-defined and that the sequence  $(x_k)_{k \geq 0}$  converges to a limit  $x \in B_r(x_0)$ , then  $x = y - \phi(x)$  and we have solved  $y = F(x)$ .

Now, using the triangular inequality and the contraction property of  $\phi$  we find

$$|x_{k+1} - x_k| \leq |\phi(x_k) - \phi(x_{k-1})| \leq \lambda |x_{k-1} - x_k| \leq \dots \leq \lambda^k |x_1 - x_0| = \lambda^k |y - F(x_0)|,$$

which proves that  $(x_k)_{k \geq 0}$  is Cauchy (if it is well defined).

But this also shows in turn that  $(x_k)_{k \geq 0}$  is well defined because the sequence never escapes from  $B_r(x_0) \subset U$ . Indeed:

$$|x_{k+1} - x_0| \leq \sum_{i=0}^k |x_{i+1} - x_i| \leq |y - F(x_0)| \sum_{i=0}^k \lambda^i < \frac{|y - F(x_0)|}{1 - \lambda} < r.$$

Hence  $x_k$  converges to  $x$  satisfying  $|x - x_0| \leq |y - F(x_0)|/(1 - \lambda) < r$ , which means that  $x \in B_r(x_0)$ , as we wanted to show.

In particular, since the point  $x_0$  is arbitrary we have shown that  $F(U)$  is open.

Finally, the second inclusion in (1) is readily checked

$$|F(y) - F(x_0)| = |y + \phi(y) - x_0 - \phi(x_0)| \leq |y - x| + \lambda |y - x| < r + \lambda r.$$

We turn to (2). First of all, if  $F(x) = F(x')$ , then  $x - x' = \phi(x) - \phi(x')$ , which is in contradiction with  $\lambda < 1$ . So  $F$  is injective and  $F^{-1}$  is a well-defined function.

Finally, for  $x, x' \in U$  putting  $y = F(x)$  and  $y' = F(x')$  we have

$$y - y' = x - x' + \phi(x) - \phi(x')$$

and by the triangle inequality

$$|y - y'| \geq |x - x'| - |\phi(x) - \phi(x')| \geq (1 - \lambda)|x - x'| = (1 - \lambda)|F^{-1}(y) - F^{-1}(y')|,$$

which proves that  $F^{-1}$  is Lipschitz with constant  $\frac{1}{1-\lambda}$ . □

**EXERCISE 12.2.** — Proof that the map  $y \mapsto F^{-1}(y) - y$  has Lipschitz constant  $\frac{\lambda}{1-\lambda}$ , so it is also a small Lipschitz perturbation of the identity when  $\lambda < 1/2$ .

**EXERCISE 12.3.** — In the proof of Lemma 12.1 we did not use the Euclidean structure of  $\mathbb{R}^n$ . In fact, this Lemma is true in any complete, normed vector space  $(V, \|\cdot\|)$ , with the same statement and the same proof. Check this claim.



**LEMMA 12.4: AUTOMATIC DIFFERENTIABILITY OF THE INVERSE**

Let  $U, V \subset \mathbb{R}^n$  be open sets and let  $f: U \rightarrow V$  and  $g: V \rightarrow U$  be bijective functions. Assume that  $f$  is differentiable at  $x_0 \in U$  and  $Df_{x_0}$  is invertible, and that  $g$  is Lipschitz in  $V$ , and that

$$f(g(y)) = y, \quad \text{for all } y \in V.$$

Then  $g$  must be differentiable at  $y_0 = f(x_0)$  and  $Dg_{y_0} = (Df_{x_0})^{-1}$ .

*Proof.* Let us set  $L := Df_{x_0}$ . By the differentiability assumption, as  $y \rightarrow y_0$ ,

$$y - y_0 = f(g(y)) - y_0 = f(x_0 + (g(y) - g(y_0))) - f(x_0) = L(g(y) - g(y_0)) + o(|g(y) - g(y_0)|).$$

Using that  $L$  is invertible, we re-write this as

$$g(y) - g(y_0) = L^{-1}(y - y_0) + o(\|L^{-1}\|_2 |g(y) - g(y_0)|). \quad (12.1)$$

We conclude noticing that, since  $g$  is Lipschitz, say with constant  $\Lambda$ ,

$$o(\|L^{-1}\|_2 |g(y) - g(\bar{y})|) = o\left(\frac{|g(y) - g(\bar{y})|}{|y - \bar{y}|} |y - \bar{y}|\right) = o(\Lambda |y - \bar{y}|) = o(|y - \bar{y}|).$$

Thus (12.1) is saying that  $g$  is differentiable at  $\bar{y}$ , with differential  $L^{-1}$ .  $\square$

15

**12.1.2 Inverse of differentiable maps****LEMMA 12.5: SMOOTHNESS OF THE INVERSE**

Let  $U \subset \mathbb{R}^{n \times n}$  be the set of invertible matrices, and let  $\theta: U \rightarrow U$  be defined as

$$\theta: X \mapsto X^{-1}.$$

Then  $\theta \in C^\infty$ .

*Proof.* Recall that the formula for the inverse matrix expresses the  $(p, q)$  entry of  $X^{-1}$  as a polynomial of  $\{X_{i,j}\}$  divided by the polynomial  $\det X$  (which is nonzero in  $U$ ).

Indeed, each entry  $(p, q)$  of the inverse matrix  $X^{-1}$  can be computed as:

$$(X^{-1})_{p,q} = \frac{C_{q,p}}{\det(X)},$$

where  $C_{q,p}$  is the cofactor of the element at position  $(q, p)$  in the matrix  $X$ . This cofactor is calculated as  $(-1)^{q+p} \cdot \det(X_{q,p})$ , where  $\det(X_{q,p})$  is the determinant of the matrix obtained by removing the  $q$ -th row and  $p$ -th column from  $X$ .

Thus, since the map  $X \mapsto X^{-1}$  is expressed as a quotient of polynomials in terms of the entries of  $X$ , it is infinitely differentiable on the domain excluding the points where the determinant, serving as the denominator, vanishes.

### 12.1.3 Inverse and Implicit function theorems

#### THEOREM 12.6: INVERSE FUNCTION

Let  $U \subset \mathbb{R}^n$  be an open set and  $f \in C^1(U, \mathbb{R}^n)$ . Let  $x_0 \in U$  be such that  $Df_{x_0} : \mathbb{R}^n \rightarrow \mathbb{R}^n$  is invertible. Then, there is  $U_0$  open containing  $x_0$  such that

(1)  $f$  is injective in  $U_0$ .

(2) The set  $V := f(U_0)$  is open.

(3) The inverse function  $g := (f|_{U_0})^{-1}$ , that is  $f: V \rightarrow U_0$  satisfying

$$g(f(x)) = x, \quad \forall x \in U_0 \quad (\text{or equivalently } f(g(y)) = y, \quad \forall y \in V)$$

is of class  $C^1$  and

$$Dg_{f(x)} = (Df_x)^{-1}, \quad \forall x \in U_0; \quad Dg_y = (Df_{g(y)})^{-1}, \quad \forall y \in V.$$

Furthermore, if  $f \in C^k(U, \mathbb{R}^n)$  for some  $k \geq 1$ , then also  $f^{-1} \in C^k(f(B_r(x_0)), U)$ .

*Proof.* It is enough to prove the result in the case  $x_0 = 0$  and  $f(x_0) = 0$ . Indeed the general case follows immediately from this “special” case applied to  $\tilde{f}(x) := f(x - x_0) - f(x_0)$ .

Let  $L := D\phi_0$ , we claim that  $F := L^{-1} \circ f|_{B_r(0)}$  is a small Lipschitz perturbation of the identity, provided  $r > 0$  is chosen small enough.

Notice that the function  $F$  is of class  $C^1$  (being the composition of a  $C^1$  function and a linear map). Write  $\text{Id} : \mathbb{R}^n \rightarrow \mathbb{R}^n$  the identity function (i.e.,  $\text{Id}(x) = x$  for all  $x \in \mathbb{R}^n$ ), and put  $\phi := F - \text{Id}$ , i.e.,  $\phi(x) := F(x) - x$ . We have

$$D\phi_0 = D(F - \text{Id})_0 = DF_0 - D\text{Id}_0 = L \circ D\phi_0 - \text{Id} = L^{-1} \circ L - \text{Id} = 0$$

so choosing  $r > 0$  small enough we will have, by continuity of the derivatives the Jacobi matrix  $J\phi$  satisfies

$$\|J\phi(x)\|_2 \leq \frac{1}{2} \quad \text{for all } x \in B_r(0).$$

Then the Mean Value Theorem (see the proof Corollary 10.26) implies that  $\phi$  is Lipschitz with constant  $1/2$  in  $B_r := B_r(0)$

By Lemma 12.1, we find that  $F(B_r(0))$  is open, that  $F|_{B_r(0)}$  is injective and that the inverse function  $F^{-1} : F(B_r) \rightarrow B_r$  is Lipschitz, with constant 2.

Now (1) holds because  $f = L \circ F$ , which is a composition of injective functions.

Also (2) holds because  $f(B_r) = L(F(B_r))$  is open, where  $F(B_r)$  is open and  $L$  linear invertible (an invertible linear map sends open sets to open sets).

For (3) we notice that  $g = F^{-1} \circ L^{-1}$  will be Lipschitz in  $f(B_r) = L(F(B_r))$ , since it is the composition of Lipschitz maps. Indeed, since  $F^{-1}$  is 2-Lipschitz:

$$|g(y) - g(y')| = |F^{-1}(L^{-1}(y)) - F^{-1}(L^{-1}(y'))| \leq 2|L^{-1}(y) - L^{-1}(y')| \leq 2\|L^{-1}\|_2|y - y'|.$$

Since  $f$  is differentiable at every point of  $B_r$ , and  $g$  is Lipschitz, then Lemma 12.4 shows that  $g$  is differentiable for all  $y \in f(B_r)$ , and

$$Dg_y = (Df_{g(y)})^{-1} \quad \Leftrightarrow \quad Jg(y) = (Jf(g(y)))^{-1},$$

for all  $y \in f(B_r)$ .

Finally, assume that  $f \in C^k$  and that we know  $f^{-1} \in C^\ell$  for some  $\ell \in \{0, 1, 2, \dots, k-1\}$ . The formula for the Jacobi matrix of the inverse  $Jg$  give it as the composition of the following functions:

$$g: y \mapsto g(y) \quad \text{which is of class } C^\ell,$$

$$Jf: x \mapsto Jf(x) \quad \text{which is of class } C^{k-1},$$

$$\theta: X \mapsto X^{-1} \quad \text{which is of class } C^\infty \text{ on the set of invertible matrices by Lemma 12.5.}$$

Thus  $Jg \in C^\ell$ , which means that  $g \in C^{\ell+1}$ . This proves the last part of the statement by induction from the  $\ell = 0$  case.  $\square$

#### DEFINITION 12.7: DIFFEOMORPHISM

Let  $U, V \subset \mathbb{R}^n$  be open. A bijective,  $C^1$  function  $f : U \rightarrow V$  with a  $C^1$  inverse  $f^{-1} : V \rightarrow U$  is called a **diffeomorphism**. If  $f$  and  $f^{-1}$  are both  $k$ -times continuously differentiable for  $k \geq 1$ , we call  $f$  a  **$C^k$ -diffeomorphism**.

12.8. — Notice that if  $f : U \rightarrow V$  is a diffeomorphism then it follows from  $f^{-1} \circ f = \text{Id}$  and the chain rule that  $D(f^{-1})_{f(x)} Df_x = \text{Id}$  for all  $x \in U$ . In particular  $DF_x$  is always invertible for all  $x \in U$ .

An important consequence of the Inverse Function Theorem is the following

## COROLLARY 12.9: IMPLICIT FUNCTION THEOREM

Let  $0 < d < n$ ,  $k \geq 1$  be integers,  $U \subset \mathbb{R}^n$  be open, let  $f \in C^1(U, \mathbb{R}^{n-d})$ . We write a point in  $\mathbb{R}^d \times \mathbb{R}^{n-d}$  as  $(x, y)$  with  $x \in \mathbb{R}^d$  and  $y \in \mathbb{R}^{n-d}$ .

Assume that we have  $(x_0, y_0) \in U$  with  $f(x_0, y_0) = 0$  such that the  $(n-d) \times (n-d)$  matrix

$$J_y f(x_0, y_0) := (\partial_{y_i} f_j(x_0, y_0))_{1 \leq j, i \leq n-d}$$

is invertible. Then, for sufficiently small  $r, s > 0$ , there is a function  $g$  from  $B_r(x_0) \subset \mathbb{R}^d$ , to  $B_s(y_0) \subset \mathbb{R}^{n-d}$ , such that, for all  $(x, y)$  in the cylinder  $U_0 := B_r(x_0) \times B_s(y_0) \subset U$ , it holds

$$f(x, y) = 0 \iff y = g(x).$$

Moreover, for all  $x \in B_r(x_0)$ ,

$$Jg(x) = -((J_y f)(x, g(x)))^{-1} (J_x f)(x, g(x)),$$

where

$$J_x f(x_0, y_0) := (\partial_{x_i} f_j(x_0, y_0))_{1 \leq j \leq n-d, 1 \leq i \leq d}$$

Furthermore, if  $f \in C^k(U, \mathbb{R}^{n-d})$  for some  $k \geq 1$ , then also  $g \in C^k(B_r(x_0), \mathbb{R}^d)$ .

*Proof.* Consider the function  $\Phi \in C^1(U, \mathbb{R}^n)$  given by

$$\Phi(x, y) := (x, f(x, y)), \quad \text{for all } (x, y) \in U.$$

By assumption the matrix  $D\Phi(x_0, y_0)$  — which has size  $n \times n$  — has a block decomposition

$$D\Phi(x_0, y_0) = \begin{pmatrix} \mathbf{1}_d & 0 \\ J_x f(x_0, y_0) & J_y f(x_0, y_0) \end{pmatrix},$$

so in particular it is invertible and we are under the assumptions of the Inverse function Theorem. Hence  $\Phi$  has a  $C^1$  inverse when restricted to a small cylinder ‘centered at  $(x_0, y_0)$ ’:

$$U_0 := B_r(x_0) \times B_s(y_0) \subset U,$$

$r, s > 0$ , which is mapped to the open set  $V := \Phi(U_0) \subset \mathbb{R}^n$ .

Let  $\Psi : V \rightarrow U_0$  denote the inverse of the restriction of  $\Phi$  to  $U_0$ . For given points  $(x, y)$  in  $U_0$  put

$$(\xi, \eta) := \Phi(x, y) = (x, f(x, y)), \quad \iff \quad (x, y) = \Psi(\xi, \eta)$$

Then  $\xi = x$ , so  $\Psi(\xi, \eta)$  is of the form

$$\Psi(\xi, \eta) = (\xi, G(\xi, \eta)), \quad \text{for all } (\xi, \eta) \in V,$$

for some  $G : V \rightarrow B_s(y_0)$  of class  $C^1$  (or of class  $C^k$  if  $f$  is of class  $C^k$ ).

Thus, since

$$(x, y) = \Psi(x, \eta) = (x, G(x, \eta))$$

we obtain

$$f(x, y) = 0 \iff \eta = 0 \iff y = G(x, 0),$$

where the last implication  $\Leftarrow$  follows using that  $\Psi$  is bijective.

In other words, defining  $g(x) := G(x, 0)$  we obtain what we need.

Finally, the formula for  $Jf(x_0)$  follows from differentiating the identity

$$f(x, g(x)) = 0$$

using chain rule. Indeed, we obtain, for all  $1 \leq i \leq d$

$$0 = \partial_i(f(x, g(x))) = \partial_{x_i}f(x, g(x)) + \sum_{\ell=1}^{n-d} \partial_{y_\ell}f(x, g(x))\partial_\ell g(x),$$

or in matricial form:

$$0 = J_x f(x, g(x)) + J_y f(x, g(x))J_x g(x).$$

□

16

## 12.2 Submanifolds of $\mathbb{R}^n$

### 12.2.1 Definition of submanifold and different representations

#### DEFINITION 12.10: SUBMANIFOLD OF $\mathbb{R}^n$

Given  $0 < d < n$  and  $k \geq 1$  integers, We say that  $M \subset \mathbb{R}^n$  nonempty is a  **$d$ -dimensional submanifold of  $\mathbb{R}^n$**  of class  $C^k$  if, for every point  $p_0 \in M$  there exists a  $U \subset \mathbb{R}^n$  open containing  $p_0$ ,  $V \subset \mathbb{R}^n$  open containing 0, and a  $C^k$ -diffeomorphism  $\Psi : U \rightarrow V$ , such that:

$$\Psi(M \cap U) = \{y \in V \mid y_{d+1} = y_{d+2} = \dots = y_n = 0\}.$$

The map  $\Psi$  is called a **submanifold chart**.

For the next result, as well as for the submanifolds section it is useful to introduce the following:

#### DEFINITION 12.11: PERMUTATION OF COORDINATES

We call a map  $P : \mathbb{R}^n \rightarrow \mathbb{R}^n$  a **permutation of the coordinates** if for some permutation  $\sigma : \{0, 1, \dots, n\} \rightarrow \{1, 2, \dots, n\}$  we have

$$P(x_1, x_2, \dots, x_n) = P(x_{\sigma(1)}, x_{\sigma(2)}, \dots, x_{\sigma(n)}), \quad \forall x \in \mathbb{R}^n$$

17

## PROPOSITION 12.12: DIFFERENT LOCAL REPRESENTATIONS OF SUBMANIFOLD

Let  $0 < d < n$  and  $k \geq 1$  be integers and  $M \subset \mathbb{R}^n$  a nonempty subset. The following four are equivalent:

(1)  $M \subset \mathbb{R}^n$  is  $d$ -dimensional submanifold.

(2) (**Implicit representation**). For all  $p_o \in M$  there are  $U \subset \mathbb{R}^n$  open containing  $p_o$  and  $f \in C^k(U, \mathbb{R}^{n-d})$  such that  $Df_{p_o}$  has maximal rank (i.e., has rank  $n - d$ )

$$M \cap U = \{x \in U \mid f(x) = 0\}.$$

(3) (**Parametric representation**). For all  $p_o \in M$  there are  $U \subset \mathbb{R}^n$  containing  $p_o$ ,  $V \subset \mathbb{R}^d$  open,  $y_o \in V$ , and  $G \in C^k(V, \mathbb{R}^n)$  such that  $G(y_o) = p_o$ , such that  $DG_{y_o}$  has maximal rank (i.e., has rank  $d$ )

$$M \cap U = G(V).$$

(4) (**Graphical representation**). There are  $U \subset \mathbb{R}^n$  open containing  $p_o$ ,  $V \subset \mathbb{R}^d$  open, and  $g \in C^k(V, \mathbb{R}^{n-d})$  such that

$$M \cap U = P(\text{graph}(g)),$$

where  $P: \mathbb{R}^n \rightarrow \mathbb{R}^n$  is some permutation of the coordinates and

$$\text{graph}(g) := \left\{ x \in U \times \mathbb{R}^{n-d} \mid (x_{d+1}, \dots, x_n) = g(x_1, \dots, x_d) \right\} \subset \mathbb{R}^n$$

*Proof.* We will show (1)  $\implies$  (2)  $\implies$  (4)  $\implies$  (3)  $\implies$  (1).

(1)  $\implies$  (2): Since  $M$  is a submanifold, there is  $U \subset \mathbb{R}^n$  open containing  $p_o$  and a diffeomorphism  $\Psi: U \rightarrow V \subset \mathbb{R}^n$  such that

$$x \in M \cap U \iff \Psi_{d+1}(x) = \dots = \Psi_n(x) = 0.$$

Hence, (2) follows defining  $f: U \rightarrow \mathbb{R}^{n-d}$  as  $f(x) := (\Psi_{d+1}(x), \dots, \Psi_n(x))^T$ . Notice that  $Df_{p_o}$  has maximal rank since it is made of columns of the invertible matrix  $D\Psi_{p_o}$ .

(2)  $\implies$  (4): If  $Df_{p_o}$  has maximal rank  $n - d$ , then for a suitable permutation of the variables  $P$ , when we consider the map  $\tilde{f} = f \circ P^{-1}$ , the  $(n - d) \times (n - d)$  matrix

$$(\partial_i \tilde{f}_j(p_o))_{1 \leq j \leq n-d, d+1 \leq i \leq n}$$

is invertible.

Hence  $\tilde{f}$  satisfies the assumptions of Implicit Function Theorem and hence there exist an open cylinder  $U_0 \subset U$  centered at  $p_o$ ,  $V \subset \mathbb{R}^d$  open,  $g: V \rightarrow \mathbb{R}^{n-d}$  of class  $C^k$ , such that for

$x \in U_0$

$$P(x) \in M \iff \tilde{f}(x) = 0 \iff (x_{d+1}, \dots, x_n) = g(x_1, \dots, x_d).$$

(4)  $\implies$  (3): By (3) we have  $M \cap U = P(\text{graph}(g))$ , for some  $g : V \rightarrow \mathbb{R}^{n-d}$ , where  $V \subset \mathbb{R}^n$  is open. So defining  $\tilde{G}(y) := (y, g(y))^T$  we have  $M \cap U = (P \circ \tilde{G})(V)$ . Let  $y_0$  be the point such that  $(P \circ \tilde{G})(y_0) = p_0$  and can take  $G(y) := (P \circ \tilde{G})$ . It is immediate to verify that  $DG_y$  has maximal rank for all  $y \in V$  (in particular at the point  $y_0$ , which is mapped to  $p_0$ ).

(3)  $\implies$  (1): Given  $G : V \rightarrow \mathbb{R}^n$ ,  $V \subset \mathbb{R}^d$  as in (3) we can pick  $n - d$  vectors  $v_{d+1}, \dots, v_n$  in  $\mathbb{R}^n$  such that the matrix

$$(DG_{y_0} \mid v_{d+1} \mid \dots \mid v_n)$$

is invertible (i.e., has full rank).

Consider the map  $\Phi \in C^1(V \times \mathbb{R}^{n-k}, \mathbb{R}^n)$  given by

$$\Phi(y_1, \dots, y_n) := f(y_1, \dots, y_d) + \sum_{j=d+1}^n y_j v_j.$$

Notice that  $D\Phi_{(y_0, 0)}$  is invertible by construction.

Then, by the Inverse function Theorem there is  $W \subset V \times \mathbb{R}^{n-k}$  open containing  $(y_0, 0)$ ,  $U \subset \mathbb{R}^n$  open containing  $p_0 = \Phi((y_0, 0))$  and  $\Psi : U \rightarrow W$  of class  $C^k$  such that

$$x = \Phi(\Psi(x)) = f(\Psi_1(x), \dots, \Psi_d(x)) + \sum_{j=d+1}^n \Psi_j(x) v_j.$$

Finally, since  $\Phi, \Psi$  are both injective it follows that

$$x \in M \cap U = f(V) \cap U \iff \Psi_{d+1}(x) = \dots = \Psi_n(x) = 0.$$

□

### 12.2.2 Some examples of submanifolds of $\mathbb{R}^3$

**Sphere ( $S^2$ ) in  $\mathbb{R}^3$**  The sphere  $S^2$  in  $\mathbb{R}^3$  serves as a quintessential example of a submanifold showcasing simple representation through implicit, parametric, and graphical forms.

- **Implicit Form:** In its implicit form, the sphere  $S^2$  can be described as the set of points  $(x, y, z) \in \mathbb{R}^3$  that satisfy the equation:

$$x^2 + y^2 + z^2 = 1.$$

This equation represents the set of all points in  $\mathbb{R}^3$  that are at a unit distance from the origin, defining the sphere.

- **Parametric Form:** A standard parametrization of the sphere is given

$$\begin{aligned}x &= \sin(\theta) \cos(\phi), \\y &= \sin(\theta) \sin(\phi), \\z &= \cos(\theta),\end{aligned}$$

where  $\theta \in [0, \pi]$  is the polar angle and  $\phi \in [0, 2\pi)$  is the azimuthal angle. This parametrization corresponds to the latitude and longitude coordinates used on the Earth.

This form, however, *does not cover the poles*  $\theta = 0$  and  $\theta = \pi$  with a single parameterization due to the singularity at the poles. To fully cover  $S^2$ , additional local parametrizations these points need to be used.

- **Graphical Form:** For the upper hemisphere, the sphere can be graphically represented as:

$$z = \sqrt{1 - x^2 - y^2},$$

for  $x^2 + y^2 \leq 1$ . This equation describes how the  $z$ -coordinate depends on  $x$  and  $y$  coordinates. However, this form *only represents the upper hemisphere*. The lower hemisphere would similarly be represented by  $z = -\sqrt{1 - x^2 - y^2}$ , indicating the need for multiple functions to describe the entire surface. To cover all the surface with smooth graphs, we must also introduce functions for the  $y$  and  $x$  coordinates as functions of the other two variables. Specifically, we can use:

$$y = \pm\sqrt{1 - x^2 - z^2}$$

for portions of the sphere where  $y$  is the dependent variable, and

$$x = \pm\sqrt{1 - y^2 - z^2}$$

for parts of the sphere with  $x$  as the dependent variable. These representations allow us to graphically depict the entire sphere by selecting the appropriate function based on the region of interest, ensuring smoothness across all points on  $S^2$ .

For the implicit representation  $x^2 + y^2 + z^2 = 1$ , the gradient vector, serving as the Jacobian matrix for scalar functions, is derived from  $f(x, y, z) = x^2 + y^2 + z^2 - 1$ . The gradient is:

$$\nabla f = \begin{pmatrix} \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial y} \\ \frac{\partial f}{\partial z} \end{pmatrix} = \begin{pmatrix} 2x \\ 2y \\ 2z \end{pmatrix}.$$

This gradient vector is non-zero everywhere on  $S^2$ , ensuring that the Jacobian has maximal rank (1 in this case) across the surface defined by  $f(x, y, z) = 0$ .



For the parametric representation:

$$\begin{aligned}x &= \sin(\theta) \cos(\phi), \\y &= \sin(\theta) \sin(\phi), \\z &= \cos(\theta),\end{aligned}$$

the Jacobian matrix  $J$  is given by:

$$J = \begin{pmatrix} \frac{\partial x}{\partial \theta} & \frac{\partial x}{\partial \phi} \\ \frac{\partial y}{\partial \theta} & \frac{\partial y}{\partial \phi} \\ \frac{\partial z}{\partial \theta} & \frac{\partial z}{\partial \phi} \end{pmatrix} = \begin{pmatrix} \cos(\theta) \cos(\phi) & -\sin(\theta) \sin(\phi) \\ \cos(\theta) \sin(\phi) & \sin(\theta) \cos(\phi) \\ -\sin(\theta) & 0 \end{pmatrix}.$$

To verify this matrix has maximal rank (2, since it's a  $3 \times 2$  matrix), consider the determinant of a  $2 \times 2$  submatrix, for example:

$$\det \begin{pmatrix} \cos(\theta) \cos(\phi) & -\sin(\theta) \sin(\phi) \\ \cos(\theta) \sin(\phi) & \sin(\theta) \cos(\phi) \end{pmatrix} = \cos^2(\theta) + \sin^2(\theta) = 1,$$

which is non-zero for all  $\theta \in (0, \pi)$ , thus demonstrating the Jacobian has maximal rank for almost all values of  $\theta$  and  $\phi$ , excluding the poles where other charts are necessary.

**Torus ( $T^2$ ) in  $\mathbb{R}^3$ .** A torus can be parametrically represented by:

$$\begin{aligned}x(u, v) &= (R + r \cos v) \cos u, \\y(u, v) &= (R + r \cos v) \sin u, \\z(u, v) &= r \sin v,\end{aligned}$$

where  $u, v \in [0, 2\pi)$ ,  $R$  is the distance from the center of the tube to the center of the torus, and  $r$  is the radius of the tube.

The Jacobi matrix is:

$$\begin{pmatrix} -\sin(u)(R + r \cos(v)) & -r \sin(v) \cos(u) \\ \cos(u)(R + r \cos(v)) & -r \sin(v) \sin(u) \\ 0 & r \cos(v) \end{pmatrix}.$$

Its rank is maximal (2) for all  $u, v$ .

**Möbius Strip in  $\mathbb{R}^3$ .** The Möbius strip is a so-called non-orientable surface that can be parametrically represented as:

$$\begin{aligned}x(u, v) &= \left(1 + \frac{v}{2} \cos \frac{u}{2}\right) \cos u, \\y(u, v) &= \left(1 + \frac{v}{2} \cos \frac{u}{2}\right) \sin u, \\z(u, v) &= \frac{v}{2} \sin \frac{u}{2},\end{aligned}$$

where  $u \in [0, 2\pi)$  and  $v \in [-1, 1]$ .

The Jacobi matrix is:

$$\begin{pmatrix} -\sin(u) \left(1 + \frac{v}{2} \cos \frac{u}{2}\right) - \frac{v}{2} \sin \frac{u}{2} \cos(u) & \frac{1}{2} \cos \frac{u}{2} \cos(u) \\ \cos(u) \left(1 + \frac{v}{2} \cos \frac{u}{2}\right) - \frac{v}{2} \sin \frac{u}{2} \sin(u) & \frac{1}{2} \cos \frac{u}{2} \sin(u) \\ \frac{1}{2} \cos \frac{u}{2} & 0 \end{pmatrix}.$$

Its rank is maximal (2).

**Catenoid in  $\mathbb{R}^3$ .** The catenoid is a so-called ‘minimal surface’. It can be represented in parametric form by:

$$x(u, v) = \cosh(v) \cos(u),$$

$$y(u, v) = \cosh(v) \sin(u),$$

$$z(u, v) = v,$$

where  $u \in [0, 2\pi)$  and  $v \in \mathbb{R}$ , with  $\cosh$  denoting the hyperbolic cosine function.

The Jacobi matrix is.

$$\begin{pmatrix} -\sin(u) \cosh(v) & \cos(u) \sinh(v) \\ \cos(u) \cosh(v) & \sin(u) \sinh(v) \\ 0 & 1 \end{pmatrix}.$$

Its rank is maximal (2) across all  $u, v$ .

---

## Chapter 13

# Multidimensional Integration

### 13.1 The Jordan measure in $\mathbb{R}^n$

In this section we introduce the Jordan measure and its first properties.

13.1. — The following notation will be useful throughout the section. If  $X \subset \mathbb{R}^n$  is a some set  $\varrho > 0$  and  $a \in \mathbb{R}^n$  we will use the notations

$$\varrho X + a \quad \text{or} \quad a + \varrho X$$

18

to refer to the set  $\{\varrho x + a \mid x \in X\}$ . This is a compact way to refer to dilations and translations of  $X$ :

Notice that with this notation we can operate with sets easily according to the usual rules of sum and multiplication, and see what is the compounded effect of some dilations and translation: for example, given  $\varrho, \varrho' > 0$  and  $a, a' \in \mathbb{R}^n$  we have

$$\varrho'(a + \varrho X) + a' = (\varrho'a + a') + \varrho'\varrho X$$

since this simply boils down to

$$\{\varrho'(a + \varrho x) + a' \mid x \in X\} = \{(\varrho'a + a') + \varrho'\varrho x \mid x \in X\}.$$

## DEFINITION 13.2: DYADIC CUBES, DYADIC SUBSETS, AND THEIR MEASURE

We call a subset of  $\mathbb{R}^n$  of the form

$$Q = 2^{-p}(a + [0, 1)^n) := \prod_{i=1}^n [2^{-p}a_i, 2^{-p}(a_i + 1)),$$

for some  $a = (a_1, a_2, \dots, a_n) \in \mathbb{Z}^n$  and  $p \in \mathbb{N}$ , **dyadic cube** of side length  $2^{-p}$ .

Any finite union (possibly empty) of dyadic cubes of the same side length is called **dyadic set** of **pixel size**  $2^{-p}$ . Notice that such set is determined by giving  $p, N \in \mathbb{N}$  and an injective map  $a : \{1, 2, \dots, N\} \rightarrow \mathbb{Z}^k$ :

$$E = \bigcup_{\ell=1}^N 2^{-k}(a(\ell) + [0, 1)^n). \quad (13.1)$$

If  $Q$  is a dyadic cube of side length  $2^{-k}$  we define its  $n$ -volume, denoted  $\mu_n$  or  $\mu$  as  $\mu(Q) := 2^{-kn}$  (we will check below that it is well-defined). Also if  $E$  is a dyadic set of the form (13.1) we define  $\mu(E) := 2^{-pn}N$ .

By definition, the empty set, denoted  $\emptyset$ , is a dyadic set. It corresponds to  $N = 0$  and hence it has zero volume:  $\mu(\emptyset) = 0$ .

18

## LEMMA 13.3: DYADIC REFINEMENTS

*Assume*

$$E = \bigcup_{\ell=1}^N 2^{-p}(a(\ell) + [0, 1)^n),$$

*for some  $a : \{1, 2, \dots, N\} \rightarrow \mathbb{Z}^n$  injective.*

*Then for every  $q > p$  we can decompose*

$$E = \bigcup_{\ell=1}^{2^{(q-p)n}N} 2^{-q}(b(\ell) + [0, 1)^n),$$

*for a suitable  $b : \{1, 2, \dots, 2^{(q-p)n}N\} \rightarrow \mathbb{Z}^n$  injective.*

*Proof.* It easily follows from the decomposition

$$[0, 1)^n = \bigcup \left\{ 2^{-(q-p)}(z + [0, 1)^n) \mid z \in (\mathbb{Z} \cap [0, 2^{q-p}))^n \right\}. \quad (13.2)$$

□

PROPOSITION 13.4: PROPERTIES OF  $\mu$ 

Unions, intersections, and differences of dyadic sets (with possibly different pixel sizes) are dyadic sets. (More precisely, if  $E$  and  $F$  are dyadic sets then also  $E \cup F$ ,  $E \setminus F$ ,  $F \setminus E$ , and  $E \cap F$  are).

The measure  $\mu = \mu_n$ , is well-defined over dyadic sets satisfies the following properties:

1. *Additivity:*  $\mu(E \cup F) = \mu(E) + \mu(F)$  for all  $E, F$  dyadic and disjoint.
2. *Normalization:*  $\mu([0, 1]^n) = 1$
3. *Translation and scaling invariance:*  $\mu(\varrho E + \tau) = \varrho^n \mu(E)$ , for all  $E$  dyadic,  $\varrho$  belonging to  $2^{-p}\mathbb{N}$  for some  $p \in \mathbb{N}$ , and  $\tau$  belonging to  $2^{-p}\mathbb{Z}^n$  for some  $p \in \mathbb{N}$ ;

*Proof.* First, we observe that the measure  $\mu(E)$  is well-defined: it does not depend on the size of the pixel chosen to represent  $E$ . Indeed, if  $E$  is initially represented using pixels of size  $2^{-p}$ , and these pixels are further subdivided into smaller pixels of size  $2^{-q}$  where  $q > p$ , then each original cube  $Q_0$  with side length  $2^{-p}$  is divided into  $2^{(q-p)n}$  smaller cubes  $Q_i$  of side length  $2^{-q}$ , as shown in (13.2). Consequently, the measure of  $Q_0$ , which is by definition is  $2^{-pn}$ , equals the aggregate measure of the smaller cubes, given by

$$2^{-pn} = 2^{(q-p)n} \cdot 2^{-qn} = \sum_{i=1}^{2^{(q-p)n}} \mu(Q_i).$$

Hence, the measure  $\mu(E)$  for the dyadic set  $E$  remains consistent, regardless of the chosen pixel size for its representation.

By applying Lemma 13.3, we can adjust refine one of the two dyadic sets, if necessary, to ensure both sets have the same pixel size, specifically by choosing the smaller pixel size from the two. Consequently, we ascertain that the union and intersection of any two given dyadic sets are themselves dyadic. Furthermore, the measure  $\mu$  exhibits additivity, which naturally stems from counting the number of pixels in each set. The normalization of  $\mu$  is established by its definitional properties.

To show the translation and rotation invariance we first consider the case  $E_\circ = 2^{-q}(a + [0, 1]^n)$ . Then, for all  $m \in \mathbb{N}$ ,  $p \in \mathbb{N}$  and  $b \in \mathbb{Z}^n$  we have

$$\begin{aligned} \varrho E_\circ + \tau &= m2^{-p}E + 2^{-p}b = m2^{-p}2^{-q}(a + [0, 1]^n) + 2^{-p}b = 2^{-(p+q)}(ma + 2^q b + m[0, 1]^n) \\ &= \bigcup_{z \in \mathbb{Z}^n \cap [0, m]^n} 2^{-(p+q)}(ma + 2^q b + z + [0, 1]^n) \end{aligned}$$

and hence

$$\mu(\varrho E_\circ + \tau) = 2^{-(p+q)n} \#\{\mathbb{Z}^n \cap [0, m]^n\} = 2^{-(p+q)n} m^n = \varrho^n 2^{-qn} = \varrho^n \mu(E).$$

From this property for cubes we deduce the property for general dyadic sets using additivity: if  $E = \bigcup_{\ell=1}^N 2^{-k}(a(\ell) + [0, 1]^n)$  then

$$\varrho E + \tau = \bigcup_{\ell=1}^N 2^{-k}(\varrho a(\ell) + \tau \varrho[0, 1]^n)$$

and the union is disjoint. Therefore,

$$\mu(\varrho E + \tau) = \sum_{\ell=1}^N \mu\left(2^{-k}(\varrho a(\ell) + \tau \varrho[0, 1]^n)\right) = \sum_{\ell=1}^N \varrho^n 2^{-kn} = \varrho^n \mu(E).$$

□

13.5. — One can show (exercise) that  $\mu$  is the only measure (i.e., a map assigning to each subset in a given collection a nonnegative real number) defined on the dyadic sets and satisfying properties (1)-(3).

EXERCISE 13.6. — Show that, for two dyadic sets  $E_1$  and  $E_2$ , the measure of their union is given by:

$$\mu(E_1 \cup E_2) = \mu(E_1) + \mu(E_2) - \mu(E_1 \cap E_2)$$

This formula ensures that the overlapping part of  $E_1$  and  $E_2$  is not counted twice.

Show that, for  $N$  dyadic sets, the measure of their union is determined by the principle of inclusion-exclusion:

$$\mu\left(\bigcup_{i=1}^N E_i\right) = \sum_{k=1}^N (-1)^{k+1} \left( \sum_{1 \leq i_1 < i_2 < \dots < i_k \leq N} \mu(E_{i_1} \cap E_{i_2} \cap \dots \cap E_{i_k}) \right).$$

#### DEFINITION 13.7: JORDAN MEASURABLE SETS AND THEIR MEASURE

Given a set  $E \subset \mathbb{R}^n$  let us define its inner and outer volumes as:

$$\mu_{\text{in}}(E) = \sup \{ \mu(F) \mid F \subset E, F \text{ dyadic set} \}$$

and

$$\mu_{\text{out}}(E) = \inf \{ \mu(F) \mid E \subset F, F \text{ dyadic set} \}$$

If  $\mu_{\text{in}}(E) = \mu_{\text{out}}(E)$  then we say that  $E$  is **Jordan measurable**. We then denote  $\mu_{\text{in}}(E) = \mu_{\text{out}}(E)$  by  $\mu(E)$ ,  $\mu_n(E)$ , or  $\text{vol}_n(E)$ .

## DEFINITION 13.8: JORDAN AND LEBESGUE NULL SETS

We say that  $E \subset \mathbb{R}^n$  is a **Jordan null** if  $\mu_{\text{out}}(E) = 0$ . Also, we say that  $E$  is **Lebesgue null** if for all  $\varepsilon > 0$  exist countably many dyadic cubes  $Q_i$  (possibly with different side sizes) such that

$$E \subset \cup_i Q_i \quad \text{and} \quad \sum_i \mu(Q_i) < \varepsilon.$$

13.9. — Notice that by definition every Jordan null is Jordan measurable (with zero volume) and is also Lebesgue null.

Clearly, Jordan null sets must be bounded, but this is not the most crucial difference with respect to Lebesgue null sets: Several bounded sets are Lebesgue null but not Jordan null. For example  $E := \mathbb{Q}^n \cap [0, 1]^n$  is a bounded subset of  $\mathbb{R}^n$  is Lebesgue null as  $\mathbb{Q}^n$  is countable so there exist a surjective map from  $f : \mathbb{N} \rightarrow \mathbb{Q}^n$ . For all  $i \in \mathbb{N}$  let  $Q_i$  be the dyadic cube of side length  $2^{-N-i-1}$  that contains  $f(i)$ . Then  $E \subset \cup_i Q_i$  but

$$\sum_i \mu(Q_i) = \sum_{i=0}^{\infty} (2^{-N-i})^n \leq 2^{-N}$$

for all  $n \geq 1$ . Since taking  $N$  large we can make  $2^{-N} < \varepsilon$ , so  $E$  is indeed Lebesgue null.

18

## LEMMA 13.10: COMPACT LEBESGUE NULL IS JORDAN NULL

If  $K \subset \mathbb{R}^n$  is compact and Lebesgue null then it is also Jordan null.

*Proof.* Suppose  $K$  is a Lebesgue null set. For any given  $\varepsilon > 0$ , it is possible to cover  $K$  with finitely many dyadic cubes  $Q_i$  such that  $K \subset \cup_{i=1}^{\infty} Q_i$  and

$$\sum_{i=1}^{\infty} \mu(Q_i) < \frac{\varepsilon}{3^n}.$$

For each dyadic cube  $Q_i$ , define  $Q_i^*$  as the open cube with the same center as  $Q_i$ , but with each side length tripled. Explicitly, if  $Q_i = 2^{-k_i}(a_i + [0, 1]^n)$ , then  $Q_i^* = \bigcup_{h \in \{-1, 0, 1\}^n} 2^{-k_i}(a_i + h + [0, 1]^n)$ . In particular we have  $\mu(Q_i^*) = 3^n \mu(Q_i)$ .

Notice now that  $K$  is contained within the union of the interiors of these expanded cubes,  $K \subset \bigcup_{i=1}^{\infty} \text{int}(Q_i^*)$ . Given the compactness of  $K$ , a finite subcover can be selected,  $K \subset \bigcup_{\ell=1}^N Q_{i_\ell}^*$ . Thus,

$$\sum_{\ell=1}^N \mu(Q_{i_\ell}^*) < \varepsilon.$$

Therefore, by demonstrating that  $K$  can be covered by a finite collection of dyadic cubes with a total measure less than any arbitrary  $\varepsilon$ , we conclude that  $K$  is Jordan null. □

EXERCISE 13.11. — Prove that finite (respectively countable) unions of Jordan (respectively Lebesgue) null sets are also null sets.

**PROPOSITION 13.12: BOUNDED JORDAN MEASURABLE SETS**

*A bounded subset  $E$  of  $\mathbb{R}^n$  is Jordan measurable if and only if  $E$  is bounded and  $\partial E$  is a Lebesgue null set.*

Before giving this proof we need three short lemmas.

**LEMMA 13.13: MEASURE OF INTERIOR AND CLOSURE OF DYADIC SET**

*Let  $E$  be a dyadic set. Then its both its interior  $\text{int}(E)$  and closure  $\overline{E}$  are Jordan measurable and  $\mu(\text{int}(E)) = \mu(\overline{E}) = \mu(E)$ .*

*Proof.* Assume  $E = \bigcup_{\ell=1}^N 2^{-p}(a(\ell) + [0, 1]^n)$ , where  $a : \{1, \dots, N\} \rightarrow \mathbb{Z}^n$  is injective and  $p \in \mathbb{N}$ . Given  $k \geq 2$ , Consider the dyadic sets

$$F_k = \bigcup_{\ell=1}^N \bigcup_{h \in \{1, 2, \dots, 2^k - 1\}^n} 2^{-p}(a(\ell) + 2^{-k}h + 2^{-k}[0, 1]^n),$$

and

$$G_k = \bigcup_{\ell=1}^N \bigcup_{h \in \{0, 1, 2, \dots, 2^k\}^n} 2^{-p}(a(\ell) + 2^{-k}h + 2^{-k}[0, 1]^n),$$

Then

$$F \subset \text{int}(E) \subset E \subset \overline{E} \subset G.$$

But

$$\mu(G_k) = N(2^k + 1)^n 2^{-kpn} \quad \text{and} \quad \mu(F) = N(2^k - 1)^n 2^{-kpn}$$

Sending  $k \rightarrow \infty$  we have  $\lim_k \mu(G_k) = \lim_k \mu(F_k) = 2^{-pn}$ . This shows that both  $\text{int}(Q)$  and  $\overline{Q}$  are Jordan measurable and have the same measure as  $Q$ .  $\square$

**LEMMA 13.14: A PIXEL IS EITHER INSIDE OR OUTSIDE A DYADIC SET**

*Suppose that  $Q$  is a dyadic cube of side length  $2^{-p}$  and  $F$  is some dyadic set of pixel size  $2^{-p}$  then either  $Q \subset F$  or  $Q \cap F = \emptyset$ .*

*Proof.* By definition  $Q = 2^{-p}(a_o + [0, 1]^n)$  for some  $a_o \in \mathbb{Z}^n$  and  $F = \bigcup_{\ell=1}^N 2^{-p}(a(\ell) + [0, 1]^n)$  with  $a : \{1, \dots, N\} \rightarrow \mathbb{Z}^n$  is injective. So either  $a_o$  belongs to the image of  $a$ , in which case  $Q \subset F$ , or not, in which case  $Q \cap F = \emptyset$ .  $\square$



## LEMMA 13.15: SANDWICH

If  $E \subset \mathbb{R}^n$  is a subset and there are sequences of sets  $F_k, G_k$  such that  $F_k \subset E \subset G_k$  such that  $\lim_k \mu_{\text{out}}(G_k) - \mu_{\text{in}}(F_k) = 0$  then  $E$  is Jordan measurable and

$$\mu(E) = \inf_k \mu_{\text{out}}(G_k) = \sup_k \mu_{\text{in}}(F_k). \quad (13.3)$$

*Proof.* For all  $\epsilon > 0$  there exist  $k \in \mathbb{N}$  and  $\tilde{F}_k \subset F_k$  and  $\tilde{G}_k \supset G_k$  dyadic sets such that

$$\mu_{\text{out}}(G_k) - \mu_{\text{in}}(F_k) < \epsilon/3$$

$$\mu(\tilde{G}_k) - \mu_{\text{out}}(G_k) < \epsilon/3$$

$$\mu_{\text{in}}(F_k) - \mu(\tilde{F}_k) < \epsilon/3.$$

Hence,  $\tilde{F}_k \subset E \subset \tilde{G}_k$  and  $\mu(\tilde{G}_k) - \mu(\tilde{F}_k) < \epsilon$ , and since  $\epsilon$  can be made arbitrarily small this proves that the set is measurable. Also, (13.3) follows from  $\mu(\tilde{G}_k) \leq \mu(E) \leq \mu(\tilde{F}_k)$ .  $\square$

*Proof of Proposition 13.12.* Observe first that the boundary  $\partial E$  of a bounded set is compact. So the boundary is Lebesgue null if and only if it is Jordan null.

For any given  $p \in \mathbb{N}$  (large), put

$$Q_p(a) := 2^{-p}(a + [0, 1)^n),$$

where  $a \in \mathbb{Z}^n$ , and define

$$F_p := \bigcup \{Q_p(a) \mid Q_p(a) \subset E, a \in \mathbb{Z}^n\}$$

$$G_p := \bigcup \{Q_p(a) \mid Q_p(a) \cap E \neq \emptyset, a \in \mathbb{Z}^n\}$$

Notice that

$$F_p \subset E \subset G_p$$

Moreover, by Lemma 13.14,  $F_p$  is the largest dyadic set of pixel size  $2^{-p}$  contained in  $E$ ; while  $G_p$  is the smallest dyadic set of pixel size  $2^{-p}$  containing  $E$ .

Suppose first that  $\partial E$  is a Jordan null set and let us show that  $E$  is Jordan measurable. Indeed, for each  $\epsilon > 0$  there is  $p \in \mathbb{N}$  and  $H_p$  dyadic with pixel size  $2^{-p}$  such that

$$\partial E \subset H_p \quad \text{and} \quad \mu(H_p) < \epsilon$$

But notice that  $G_p \setminus F_p \subset H_p$ . Indeed, if  $Q_p(a) \subset G_p \setminus F_p$  then  $Q_p(a)$  contains simultaneously a point  $x \in E$  and a point  $y \in \mathbb{R}^n \setminus E$ . Then  $Q_p(a)$  also contains the point

$$z_* := (1 - t_*)x + t_*y \quad \text{where} \quad \sup \{t \in (0, 1) \mid (1 - t)x + ty \in E\}.$$

But  $z_*$  belongs to  $\partial E$  being an accumulation point of both  $E$  and  $\mathbb{R}^n \setminus E$ . Hence  $z_* \in Q_p(a) \cap \partial E$  and thus  $Q_p(a) \subset I_p$  (by Lemma 13.14).

Then, since  $\mu$  is additive on dyadic sets, we have

$$\mu_{\text{out}}(E) - \mu_{\text{in}}(E) \leq \mu(G_p) - \mu(F_p) \leq \mu(H_p) < \epsilon$$

Since  $\epsilon > 0$  can be made arbitrarily small it follows that  $E$  is Jordan measurable.

To establish the converse implication, we notice that if  $E$  is Jordan measurable, then there exists  $p$  such that

$$\mu(G_p) - \mu(F_p) \leq \epsilon.$$

The only issue now to conclude that that  $\partial E$  is a nulls set, is that  $\partial E$  may not be contained in  $G_p \setminus F_p$ . (For example, when  $E = [0, 1]^n$  then  $F_p = E$  and  $G_p = \emptyset$  for all  $p$ )

However, by Lemma 13.13, the  $\text{int}(F_p)$  and  $\overline{G_p}$  are Jordan measurable and have the same measure as  $F_p$  and  $G_p$ , respectively. But  $\partial E \subset \overline{G_p} \setminus \text{int}(F_p)$  and  $\mu(\overline{G_p} \setminus \text{int}(F_p)) = \mu(\overline{G_p}) - \mu(\text{int}(F_p)) < \epsilon$

Since  $\epsilon > 0$  can be made arbitrarily small it follows that  $\partial E$  is a Jordan null set.  $\square$

In order to proof the proposition we will need the following

**LEMMA 13.16: LIPSCHITZ MAPS PRESERVE NULL SETS**

*Suppose that  $m \leq n$ , and let  $U \subset \mathbb{R}^m$  open,  $E \subset U \subset \mathbb{R}^m$  a Jordan null set (in  $\mathbb{R}^m$ , i.e. with respect to the measure  $\mu_m$ ), and  $f : U \rightarrow \mathbb{R}^n$  a Lipschitz map. Then  $f(E) \subset \mathbb{R}^n$  is also Jordan null*

*Proof.* Since  $\mu(E) = 0$ , given  $\epsilon > 0$  exist disjoint dyadic cubes  $Q_\ell := 2^{-p}(a(\ell) + [0, 1]^m) \subset \mathbb{R}^m$  covering  $E$  such that  $\sum_{\ell=1}^N \mu_m(Q_\ell) < \epsilon$ .

If  $\Lambda$  is the Lipschitz constant of  $f$  then for all  $x \in Q_\ell \cap U$  we have

$$|f(x) - f(2^{-p}a(\ell))| \leq \Lambda|x - 2^{-p}a(\ell)| \leq \sqrt{n}\Lambda 2^{-p},$$

where  $\Lambda$  is the Lipschitz constant of  $f$  (in  $U$ ).

Hence, fixing  $p_0 \geq 1$  (depending only on  $n$  and  $\Lambda$ ) so that  $\sqrt{n}(\Lambda + 1) < 2^{p_0}$  and picking  $y_\ell \in 2^{-p}\mathbb{Z}^n$  such that  $y_\ell \in f(2^{-p}a(\ell)) + 2^{-p}[0, 1]^n$ , we obtain  $f(Q_\ell \cap U) \subset \tilde{Q}_\ell := y_\ell + 2^{p_0-p}[0, 1]^n$  (Indeed, for all  $x \in Q_\ell \cap U$  we have  $|f(x) - y_\ell| \leq |f(x) - f(2^{-p}a(\ell))| + \sqrt{n}2^{-p} \leq \sqrt{n}(\Lambda + 1)2^{-p}$ .)

Hence,

$$f(E) \subset \bigcup_{\ell=1}^N \tilde{Q}_\ell$$

But notice that each  $\tilde{Q}_\ell$  is a union of dyadic cubes of  $\mathbb{R}^n$  and

$$\sum_{\ell=1}^N \mu_n(\tilde{Q}_\ell) \leq N 2^{(p_0-p)n} \leq N 2^{p_0 n} 2^{-pm} = 2^{p_0 n} \sum_{\ell=1}^N \mu_m(Q_\ell) < 2^{p_0 n} \epsilon.$$

Hence, since  $\epsilon > 0$  can be made arbitrarily small we conclude that  $f(E)$  is a null set (in  $\mathbb{R}^n$ ).  $\square$

**LEMMA 13.17: GRAPHS OF UNIFORMLY CONTINUOUS MAPS ARE NULL**

Suppose that  $E \subset \mathbb{R}^n$  is a bounded set and  $f : E \rightarrow \mathbb{R}$  is a uniformly continuous map (i.e., for all  $\epsilon > 0$  exists  $\delta > 0$  such that  $|x - x'| < \delta$ ,  $x, x' \in E \implies |f(x) - f(x')| < \epsilon$ ). Then,  $\{(x, f(x)) \mid x \in E\} \subset \mathbb{R}^{n+1}$  is Jordan null.

*Proof.* Since  $E$  is bounded we have  $E \subset [-N_o, N_o]^n$ , for some  $N_o \in \mathbb{N}$ .

Fix  $\epsilon = 2^{-q} > 0$ , where  $q \in \mathbb{N}$ . Since  $f$  is uniformly continuous in  $E$ , there exists  $\delta > 0$  such that if  $x, x' \in E$  are such that  $|x - x'| < \delta$ , then  $|f(x) - f(x')| < \epsilon$ . Take  $p \in \mathbb{N}$  such that  $\sqrt{n}2^{-p} < \delta$ . Assume without loss of generality that  $p \geq q$ .

We can then cover  $E$  by (a number  $N \leq (2^{p+1}N_o)^n$  of) disjoint union of cubes  $Q_\ell = 2^{-p}(a(\ell) + [0, 1]^n)$ , where  $a(\ell) \in \mathbb{Z}^n$ . We can assume that each cube  $Q_\ell$  has nonempty intersection with  $E$ . For each  $\ell$ , let  $x_\ell$  be some point belonging to  $E \cap Q_\ell$  and let  $b_\ell \in \mathbb{Z}$  be such that  $f(x_\ell) \in 2^{-q}(b_\ell + [0, 1])$ .

Then, for all  $x \in E \cap Q_\ell$  we have

$$|f(x) - 2^{-q}b_\ell| < |f(x) - f(x_\ell)| + |f(x_\ell) - 2^{-q}b_\ell| \leq 2^{-q},$$

where we used that  $|x - x_\ell| \leq \sqrt{n}2^{-p}$  for all  $x \in Q_\ell$  (and that  $\sqrt{n}2^{-p} < \delta$ ).

Therefore,

$$\{(x, f(x)) \mid x \in E\} \subset F := \bigcup_{\ell=1}^N Q_\ell \times 2^{-q}[b_\ell - 2, b_\ell + 2]$$

Now, for each  $\ell$ , the ‘vertical column above  $Q_\ell$ ’,  $Q_\ell \times 2^{-q}[b_\ell - 2, b_\ell + 2]$ , can be written as:

$$\bigcup_{-22^{p-q} \leq k < 22^{p-q}} 2^{-p}((a_\ell, k) + [0, 1]^{n+1}).$$

That is, this ‘vertical column above  $Q_\ell$ ’ is the union of  $42^{p-q}$  cubes of side length  $2^{-p}$  in  $\mathbb{R}^{n+1}$ , the set  $F$  is dyadic and

$$\mu_{n+1}(F) = 4N2^{p-q}2^{-pn+1} = 4(2N_o)^n2^{-q} < 4(2N_o)^n\epsilon.$$

Sending  $\epsilon \rightarrow 0$  we obtain that the graph of  $f$  is Jordan null.  $\square$

## THEOREM 13.18: FIRST PROPERTIES OF THE JORDAN MEASURE

The Jordan measure  $\mu = \text{vol}_n$ , defined over all Jordan measurable sets  $E \subset \mathbb{R}^n$  satisfies the following properties.

1. *Additivity:* If  $E, F$  are Jordan measurable  $E \cup F, E \cap F, E \setminus F$  and  $F \setminus E$  are Jordan measurable. Moreover, if  $E$  and  $F$  are disjoint:  $\mu(E \cup F) = \mu(E) + \mu(F)$ .
2. *Normalization:*  $\text{vol}_n([0, 1]^n) = 1$
3. *Volume of boxes:* if

$$E = [a_1, b_1) \times [a_2, b_2) \times \cdots \times [a_n, b_n),$$

with  $b_i > a_i$  for  $1 \leq i \leq n$  then  $\mu(E) = (b_1 - a_1)(b_2 - a_2) \cdots (b_n - a_n)$ .

*Proof.* The boundary of any of the four sets  $E \cup F, E \cap F, E \setminus F$  and  $F \setminus E$  is a subset of  $\partial E \cup \partial F$  (exercise). If  $E, F$  are Jordan measurable then  $\partial E$  and  $\partial F$  are Jordan null, by Lemma 13.12, and hence their union and the boundaries of the previous four sets are Jordan null. Hence, using Lemma 13.12 again each of the previous four sets is Jordan measurable.

Now assume that  $E_1, E_2$  are Jordan measurable and disjoint. For any given  $\epsilon > 0$  there are dyadic sets  $F_i, G_i$  such that

$$F_i \leq E_i \leq G_i, \quad \mu(F_i) \leq \mu(E_i) \leq \mu(G_i) \leq \mu(F_i) + \epsilon.$$

But then

$$F_1 \cup F_2 \subset E_1 \cup E_2 \subset G_1 \cup G_2$$

and  $F_1, F_2$  are disjoint. Thus,

$$\mu(E_1 \cup E_2) \geq \mu(F_1) + \mu(F_2) \geq \mu(E_1) - \epsilon + \mu(E_2) - \epsilon;$$

and

$$\mu(E_1 \cup E_2) \leq \mu(G_1) + \mu(G_2) \leq \mu(E_1) + \epsilon + \mu(E_2) + \epsilon.$$

Therefore,

$$-2\epsilon \leq \mu(E_1) + \mu(E_2) - \mu(E_1 \cup E_2) \leq 2\epsilon.$$

Sending  $\epsilon \rightarrow 0$  we conclude that additivity property  $\mu(E_1) + \mu(E_2) - \mu(E_1 \cup E_2) = 0$ .

Finally suppose that

$$E = [a_1, b_1) \times [a_2, b_2) \times \cdots \times [a_n, b_n),$$

with  $b_i > a_i$  for  $1 \leq i \leq n$ . For  $t \in \mathbb{R}$ , recall that the floor and ceiling of  $t$  are respectively defined as:

$$\lfloor t \rfloor = \max \{k \in \mathbb{Z} \mid k \leq t\} \quad \text{and} \quad \lceil t \rceil = \min \{k \in \mathbb{Z} \mid k \geq t\}.$$

Given  $p \in \mathbb{N}$  let us define

$$\lfloor t \rfloor_p := 2^{-p} \lfloor 2^p t \rfloor \quad \text{and} \quad \lceil t \rceil_p := 2^{-p} \lceil 2^p t \rceil.$$

For given  $p \in \mathbb{N}$  consider the two dyadic sets

$$F_p = \prod_{i=1}^n [\lfloor a_i \rfloor_p, \lfloor b_i \rfloor_p)$$

and

$$G_p = \prod_{i=1}^n [\lfloor a_i \rfloor_p, \lceil b_i \rceil_p).$$

Then,  $F_p \subset E \subset G_p$  and (for  $p$  large enough so that  $2^{1-p} < \min_i (b_i - a_i)$ )

$$\prod_{i=1}^n (b_i - a_i - 2^{1-p}) \leq \mu(F_p) \leq \mu(E) \leq \mu(G_p) \leq \prod_{i=1}^n (b_i - a_i + 2^{1-p}).$$

Sending  $p \rightarrow \infty$  we conclude  $\mu(E) = \prod_{i=1}^n (b_i - a_i)$ .  $\square$

EXERCISE 13.19. — Prove that if  $f : \mathbb{R}^n \rightarrow \mathbb{R}^n$  is a continuous map with continuous inverse then for every set  $E \subset \mathbb{R}^n$  we have  $f(\overline{E}) = \overline{f(E)}$ ,  $f(\text{int}(E)) = \text{int}(f(E))$  and  $f(\partial E) = \partial f(E)$ .

**LEMMA 13.20: TRANSLATION AND DILATION INVARIANCE**

For all  $E$  Jordan measurable,  $\varrho > 0$  and  $\tau \in \mathbb{R}^n$ ,  $\varrho E + \tau$  is also Jordan measurable and  $\mu(\varrho E + \tau) = \varrho^n \mu(E)$

*Proof.* Since the map  $x \mapsto \varrho x + \tau$  is Lipschitz and  $\partial(\varrho E + \tau) = \varrho \partial E + \tau$  we obtain, using Lemma 13.12 that  $\varrho E + \tau$  is Jordan measurable.

Now, putting  $Q_p := 2^{-p}[0, 1]^n$  the formula for the measure of boxes implies that  $\mu(\varrho Q_p + \tau) = \mu([0, \varrho 2^{-p}]^n) = (\varrho 2^{-p})^n = \varrho^n \mu(Q_p)$  for every translation  $\tau \in \mathbb{R}^n$  and dilation factor  $\varrho > 0$ . Then, since  $\mu$  is additive,  $\mu(\varrho F + \tau) = \varrho^n \mu(F)$  for all  $F$  dyadic (since  $F$  is, by definition, a disjoint union of dyadic cubes).

Finally, for all  $E$  Jordan measurable, given  $\epsilon > 0$  there exist dyadic sets  $F_\epsilon$  and  $G_\epsilon$  such that  $F_\epsilon \subset E \subset G_\epsilon$  and  $\mu(F_\epsilon) \leq \mu(E) \leq \mu(G_\epsilon) \leq \mu(F_\epsilon) + \epsilon$ . But then  $(\varrho F_\epsilon + \tau) \subset (\varrho E + \tau) \subset (\varrho G_\epsilon + \tau)$  and thus  $\varrho^n \mu(F_\epsilon) = \mu(\varrho F_\epsilon + \tau) \leq \mu(\varrho E + \tau) \leq \mu(\varrho G_\epsilon + \tau) = \varrho^n \mu(G_\epsilon) \leq \varrho^n (\mu(F_\epsilon) + \epsilon)$ . Sending  $\epsilon \rightarrow 0$  (since  $\mu(F_\epsilon) \rightarrow \mu(E)$ ) we conclude that  $\mu(\varrho E + \tau) = \varrho^n \mu(E)$ .  $\square$

## PROPOSITION 13.21: AFFINE TRANSFORMATIONS

Given some invertible linear map  $L : \mathbb{R}^n \rightarrow \mathbb{R}^n$ . If a set  $E \subset \mathbb{R}^n$  is Jordan measurable then so is its image under  $L$ , which we denote  $LE$ . Moreover there exist a positive factor  $\lambda(L)$  such that

$$\mu(LE) = \lambda(L)\mu(E) \quad \text{for all } E \text{ Jordan measurable.}$$

*Proof of Proposition 13.21.* By the exact same argument as in Lemma 13.20. If  $E \subset \mathbb{R}^n$  is Jordan measurable the  $L(E)$  is also Jordan measurable (since  $L$  is Lipschitz).

Let  $Q_p := [0, 2^{-p}]^n$ . We know that  $L(Q_0)$  is measurable. Since  $L^{-1}$  is Lipschitz and maps  $L(Q_0)$  to  $Q_0$  (and  $\mu(Q_0) = 1 > 0$ ) the set  $L(Q_0)$  cannot be a null set. Let us put  $\lambda(L) := \mu(L(Q_0)) > 0$

Now since  $L$  is linear it follows that  $L(Q_p + a) = 2^{-p}L(Q_0) + L(a)$ . But then by the scaling and translation invariance of  $\mu$  we obtain  $\mu(Q_p + a) = 2^{-pn}\lambda(L)$ .

But then using the additivity of  $\mu$  we deduce that for all  $F$  dyadic

$$\mu(L(F)) = \lambda(L)\mu(F).$$

Finally, the validity previous formula is extended to every Jordan measurable set exactly as in the proof of Lemma 13.20.  $\square$

20

## PROPOSITION 13.22: THE FACTOR IS THE DETERMINANT

For every invertible linear map  $L : \mathbb{R}^n \rightarrow \mathbb{R}^n$  we have  $\lambda(L) = |\det L|$

The proof of Proposition 13.22 uses the following three lemmas.

## LEMMA 13.23: THE FACTOR FOR SPECIAL STRETCHINGS

Given  $\lambda_1, \dots, \lambda_n$  given positive real numbers, consider the 'special stretching' (affine transformation)  $Sx = (\lambda_1 x_1, \dots, \lambda_n x_n)$ . Then  $\lambda(S) = \prod_{i=1}^n \lambda_i$

*Proof.* It is an immediate consequence of the formula of the volume of boxes.  $\square$

## LEMMA 13.24: BALLS ARE JORDAN MEASURABLE

The unit ball  $B_1(0) \subset \mathbb{R}^n$  is Jordan measurable.

*Proof.* The boundary of the ball is covered by the union of the two graphs

$$x_n = f(x_1, \dots, x_{n-1}) \quad x_n = -f(x_1, \dots, x_{n-1})$$

where

$$f(x_1, \dots, x_{n-1}) = \sqrt{1 - (x_1^2 + \dots + x_{n-1}^2)},$$

is uniformly continuous in the closed unit ball of  $\mathbb{R}^{n-1}$ .

Hence, using Lemma 13.17 we obtain,  $\partial B_1(0)$  is a null set and thus, by Lemma 13.12 we obtain that  $B_1(0)$  is Jordan measurable.  $\square$

The next result is a standard consequence of the Spectral Theorem (11.9)

**LEMMA 13.25: POLAR DECOMPOSITION**

Let  $L : \mathbb{R}^n \rightarrow \mathbb{R}^n$  an invertible linear map, i.e., an invertible  $n \times n$  matrix with real entries acting on vectors). Then  $L$  admits the factorization

$$L = R_2 S R_1,$$

where  $R_i$  are  $n \times n$  orthogonal matrices (i.e.,  $R_i^T R_i = I_n$ ) and  $S$  is a diagonal matrix with positive entries.

*Proof.* Notice that the matrix  $A = L^T L$  is symmetric and nonnegative definite (since  $v^T A v = |Lv|^2 \geq 0$  for all  $v \in \mathbb{R}^n$ ). Thus, applying the Spectral Theorem to  $A$  we obtain  $O^T A O = D$ , where  $D = \text{diag}(\lambda_1^2, \dots, \lambda_n^2)$  is a diagonal matrix with nonnegative entries and  $O$  is orthogonal. Since  $L$  is invertible the entries of  $D$  must be all positive.

Define then  $S := \text{diag}(|\lambda_1|, \dots, |\lambda_n|)$  and notice that  $S^2 = D$ . We thus find  $O^T A O = S^2$  and thus  $S^{-1} O^T L^T L O S^{-1} = I_n$ . But since  $S^{-1} = (S^{-1})^T$  (because  $S$  and  $S^{-1}$  are diagonal) we have shown  $(L O S^{-1})^T L O S^{-1} = I_n$ . In other words  $L O S^{-1}$  is an orthogonal matrix. Hence,  $L O S^{-1} =: R_2$ , is orthogonal. Thus,  $L = R_2 S O^T$ , and the lemma follows putting  $R_1 = O^T$ .  $\square$

*Proof of Proposition 13.22.* By Lemma 13.25 we have that  $L = R_2 S R_1$ , where  $S$  is a special stretching and  $R_i$  are orthogonal matrices (i.e., rotations and symmetries). Since rotations leave the unit ball invariant we have  $\mu(B_1) = \mu(R_i B_1) = \lambda(R_i) \mu(B_1)$  and hence  $\lambda(R_i) = 1$ . On the other hand, it is an immediate consequence of the defining property of  $\lambda$  that it must be clearly multiplicative:

$$\lambda(L_2 \circ L_1) = \lambda(L_2) \lambda(L_1).$$

(Indeed for all  $E$  Jordan measurable,  $\lambda(L_2 \circ L_1) \mu(E) = \mu((L_2 \circ L_1)(E)) = \mu((L_2(L_1(E)))) = \lambda(L_2) \mu(L_1(E)) = \lambda(L_2) \lambda(L_1) \mu(E)$ .)

Hence,  $\lambda(L) = \lambda(R_2 S R_1) = \lambda(R_2) \lambda(S) \lambda(R_1) = \lambda S = |\det S|$ .

Since orthogonal matrices have absolute value of determinant equal to one, and the determinant is also multiplicative,  $|\det S| = |\det R_2 S R_1| = |\det L|$  and the lemma follows.  $\square$

**COROLLARY 13.26: ISOTROPY OF THE VOLUME**

For all  $E$  Jordan measurable and  $R$  orthogonal  $\mu(RE) = \mu(E)$ . In other words,  $\mu$  is invariant by rotations and reflections.

## 13.2 Riemann integral, change of variables formula

We next define the Riemann integral for real functions of several variables.

First we define the characteristic functions of a set in  $\mathbb{R}^n$ .

### DEFINITION 13.27: CHARACTERISTIC FUNCTION OF A SET

Given  $A \subset \mathbb{R}^n$  the **characteristic function** of  $A$ , that we will denote  $\mathbf{1}_A; \mathbb{R}^n \rightarrow \mathbb{R}$  is the function defined as:

$$\mathbf{1}_A(x) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{if } x \in \mathbb{R}^n \setminus A. \end{cases}$$

20

### DEFINITION 13.28: DYADIC STEP FUNCTIONS AND THEIR INTEGRAL

We say that a function  $g : \mathbb{R}^n \rightarrow \mathbb{R}$  is a dyadic step function if for some  $p, N \in \mathbb{N}$  and  $a : \{1, \dots, N\} \rightarrow \mathbb{Z}^n$  injective and  $b : \{1, \dots, N\} \rightarrow \mathbb{Z}$  we have

$$g(x) = \sum_{\ell=1}^N 2^{-pb(\ell)} \mathbf{1}_{Q_p(a(\ell))}(x),$$

where we use the notation  $Q_p(a) := 2^{-p}(a + [0, 1)^N)$  for dyadic cubes.

If  $g : \mathbb{R}^n \rightarrow \mathbb{R}$  is a dyadic step function we define its integral, denoted  $\int g$  as:

$$\int g = \sum_{\ell=1}^N 2^{-pb(\ell)} \mu_n(Q_p(a(\ell))) = 2^{-np} \sum_{\ell=1}^N 2^{-pb(\ell)}$$

21 **Cancelled.**



## DEFINITION 13.29: RIEMANN INTEGRABLE FUNCTIONS AND THEIR INTEGRAL

Let  $A \subset \mathbb{R}^n$  be subset and  $f : A \rightarrow \mathbb{R}$  be a function.

We say that  $f$  is **Riemann integrable** over  $A$  if lower and upper sums, defined as

$$I_{\text{low}}(f, A) := \sup \left\{ \int g \mid g : \mathbb{R}^n \rightarrow \mathbb{R} \text{ dyadic step function such that } g \leq f \mathbf{1}_A \right\}$$

and

$$I_{\text{up}}(f, A) := \inf \left\{ \int h \mid h : \mathbb{R}^n \rightarrow \mathbb{R} \text{ dyadic step function such that } h \geq f \mathbf{1}_A \right\},$$

coincide.

The common value is termed the **integral of  $f$  over  $A$** , denoted by  $\int_A f$ .

To explicitly indicate the variables of integration, we use notations like  $\int_A f(x) dx$  or  $\int_A f(y) dy$ , where  $dx = dx_1 \dots dx_n$  and  $dy = dy_1 \dots dy_n$ , respectively. To explicitly indicate that the reference measure to define the integral is the Jordan measure, denoted  $\mu$ ,  $\mu_n$ , or  $\text{vol}_n$  we will sometimes write  $\int_A f d\mu$ ,  $\int_A f d\mu_n$ , or  $\int_A f d\text{vol}_n$

22

## LEMMA 13.30: FIRST PROPERTIES OF THE RIEMANN INTEGRAL

If  $A$  is Jordan measurable and  $c : A \rightarrow \mathbb{R}$  is a constant function then it is Riemann integrable and  $\int_A c = c\mu_n(A)$ .

Also if,  $f_1, f_2 : A \rightarrow \mathbb{R}$  are Riemann integrable and  $c_1, c_2 \in \mathbb{R}$  then  $c_1 f_1 + c_2 f_2$  is Riemann integrable and:

$$\int_A (c_1 f_1 + c_2 f_2) = c_1 \int_A f_1 + c_2 \int_A f_2.$$

*Proof.* For the first part of the lemma notice that if  $E$  is Jordan measurable then given  $\epsilon > 0$  there exists dyadic sets  $G_p \subset E$  and  $H_p \supset E$  such that  $\mu(G_p) \leq \mu(E) \leq \mu(H_p) \leq \mu(G_p) + \epsilon$ .

Then the step function  $g := \lfloor c \rfloor_p \mathbf{1}_{G_p}$  and  $h := \lceil c \rceil_p \mathbf{1}_{H_p}$  satisfy  $g \leq c \mathbf{1}_E$ ,  $h \geq c \mathbf{1}_E$ , and  $\int g \leq c\mu(E) \leq \int h \leq \int g + c\epsilon + 2^{-p}$ . Sending  $\epsilon \rightarrow 0$  and  $p \rightarrow \infty$  we obtain that  $(I) = (II) = c\mu(E)$ .

For the linearity property, given that  $f_1$  and  $f_2$  are Riemann integrable over  $A$ , and  $c_1, c_2$  are constants, the function  $(c_1 f_1 + c_2 f_2) \mathbf{1}_A$  can be approximated (by below and by above) by linear combinations of dyadic step functions approximating  $f_1$  and  $f_2$ . Since by definition the integral is linear over step functions we conclude. The details are left to the reader.  $\square$

**DEFINITION 13.31: POSITIVE AND NEGATIVE PARTS OF A FUNCTION**

If  $X$  is a set and  $f : X \rightarrow \mathbb{R}$  is a function, we define the positive part  $f^+ : X \rightarrow [0, +\infty)$  and negative part  $f^- : X \rightarrow [0, +\infty)$  of  $f$  as:

$$f^+(x) := \max(f(x), 0), \quad f^-(x) := \max(-f(x), 0) \quad x \in X.$$

Notice that we can write  $f = f^+ - f^-$ .

**LEMMA 13.32: ALL BOILS DOWN TO INTEGRATING NONNEGATIVE FUNCTIONS**

Given  $A \subset \mathbb{R}^n$ , a function  $f : A \rightarrow \mathbb{R}$  is Riemann integrable (over  $A$ ) if and only if  $f^+$  and  $f^-$  are. Moreover:

$$\int_A f = \int_A f^+ - \int_A f^-$$

*Proof.* Left as an exercise. □

**LEMMA 13.33: RIEMANN INTEGRAL AS JORDAN MEASURE**

Let  $A \subset \mathbb{R}^n$  be a set. A function  $f : A \rightarrow [0, \infty)$  is Riemann integrable if and only if the hypograph

$$\Gamma_A(f) := \{(x, y) \in \mathbb{R}^n \times \mathbb{R} \mid x \in A, \quad 0 \leq y < f(x)\}$$

is Jordan measurable in  $\mathbb{R}^{n+1}$ .

In such case:

$$\int_A f = \mu_{n+1}(\Gamma_A(f)).$$

*Proof.* The lemma follows easily from the following observation: Given any dyadic step functions

$$g(x) = \sum_{\ell=1}^N 2^{-p} b(\ell) \mathbf{1}_{Q_p(a(\ell))}(x),$$

such that  $g \leq f$ ; then the dyadic set

$$G = \bigcup \{2^{-p}((a(\ell), k) + [0, 1)^{n+1}) \mid 1 \leq \ell \leq N, \quad 0 \leq k < b(\ell)\}$$

is contained in  $\Gamma(f)$  and satisfies  $\mu_{n+1}(G) = \int g$ .

Similarly, given any dyadic step function  $h$  such that  $f \leq h$  we have an associated dyadic set  $H$  containing  $\Gamma_A(f)$  and such that  $\mu_{n+1}(H) = \int h$ . □

**COROLLARY 13.34: INTEGRABILITY OF UNIFORMLY CONTINUOUS FUNCTIONS**

Suppose that  $E \subset \mathbb{R}^n$  is a Jordan measurable set and  $f : \overline{E} \rightarrow [0, \infty)$  is a continuous function. Then,  $f$  is Riemann integrable over  $E$ .

*Proof.* Since  $E$  Jordan measurable it is bounded. Hence its closure  $\overline{E}$  is compact. Therefore  $f$  is a continuous function on a compact set, hence it is bounded and uniformly continuous. Let  $C := \max_{x \in \overline{E}} f(x)$ .

Lemma 13.17 implies that

$$A_1 := \{(x, y) \in \mathbb{R}^n \times \mathbb{R} \mid x \in \overline{E}, y = f(x)\}$$

is Jordan null in  $\mathbb{R}^{n+1}$ . And the set

$$A_2 := \{(x, 0) \in \mathbb{R}^n \times \mathbb{R} \mid x \in \overline{E}\}$$

is also Jordan null in  $\mathbb{R}^{n+1}$ .

Finally, using that  $\partial E$  is Jordan null in  $\mathbb{R}^n$  is not difficult to show that the set

$$A_3 := \{(x, y) \in \mathbb{R}^n \times \mathbb{R} \mid x \in \partial E, 0 \leq y \leq C\}$$

is also Jordan null in  $\mathbb{R}^{n+1}$  (left as an exercise).

Now since,

$$\partial \Gamma_E(f) \subset A_1 \cup A_2 \cup A_3$$

we conclude that  $\partial \Gamma(f)$  is Jordan null and hence  $\Gamma(f)$  is Jordan measurable or, equivalently,  $f$  is Riemann integrable.  $\square$

22

We now give the change of variables formula

**THEOREM 13.35: CHANGE OF VARIABLES FORMULA**

Suppose that  $U, V \subset \mathbb{R}^n$  are open sets and  $\Phi : U \rightarrow V$  be a  $C^1$  diffeomorphism. Then, if  $A \subset U$  is a Jordan measurable set with  $\overline{A} \subset U$  and  $f : \overline{A} \rightarrow [0, +\infty)$  is a continuous function then

$$\int_A f(x) dx = \int_{\Phi(A)} \frac{f(\Phi^{-1}(y))}{|\det J\Phi(\Phi^{-1}(y))|} dy. \quad (13.4)$$

where  $|\det J\Phi(x)|$  is the absolute value of the determinant of the Jacobi matrix of  $\Phi$  at  $x \in U$ .

*Proof.* Since  $\Phi \in C^1(U, \mathbb{R}^n)$  we have that  $\Phi$  is locally Lipschitz continuous.

Let us show that  $\Phi(\overline{A}) = \overline{\Phi(A)}$  is Jordan measurable.

Indeed, since  $A$  is Jordan measurable so is the compact set  $\overline{A}$  (they have the same boundary). Now, since  $\Phi$  is locally Lipschitz continuous for every  $x \in \overline{A}$  there is  $r_x > 0$  such that  $f|_{B_{r_x}(x)}$  is Lipschitz continuous. But since the collection of open balls  $\{B_{r_x}(x) \mid x \in \overline{A}\}$  is an open cover of  $A$  we can extract a finite subcover  $B_{r_1}(x_1), \dots, B_{r_N}(x_N)$ .

Now since  $\mu(\partial A) = 0$  we have  $\mu(\partial A \cap B_{r_i}(x_i)) = 0$  for all  $i = 1, \dots, N$  and hence  $\Phi(\partial A \cap B_{r_i}(x_i)) = 0$  being the Lipschitz image of a Jordan null set.

But then since,  $\partial\Phi(A) = \Phi(\partial A) = \bigcup_{i=1}^N \Phi(\partial A \cap B_{r_i}(x_i))$  we obtain that  $\partial\Phi(A)$  is Jordan null and hence  $\Phi(A)$  is Jordan measurable.

Therefore by Corollary 13.34 the integrals at the two sides of (13.4) are well-defined Riemann integrals.

For  $p \in \mathbb{N}$  and  $\delta \in (0, 1/2)$  define

$$Q_{p,\delta} := 2^{-p}[\delta, 1 - \delta]^n.$$

Now, a key part of the proof is to show that for any given  $\delta > 0$  (small) there exists  $p_\delta \in \mathbb{N}$  such that for every  $p \geq p_\delta$  and  $x_o \in \bar{A}$  we have

$$(1 - 2\delta)^n 2^{-pn} |\det J\Phi(x_o)| = \mu(D\Phi_{x_o}(Q_{p,\delta})) \leq \mu(\Phi(x_o + [0, 2^{-p}]^n)) \quad (13.5)$$

and

$$\sup \{ |\det J\Phi(x)| \mid x \in x_o + [0, 2^{-p}]^n \} \leq (1 + \delta) |\det J\Phi(x_o)|. \quad (13.6)$$

We notice that the equality in (13.5) follows from (13.22). To establish the inequality we will show that

$$(\Phi(x_o) + D\Phi_{x_o}(Q_{p,\delta})) \subset \Phi(x_o + [0, 2^{-p}]^n)$$

Similarly to the proof of the Inverse Function Theorem (IFT), for every  $\lambda > 0$  (small) there exists  $r_\lambda > 0$  such that for all  $x_o \in \bar{A}$  that the map

$$\Psi_{x_o}(h) := \Phi^{-1}(\Phi(x_o) + D\Phi_{x_o}(h)) - x_o$$

satisfies that  $\Psi_{x_o} - \text{Id}$  is  $\lambda$ -Lipchitz for  $h \in B_{r_\lambda}(0)$ . Indeed, by the chain rule we have

$$D(\Psi_{x_o} - \text{Id})_0 = 0$$

and hence, reasoning exactly as in the proof of IFT, given  $\lambda > 0$  there is  $r_\lambda > 0$  such that

$$\Psi_{x_o} - \text{Id} \quad \text{is } \lambda\text{-Lipschitz for } h \in B_{r_\lambda}(0).$$

The fact that  $r_\lambda > 0$  can be chosen independent of  $x_o$  is a consequence of the compactness of  $\bar{A}$  —one can use similar argument to the proof of Proposition 9.78 (‘Continuous function in compact is uniformly continuous’).

Since  $\Psi_{x_o}(0) = 0$  for  $h \in Q_{p,\delta} \subset B_{r_o}$  we have

$$\|\Psi_{x_o}(h) - h\| \leq \lambda|h| \leq \lambda\sqrt{n}2^{-p}$$

Therefore, given  $\delta > 0$  we can choose  $\lambda$  such that  $\lambda\sqrt{n} < \delta$  and  $p_\delta$  such that  $2^{-p_\delta} \leq r_\lambda$  so that

$$\Psi_{x_o}(Q_{p,\delta}) \subset [0, 2^{-p}]^n \quad \text{for all } p \geq p_\delta.$$

That is:

$$\Phi^{-1}(\Phi(x_o) + D\Phi_{x_o}(Q_{p,\delta})) \subset [0, 2^{-p}]^n + x_o \text{ for all } p \geq p_\delta.$$

Applying  $\Phi$  both sides we conclude (13.5).

On the other hand (13.6) follows from the fact that  $|\det J\Phi(x)|$  is positive (since  $\Phi$  is a diffeomorphism) and uniformly continuous on the compact set  $\bar{A}$ . Indeed, we have

$$c := \inf_{x \in \bar{A}} |\det J\Phi(x)| > 0$$

and then by uniform continuity for every  $\delta > 0$  exist  $r_\delta > 0$  such that for all  $x_o \in \bar{A}$  we have

$$||\det J\Phi(x)| - |\det J\Phi(x_o)|| < \delta c \quad \text{for } x \in B_{r_\delta}(x_o) \cap \bar{A}$$

and hence

$$|\det J\Phi(x)| \leq |\det J\Phi(x_o)| + \delta c \leq (1 + \delta)|\det J\Phi(x_o)| \quad \text{for } x \in B_{r_\delta}(x_o) \cap \bar{A}.$$

Suppose now that  $g : \mathbb{R}^n \rightarrow \mathbb{R}$  is a dyadic step function approximating  $f\mathbf{1}_A$  by below. More precisely, given  $\epsilon > 0$  let  $g = \sum_{\ell=1}^N g_\ell \mathbf{1}_{Q_\ell}$  where  $Q_\ell = 2^{-p}(a(\ell) + [0, 1]^n)$ ,  $a : \{1, \dots, N\} \rightarrow \mathbb{Z}^n$  injective, satisfies  $g \leq f\mathbf{1}_A$  in all of  $\mathbb{R}^n$  and  $\int_A f - \epsilon \leq \int g$ . Put  $x_\ell := 2^{-p}a(\ell)$ .

Finally, combining (13.5)-blabla2 we obtain:

$$\sup \{ |\det J\Phi(x)| \mid x \in x_o + [0, 2^{-p}]^n \} 2^{-pn} \leq \frac{1 + \delta}{(1 - 2\delta)^n} \mu(\Phi(x_o + [0, 2^{-p}]^n))$$

for all  $x_o \in \bar{A}$ , provided  $p \geq p_\delta$ .

Therefore, observing

$$g_\ell \leq \inf_{x \in Q_\ell} f(x) = \inf_{y \in \Phi(Q_\ell)} f(\Phi^{-1}(y))$$

we obtain:

$$\begin{aligned} \int_A f - \epsilon &\leq \int g = \sum_{\ell=1}^N g_\ell 2^{-np} \\ &\leq \frac{1 + \delta}{(1 - 2\delta)^n} \sum_{\ell=1}^N \inf_{y \in \Phi(Q_\ell)} \frac{f(\Phi^{-1}(y))}{|\det J\Phi(\Phi^{-1}(y))|} \mu(\Phi(Q_\ell)) \\ &\leq \frac{1 + \delta}{(1 - 2\delta)^n} \int_{\Phi(A)} \frac{f(\Phi^{-1}(y))}{|\det J\Phi(\Phi^{-1}(y))|} dy \end{aligned}$$

Since  $\epsilon$  and  $\delta$  can be made arbitrarily small (by taking  $p$  large enough), we obtain:

$$\int_A f \leq \int_{\Phi(A)} \frac{f(\Phi^{-1}(y))}{|\det J\Phi(\Phi^{-1}(y))|} dy.$$

This inequality also holds with  $A$ ,  $f$ , and  $\Phi$  replaced by  $\tilde{A} = \Phi(A)$ ,  $\tilde{f}(y) = \frac{f(\Phi^{-1}(y))}{|\det J\Phi(\Phi^{-1}(y))|}$ , and  $\tilde{\Phi} = \Phi^{-1}$ , yielding:

$$\int_{\tilde{A}} \tilde{f} \leq \int_{\tilde{\Phi}(\tilde{A})} \frac{\tilde{f}(\tilde{\Phi}^{-1}(x))}{|\det J\tilde{\Phi}(\tilde{\Phi}^{-1}(x))|} dx. \quad (13.7)$$

However, considering the definitions of  $\tilde{f}$  and  $\tilde{\Phi}$ , we perform the computation:

$$\frac{\tilde{f}(\tilde{\Phi}^{-1}(x))}{|\det J\tilde{\Phi}(\tilde{\Phi}^{-1}(x))|} = \frac{f(\Phi^{-1}(\tilde{\Phi}^{-1}(x)))}{|\det J\Phi(\Phi^{-1}(\tilde{\Phi}^{-1}(x)))|} = \frac{f(x)}{|\det J\Phi(x)|} = f(x)$$

because  $\tilde{\Phi}^{-1}(x) = \Phi(x)$  and  $\det J\Phi(x) \det J(\Phi^{-1})(\Phi(x)) = 1$  (since  $(J\Phi)^{-1}(x) = J(\Phi^{-1})(\Phi(x))$ ).

Therefore, the (13.7) can be rewritten as:

$$\int_{\tilde{A}} \tilde{f} = \int_{\Phi(A)} \frac{f(\Phi^{-1}(y))}{|\det J\Phi(\Phi^{-1}(y))|} dy \leq \int_{\tilde{\Phi}(\tilde{A})} \frac{\tilde{f}(\tilde{\Phi}^{-1}(x))}{|\det J\tilde{\Phi}(\tilde{\Phi}^{-1}(x))|} dx = \int_A f(x) dx.$$

This establishes the equality (13.4) and hence the theorem.  $\square$

EXERCISE 13.36. — Prove the following equivalent version of the change of variables formula: If  $\Psi : V \rightarrow U$  is a diffeomorphism,  $f : U \rightarrow \mathbb{R}$  a continuous function and  $\bar{A} \subset U$  a compact Jordan measurable subset, then:

$$\int_A f(x) dx = \int_{\Psi(A)} f(\Psi(y)) |\det J\Psi(y)| dy.$$

### 13.3 Fubini Theorem, differentiation under the integral sign

We next give the Fubini Theorem

## THEOREM 13.37: SLICING FORMULA (OR CAVALIERI'S PRINCIPLE)

Suppose that  $E \subset \mathbb{R}^{n-1} \times (-C, C) \subset \mathbb{R}^n$  is a Jordan measurable set. For  $y \in \mathbb{R}$  let the 'slice'  $S_y(E) \subset \mathbb{R}^{n-1}$  be defined as:

$$S_y(E) := \{x \in \mathbb{R}^{n-1} \mid (x, y) \in E\}.$$

Then functions

$$\varphi(y) := \mu_{n,\text{in}}(S_y(E)) \quad \text{and} \quad \psi(y) := \mu_{n,\text{out}}(S_y(E)),$$

which satisfy  $0 \leq \varphi \leq \psi$  are both Riemann integrable in  $\mathbb{R}$  and

$$\int_{-C}^C \varphi(y) dy = \int_{-C}^C \psi(y) dy = \mu_{n+1}(E).$$

*Proof.* Since  $E$  is Jordan measurable, for any  $\epsilon > 0$  there  $p \in \mathbb{N}$  and  $G_p, H_p$  dyadic with pixel size  $2^{-p}$  such that  $G_p \subset E \subset H_p$  and  $\mu_n(G_p) \leq \mu_n(E) \leq \mu_n(H_p) < \mu_n(G_p) + \epsilon$ .

Notice that for all  $y \in \mathbb{R}$  the sets  $S_y(G)$  and  $S_y(H)$  are dyadic (also with pixel size  $2^{-p}$ ) and satisfy  $S_y(G) \subset S_y(E) \subset S_y(H)$ .

Notice also that the sets  $S_y(G)$  and  $S_y(H)$  remain constant for  $y$  within intervals of the form  $y \in 2^{-p}[k, k+1)$ , for all  $k \in \mathbb{Z}$ .

Hence the functions

$$g_p(y) := \mu_{n-1}(S_y(G)) \quad \text{and} \quad h_p(y) := \mu_{n-1}(S_y(H))$$

are dyadic step functions in  $\mathbb{R}$ . Notice also that

$$\int_{\mathbb{R}} g_p = \mu_n(G_p) \quad \text{and} \quad \int_{\mathbb{R}} h_p = \mu_n(H_p).$$

Indeed, as stated above, within each interval  $y \in 2^{-p}[k, k+1)$ , the function  $g_p(y)$  remains constant. It is calculated as  $2^{-p(n-1)}$  times the number of  $n$ -dimensional cubes in  $G_p$  that intersect the hyperplane  $\{x \in \mathbb{R}^n \mid x_n = t\}$ . As  $t$  varies, this hyperplane effectively sweeps through all the cubes within  $G_p$ . Within this setup, the identity  $\mu_n(G_p) = \int_{\mathbb{R}} g_p$  amounts to an elementary discrete counting argument: that the total number of cubes in  $G_p$  is equal to the sum, across all vertical levels, of the number of cubes at each level. A similar argument applies to  $h_p$ .

Since  $S_y(G) \subset S_y(E) \subset S_y(H)$  we have

$$g_p(y) \leq \varphi(y) = \mu_{n,\text{in}}(S_y(E)) \leq \mu_{n,\text{out}}(S_y(E)) = \psi(y) \leq h_p(y)$$

for all  $y \in \mathbb{R}$  and therefore:

$$\int g_p \leq I_{\text{low}}(\varphi, (-C, C)) \leq I_{\text{up}}(\varphi, (-C, C)) \leq \int h_p = \int g_p + \epsilon$$

and

$$\int g_p \leq I_{\text{low}}(\psi, (-C, C)) \leq I_{\text{up}}(\psi, (-C, C)) \leq \int h_p \leq \int g_p + \epsilon.$$

Since  $\epsilon > 0$  can be made arbitrarily small by taking  $p \rightarrow \infty$  (and using that  $\int g_p = \mu_{n+1}(G_p)$  converges to  $\mu_{n+1}(E)$ ) it follows that both  $\varphi$  and  $\psi$  are Riemann integrable with

$$\int_{-C}^C \varphi = \int_{-C}^C \psi = \mu_{n+1}(E)$$

as claimed.  $\square$

With a similar proof we have

**THEOREM 13.38: SLICING FORMULA FOR FUNCTIONS (FUBINI)**

Suppose  $A \subset \mathbb{R}^n$  is a bounded subset, i.e.  $A \subset (-C, C)$  for some  $C$ , and that  $f : A \rightarrow \mathbb{R}$  is Riemann integrable function. Let  $\tilde{f} : (-C, C)^n \rightarrow \mathbb{R}$  be defined as  $\tilde{f} = f \mathbf{1}_A$ . For a given  $i \in \{1, 2, \dots, n\}$  and  $y \in \mathbb{R}$ , define the ‘slice’  $f_{i,y} : \mathbb{R}^{n-1} \rightarrow \mathbb{R}$  as follows:

$$f_{i,y}(x_1, \dots, x_{n-1}) := \tilde{f}(x_1, \dots, x_{i-1}, y, x_i, \dots, x_{n-1}),$$

where  $y$  is inserted in the  $i$ -th position of the function argument, replacing the  $i$ -th variable. Then functions

$$\varphi(y) := I_{\text{low}}(f_{i,y}, (-C, C)^{n-1}) \quad \text{and} \quad \psi(y) := I_{\text{up}}(f_{i,y}, (-C, C)^{n-1}),$$

which satisfy  $\varphi \leq \psi$ , are both Riemann integrable (in  $\mathbb{R}$ ) and

$$\int_{-C}^C \varphi(y) dy = \int_{-C}^C \psi(y) dy = \int_A f.$$

*Proof.* Is a variation of the proof of Theorem 13.37. Left to the interested reader.  $\square$



## THEOREM 13.39: FUBINI IN BOXES

Suppose that  $f : K \rightarrow [0, \infty)$  is a continuous function in the compact box

$$K = [a_1, b_1] \times [a_2, b_2] \times \cdots \times [a_n, b_n].$$

Then,

$$\begin{aligned} \int_K f(x) dx &= \int_{a_n}^{b_n} \left( \int_{a_{n-1}}^{b_{n-1}} \left( \cdots \int_{a_2}^{b_2} \left( \int_{a_1}^{b_1} f(x_1, x_2, \dots, x_n) dx_1 \right) dx_2 \cdots dx_{n-1} \right) dx_n \right) \\ &= \int_{a_n}^{b_n} \int_{a_{n-1}}^{b_{n-1}} \cdots \int_{a_2}^{b_2} \int_{a_1}^{b_1} f(x_1, x_2, \dots, x_n) dx_1 dx_2 \cdots dx_{n-1} dx_n. \end{aligned}$$

Moreover, we can integrate with respect to the different variables in any order: For any permutation  $\sigma$  of  $\{1, 2, \dots, n\}$ , we have:

$$\int_{a_{\sigma(n)}}^{b_{\sigma(n)}} \cdots \int_{a_{\sigma(2)}}^{b_{\sigma(2)}} \int_{a_{\sigma(1)}}^{b_{\sigma(1)}} f(x_1, x_2, \dots, x_n) dx_{\sigma(1)} dx_{\sigma(2)} \cdots dx_{\sigma(n)}.$$

23

*Proof.* It follows applying  $n$  times the slicing formula for functions, each time in one dimension less.  $\square$

13.40. — Observe that in the Slicing formula we have used the variable  $x_{n+1}$  to slice just for convenience, but we could have sliced in any other direction  $x_i$ . Indeed, if  $P$  is a permutation of the coordinates then we can apply our slicing results to  $\tilde{f} = f \circ P$  after noticing that  $\int_{P^{-1}(A)} f \circ P = \int_A P$ .

As a consequence in Fubini's Theorem the order in which we perform the integrations is irrelevant. For example if  $f = f(x_1, x_2, x_3)$  is a continuous function in  $[a_1, b_1] \times [a_2, b_2] \times [a_3, b_3]$  then

$$\begin{aligned} \int_{a_3}^{b_3} \int_{a_2}^{b_2} \int_{a_1}^{b_1} f(x_1, x_2, x_3) dx_1 dx_2 dx_3 &= \int_{a_2}^{b_2} \int_{a_1}^{b_1} \int_{a_3}^{b_3} f(x_1, x_2, x_3) dx_3 dx_1 dx_2 \\ &= \cdots = \int_{a_1}^{b_1} \int_{a_2}^{b_2} \int_{a_3}^{b_3} f(x_1, x_2, x_3) dx_3 dx_2 dx_1 \end{aligned}$$

24

Another very useful result is the following

**THEOREM 13.41: DIFFERENTIATION UNDER THE INTEGRAL SIGN**

Suppose that  $U \subset \mathbb{R}^n \times \mathbb{R}$  is open and that  $f \in C^1(U)$ . Let  $K \subset \mathbb{R}^n$  be compact and Jordan measurable and suppose that  $K \times [-a, b] \subset U$  for some  $a < b$ . Then for  $y \in (a, b)$  the function

$$y \mapsto \int_K f(x, y) dx$$

is continuously differentiable in  $(a, b)$  and

$$\frac{d}{dy} \int_K f(x, y) dx = \int_K \frac{\partial}{\partial y} f(x, y) dx.$$

*Proof.* Fix  $y \in (a, b)$  and let  $h > 0$  be small enough so that  $y + h \in (a, b)$ . The function  $\frac{\partial}{\partial y} f(x, y)$  is continuous and hence uniformly continuous in the compact  $K \times [a, b]$ .

On the other hand, by the Intermediate Value Theorem we have  $f(x, y + h) - f(x, y) = \frac{\partial}{\partial y} f(x, \xi_{x,y,h})$  for some  $\xi_{x,y,h}$  in the interval joining  $y$  and  $y + h$ .

Thus, for any given  $\epsilon > 0$  exists  $\delta > 0$  such that if  $0 < |h| < \delta$  and  $x \in K$  we have,

$$\left| \frac{f(x, y + h) - f(x, y)}{h} - \frac{\partial}{\partial y} f(x, y) \right| = \left| \frac{\partial}{\partial y} f(x, \xi_{x,y,h}) - \frac{\partial}{\partial y} f(x, y) \right| < \epsilon.$$

Then, using the linearity of the integral

$$\frac{1}{h} \left( \int_K f(x, y + h) dx - \int_K f(x, y) dx \right) = \int_K \frac{f(x, y + h) - f(x, y)}{h} dx \quad (13.8)$$

but

$$\begin{aligned} \left| \int_K \frac{f(x, y + h) - f(x, y)}{h} dx - \int_K \frac{\partial}{\partial y} f(x, y) dx \right| &\leq \\ &\leq \int_K \left| \frac{f(x, y + h) - f(x, y)}{h} - \frac{\partial}{\partial y} f(x, y) \right| dx \leq \epsilon \mu_n(K) \end{aligned} \quad (13.9)$$

Sending  $h$  (and also hence  $\delta, \epsilon$ ) to zero we find

$$\lim_{h \rightarrow 0} \int_K \frac{f(x, y + h) - f(x, y)}{h} dx = \int_K \frac{\partial}{\partial y} f(x, y) dx$$

But then in view of (13.8) we obtain that  $\frac{d}{dy} \int_K f(x, y)$  exists and equals  $\int_K \frac{\partial}{\partial y} f(x, y) dx$ . That the function  $y \mapsto \int_K \frac{\partial}{\partial y} f(x, y) dx$  is continuous follows from (13.9).  $\square$

**EXERCISE 13.42.** — Show that Theorem 13.41 still holds (with the same proof) if we replace the assumption  $f \in C^1(U)$  by the assumption that  $\partial_y f$  exists and is continuous in  $U$ .

## 13.4 Examples: Change of variables and Fubini in practice

### Integrals polar coordinates in $\mathbb{R}^2$

Cartesian coordinates  $(x, y)$  and polar coordinates  $(r, \theta)$  in  $\mathbb{R}^2$  are related by the transformation:

$$x = r \cos \theta, \quad y = r \sin \theta, \quad \theta \in [0, 2\pi), \quad r \geq 0,$$

in other words

$$(x, y) = \Psi(r, \theta) := (r \cos \theta, r \sin \theta).$$

The Jacobian matrix of this transformation is

$$J_{\text{polar}} = J\Psi(r, \theta) = \begin{bmatrix} \cos \theta & -r \sin \theta \\ \sin \theta & r \cos \theta \end{bmatrix},$$

and the determinant of this Jacobian matrix is:

$$\det(J_{\text{polar}}) = r$$

EXAMPLE 13.43. — To compute the integral of the function  $(x, y) \mapsto e^{-x^2-y^2}$  in the ball of radius one  $B_1 \subset \mathbb{R}^2$  we notice that that the set  $\Psi$  is a diffeomorphism when restricted to  $V := (0, 2\pi) \times (0, 1)$  and that  $U := \Psi((0, 2\pi) \times (0, 1))$  is such that  $B_1 \setminus U$  is a null set.

Hence the change of variables formula gives

$$\begin{aligned} \int_{B_1} e^{-x^2-y^2} dx dy &= \int_U e^{-x^2-y^2} dx dy = \int_V e^{-(r \cos \theta)^2 - (r \sin \theta)^2} r dr d\theta \\ &= \int_0^1 \int_0^{2\pi} e^{-r^2} r dr = \pi \int_0^1 e^{-r^2} 2r dr = \pi(1 - e^{-1}). \end{aligned}$$

### Integrals in Spherical Coordinates in $\mathbb{R}^3$

Cartesian coordinates  $(x, y, z)$  and spherical coordinates  $(r, \phi, \theta)$  in  $\mathbb{R}^3$  are related by the transformation:

$$x = r \sin \phi \cos \theta, \quad y = r \sin \phi \sin \theta, \quad z = r \cos \phi.$$

The Jacobian matrix of this transformation is:

$$J_{\text{spherical}} = \begin{bmatrix} \sin \phi \cos \theta & r \cos \phi \cos \theta & -r \sin \phi \sin \theta \\ \sin \phi \sin \theta & r \cos \phi \sin \theta & r \sin \phi \cos \theta \\ \cos \phi & -r \sin \phi & 0 \end{bmatrix},$$

and the determinant of this Jacobian matrix is:

$$\det(J_{\text{spherical}}) = r^2 \sin \phi.$$

24

EXAMPLE 13.44. — To compute the volume of a sphere of radius  $R$ , we integrate in spherical coordinates:

$$\text{Volume} = \int_0^{2\pi} \int_0^\pi \int_0^R \rho^2 \sin \phi \, d\rho \, d\phi \, d\theta = \frac{4}{3}\pi R^3.$$

### Cylindrical Coordinates in $\mathbb{R}^3$

Cylindrical coordinates are a generalization of polar coordinates to three dimensions where a point in space is represented by  $(r, \theta, z)$ . Here,  $r$  and  $\theta$  have the same interpretation as in polar coordinates, representing the radial distance from the origin and the angle from a reference direction in the plane, respectively, while  $z$  represents the height above the plane.

The transformation from Cartesian coordinates  $(x, y, z)$  to cylindrical coordinates is given by:

$$x = r \cos \theta, \quad y = r \sin \theta, \quad z = z.$$

The Jacobian matrix for this transformation is:

$$J_{\text{cylindrical}} = \begin{bmatrix} \cos \theta & -r \sin \theta & 0 \\ \sin \theta & r \cos \theta & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

and the determinant of this Jacobian matrix is:

$$\det(J_{\text{cylindrical}}) = r.$$

In cylindrical coordinates, the volume element is  $r \, dr \, d\theta \, dz$ , which is used for integrating functions over a volume in  $\mathbb{R}^3$ .

EXAMPLE 13.45. — To compute the volume of a right circular cone in cylindrical coordinates, consider a cone with height  $h$  and base radius  $R$ . The cone is parametrized as follows:

$$0 \leq r \leq R, \quad 0 \leq \theta < 2\pi, \quad 0 \leq z \leq \frac{h}{R}r.$$

The volume element in cylindrical coordinates is  $r \, dr \, d\theta \, dz$ , so the volume  $V$  of the cone can be computed by the integral:

$$\text{Volume} = \int_0^{2\pi} \int_0^R \int_0^{\frac{h}{R}r} r \, dz \, dr \, d\theta.$$

First, integrate with respect to  $z$ :

$$\text{Volume} = \int_0^{2\pi} \int_0^R r \left[ \frac{h}{R}r \right] \, dr \, d\theta = \frac{h}{R} \int_0^{2\pi} \int_0^R r^2 \, dr \, d\theta.$$

Next, integrate with respect to  $r$ :

$$\text{Volume} = \frac{h}{R} \int_0^{2\pi} \left[ \frac{1}{3} r^3 \right]_0^R d\theta = \frac{h}{R} \int_0^{2\pi} \frac{1}{3} R^3 d\theta = \frac{h}{R} \cdot \frac{1}{3} R^3 \cdot 2\pi.$$

which is the classical formula for the volume of a cone.

## Volume Computation of a Torus

Consider a torus with a major radius  $R$  and a minor radius  $r$ . To compute the volume, we parametrize the torus and calculate the Jacobian determinant for the transformation from toroidal to Cartesian coordinates.

The torus can be parametrized by:

$$\begin{aligned} x &= (R + \rho \cos \theta) \cos \varphi, \\ y &= (R + \rho \cos \theta) \sin \varphi, \\ z &= \rho \sin \theta, \end{aligned}$$

where  $0 \leq \rho \leq r$ ,  $0 \leq \theta < 2\pi$ , and  $0 \leq \varphi < 2\pi$ .

To transform from toroidal coordinates  $(\rho, \theta, \varphi)$  to Cartesian coordinates  $(x, y, z)$ , we compute the Jacobian determinant. The partial derivatives of  $x, y, z$  with respect to  $\rho, \theta, \varphi$  form the Jacobian matrix:

$$J = \begin{bmatrix} \frac{\partial x}{\partial \rho} & \frac{\partial x}{\partial \theta} & \frac{\partial x}{\partial \varphi} \\ \frac{\partial y}{\partial \rho} & \frac{\partial y}{\partial \theta} & \frac{\partial y}{\partial \varphi} \\ \frac{\partial z}{\partial \rho} & \frac{\partial z}{\partial \theta} & \frac{\partial z}{\partial \varphi} \end{bmatrix} = \begin{bmatrix} \cos \theta \cos \varphi & -\rho \sin \theta \cos \varphi & -(R + \rho \cos \theta) \sin \varphi \\ \cos \theta \sin \varphi & -\rho \sin \theta \sin \varphi & (R + \rho \cos \theta) \cos \varphi \\ \sin \theta & \rho \cos \theta & 0 \end{bmatrix}.$$

The Jacobian determinant  $|\det J|$  is then computed as:

$$\begin{aligned} |\det J| &= \rho ((R + \rho \cos \theta)(\cos^2 \theta + \sin^2 \theta)) \\ &= \rho(R + \rho \cos \theta). \end{aligned}$$

Using this Jacobian, the volume of the torus is given by integrating over the parametric domain:

$$\text{Volume} = \int_0^{2\pi} \int_0^{2\pi} \int_0^r \rho(R + \rho \cos \theta) d\rho d\theta d\varphi.$$

First, we solve the integral with respect to  $\rho$ :

$$\int_0^r \rho(R + \rho \cos \theta) d\rho = \int_0^r (R\rho + \rho^2 \cos \theta) d\rho.$$

This evaluates to

$$\left[ \frac{1}{2} R\rho^2 + \frac{1}{3} \rho^3 \cos \theta \right]_0^r = \frac{1}{2} Rr^2 + \frac{1}{3} r^3 \cos \theta.$$

Now, integrate with respect to  $\theta$ :

$$\int_0^{2\pi} \left( \frac{1}{2} Rr^2 + \frac{1}{3} r^3 \cos \theta \right) d\theta = \pi Rr^2,$$

because the integral of  $\cos \theta$  over one period is 0.

Finally, integrate with respect to  $\varphi$ :

$$\int_0^{2\pi} \pi Rr^2 d\varphi = 2\pi^2 Rr^2.$$

Thus, the volume of the torus is  $2\pi^2 Rr^2$ .

## 13.5 Improper integrals: definition and examples

### DEFINITION 13.46:

Suppose that  $U \subset \mathbb{R}^n$  is open and  $f \in C(U)$  is nonnegative. We define the **improper integral** of  $f$  over  $U$  as  $\int_U f$  as

$$\sup \left\{ \int_K f \mid K \text{ compact, Jordan measurable, and contained in } U \right\}.$$

Notice that  $\int_U f$  is always well-defined, although it can be  $+\infty$ .

### LEMMA 13.47: IMPROPER INTEGRALS AS LIMITS

Suppose that  $U \subset \mathbb{R}^n$  is open and  $f \in C(U)$  is nonnegative.

Assume that  $K_\ell$ ,  $\ell \geq 0$  is a increasing sequence of nested compact Jordan measurable sets with

$$K_1 \subset K_2 \subset K_3 \subset \dots$$

such that

$$\bigcup_{\ell=0}^{\infty} \text{int}(K_\ell) = U.$$

Then

$$\int_U f = \lim_{\ell \rightarrow \infty} \int_{K_\ell} f.$$

(In particular the limit exists.)

*Proof.* Notice that since

$$K_1 \subset K_2 \subset K_3 \subset \dots$$

and  $f \geq 0$  we have  $\int_{K_\ell} f \leq \int_{K_{\ell+1}} f$  for all  $\ell \geq 0$ . Hence the sequence of integrals is monotone and thus has a limit.

Nowe, for any  $K \subset U$  compact and Riemann integrable, since  $K \subset \bigcup_{\ell=0}^{\infty} \text{int}(K_{\ell})$  is an open cover of  $K$  (and the sets are nested) there exists  $K_{\ell_0}$  such that  $K \subset K_{\ell_0}$ .

But then

$$\int_K f \leq \int_{K_{\ell_0}} f \leq \lim_{k \rightarrow \infty} \int_{K_{\ell}} f$$

Taking the supremum over  $K \subset U$  compact and Jordan measurable we obtain

$$\int_U f \leq \lim_{k \rightarrow \infty} \int_{K_{\ell}} f.$$

The opposite inequality is trivial as each  $K_{\ell}$  is an admissible set in the supremum defining the improper integral.  $\square$

24

EXAMPLE 13.48. — Let us compute the integral of the  $I := \int_{-\infty}^{\infty} e^{-x^2} dx = \lim_{A \rightarrow +\infty} \int_{-A}^A e^{-x^2} dx$ .

To compute it, we can use the following classical trick of doubling the variables:

$$\left( \int_{-A}^A e^{-x^2} dx \right)^2 = \int_{-A}^A e^{-x^2} dx \int_{-A}^A e^{-y^2} dy = \int_{[-A,A]^2} e^{-x^2-y^2} dx dy,$$

and hence, sending  $A \rightarrow +\infty$ ,

$$\begin{aligned} I^2 &= \int_{\mathbb{R}^2} e^{-x^2-y^2} dx dy = \lim_{R \rightarrow \infty} \int_{B_R} e^{-x^2-y^2} dx dy = \lim_{R \rightarrow \infty} \int_0^R \int_0^{2\pi} e^{-r^2} r d\theta dr \\ &= \pi \lim_{R \rightarrow \infty} \int_0^R e^{-r^2} 2r dr = \pi. \end{aligned}$$

## 13.6 Length, Area, and integrals over submanifolds

The following definition is standard, convenient, and will be used from now on.

### DEFINITION 13.49: $C^k$ FUNCTIONS ON NON-OPEN DOMAINS

If  $A \subset \mathbb{R}^n$  is not open we say that  $f \in C^k(A, \mathbb{R}^m)$  if there exist  $U \subset \mathbb{R}^n$  open and  $\tilde{f} \in C^k(U, \mathbb{R}^m)$  such that  $\tilde{f} = f$  when restricted to  $A$ .

(In practice, assuming  $f \in C^k(A, \mathbb{R}^n)$  is the same as assuming that  $f \in C^k(U, \mathbb{R}^n)$  for some  $U$  open containing  $A$ , but it is shorter to write.)

25

Here only length with total variation formula. Prove finite for Lipschitz curve.

**DEFINITION 13.50: LENGTH OF A  $C^1$  CURVE**

Let  $\gamma \in C^1([a, b], \mathbb{R}^n)$  be a curve. We define the length of  $\gamma$  as

$$L(\gamma) := \int_a^b |\gamma'(t)| dt,$$

where the vector  $\gamma'(t) = (\gamma'_1(t), \dots, \gamma'_n(t))$  is the **velocity** and the number  $|\gamma'(t)| = \sqrt{\sum_{i=1}^n (\gamma'_i(t))^2}$  is the **speed** of the path at time  $t$ .

**EXERCISE 13.51.** — Show that if  $s: [c, d] \rightarrow [a, b]$  is a  $C^1$  bijective map with  $C^1$  inverse, then  $L(\gamma \circ s) = L(\gamma)$ .

The following result brings some intuition into our definition of length

**THEOREM 13.52: LENGTH AS TOTAL VARIATION**

Let  $\gamma \in C^1([a, b], \mathbb{R}^n)$  be a curve and define

$$V(\gamma) := \sup \left\{ \sum_{j=0}^N |\gamma(t_{j+1}) - \gamma(t_j)| : N \in \mathbb{N}, a = t_0 \leq t_1 \leq \dots \leq t_{N-1} \leq t_N = b \right\}.$$

Then  $L(\gamma) = V(\gamma)$ .

25

*Proof.* (Extra material.) We start showing  $L(\gamma) \geq V(\gamma)$ . Denote  $\gamma(t) = (\gamma_1(t), \dots, \gamma_n(t))$ , fix  $N \in \mathbb{N}$ , some partition  $0 = t_0 \leq t_1 \leq \dots \leq t_{N-1} \leq t_N = 1$  and any set of  $N$  unit vectors in  $\nu_1, \dots, \nu_N \in \mathbb{R}^n$ :

$$\begin{aligned} \int_0^1 |\gamma'(s)| ds &= \sum_{j=0}^N \int_{t_j}^{t_{j+1}} |\gamma'(s)| ds \geq \sum_{j=0}^N \int_{t_j}^{t_{j+1}} \nu_j \cdot \gamma'(s) ds \\ &= \sum_{j=0}^N (\nu_j \cdot \gamma(t_{j+1}) - \nu_j \cdot \gamma(t_j)), \end{aligned}$$

where we used that  $\frac{d}{ds}(\nu \cdot \gamma(s)) = \nu \cdot \gamma'(s)$  and the Fundamental Theorem of Calculus.

With the choice

$$\nu_j := \frac{\gamma(t_{j+1}) - \gamma(t_j)}{|\gamma(t_{j+1}) - \gamma(t_j)|}$$

the last term becomes

$$\sum_{j=0}^N \nu_j \cdot (\gamma(t_{j+1}) - \gamma(t_j)) = \sum_{j=0}^N |\gamma(t_{j+1}) - \gamma(t_j)|,$$

concluding the proof.

To prove the opposite inequality, let us show that  $L(\gamma) \leq V(\gamma) + \varepsilon$  for all  $\varepsilon > 0$  small.



Observe that the function  $\gamma' : [a, b] \rightarrow \mathbb{R}^n$  is uniformly continuous (it is continuous on a compact) so there is  $\delta > 0$  such that if  $s, t \in [a, b]$  and  $|s - t| < \delta$  we have

$$|\gamma'(s) - \gamma'(t)| < \frac{\varepsilon}{2\sqrt{n}(b-a)}$$

Choose a partition  $a = t_0 < t_1 < t_2 < \dots < t_{N-1} \leq t_N = b$  of  $[a, b]$  such that  $t_j - t_{j-1} < \delta$  and  $\xi_j \in [t_{j-1}, t_j]$  for  $1 \leq j \leq N$  such that

$$L(\gamma) < \sum_{j=1}^N (t_j - t_{j-1}) |\gamma'(\xi_j)| + \frac{\varepsilon}{2}. \quad (13.10)$$

By the Mean Value Theorem for all  $1 \leq i \leq n$  and  $1 \leq j \leq N$  we have

$$\gamma_i(t_j) - \gamma_i(t_{j-1}) = \gamma'_i(\eta_{i,j})(t_j - t_{j-1}),$$

for some  $\eta_{i,j} \in [t_{j-1}, t_j]$ . But then since  $|\eta_{i,j} - \xi_j| < \delta$

$$|\gamma'_i(\eta_{i,j}) - \gamma'_i(\xi_j)| \leq |\gamma'(\eta_{i,j}) - \gamma'(\xi_j)| \leq \frac{\varepsilon}{2\sqrt{n}(b-a)},$$

Therefore,

$$|\gamma'_i(\xi_j)|(t_j - t_{j-1}) \leq |\gamma_i(t_j) - \gamma_i(t_{j-1})| + \frac{\varepsilon}{2\sqrt{n}(b-a)}$$

and thus (summing the squares for  $i = 1, \dots, n$ )

$$|\gamma'(\xi_j)|(t_j - t_{j-1}) \leq |\gamma(t_j) - \gamma(t_{j-1})| + \frac{\varepsilon(t_j - t_{j-1})}{2(b-a)}.$$

Inserting this in (13.10) we obtain:

$$L(\gamma) < \sum_{j=1}^N |\gamma(t_j) - \gamma(t_{j-1})| + \varepsilon \leq V(\gamma) + \varepsilon,$$

which concludes the proof.  $\square$

For the discussion that follows it is convenient to recall the symmetries of the Euclidean space:

#### DEFINITION 13.53: EUCLIDEAN ISOMETRIES (OR RIGID MOTIONS)

A map  $F : \mathbb{R}^n \rightarrow \mathbb{R}^n$  of the form  $F(p) = Rp + a$  where  $R$  is a orthogonal  $n \times n$  matrix and  $a \in \mathbb{R}^n$ , is termed **Euclidean isometry** or **rigid motion**.

Notice that Euclidean are the maps that preserve the Euclidean distance:  $d(F(p), F(q)) = d(p, q)$  for every pair of points  $p, q \in \mathbb{R}^n$

13.54. — Besides Theorem (13.52), the appropriateness of  $L(\gamma)$  as a definition of length is supported by the following properties:

1. **Invariance under Reparametrization:**  $L(\gamma)$  remains unchanged under any smooth and bijective reparametrization of  $\gamma$ , indicating that it fundamentally depends on the path's image, not on the parametrization. This property confirms that  $L(\gamma)$  reflects the geometric nature of the curve.
2. **Invariance under Euclidean Isometries:** For any rigid motion  $F : \mathbb{R}^n \rightarrow \mathbb{R}^n$ , characterized by  $|F(x) - F(y)| = |x - y|$  for all  $x, y \in \mathbb{R}^n$ , it holds that  $L(F \circ \gamma) = L(\gamma)$ .  
Indeed, since  $F(p) = Rp + a$ , where  $R$  is an orthogonal matrix, the chain rule gives  $|(F \circ \gamma)'(t)| = |R\gamma'(t)| = |\gamma'(t)|$  for all  $t$ .
3. **Additivity:** The length function  $L$  is additive. Specifically, if  $\gamma : [a, b] \rightarrow \mathbb{R}^n$  and  $c \in (a, b)$ , then  $L(\gamma|_{[a,b]}) = L(\gamma|_{[a,c]}) + L(\gamma|_{[c,b]})$ .
4. **Normalization** The length of the segment  $[0, 1] \times \{0\} \times \cdots \times \{0\}$  is 1.

These properties serve as the foundational guidelines for defining the area of surfaces (i.e. 2-dimensional submanifolds) and, more generally, the  $d$ -volume for  $d$ -dimensional submanifolds.

**DEFINITION 13.55: GRAM DETERMINANT**

Suppose that  $L : \mathbb{R}^m \rightarrow \mathbb{R}^n$  is a linear map where with  $m < n$  (equivalently a  $n \times m$  matrix acting on column  $m$ -vectors by matricial multiplication).

We define the Gram determinant of  $L$  as the square root of the determinant of the  $d \times d$  matrix  $LL^T$ , that is:

$$\sqrt{\det(LL^T)}$$

13.56. — If  $L$  is a  $n \times m$  matrix we can essentially repeat the proof of the polar decomposition theorem starting from the nonnegative definite  $m \times m$  matrix  $LL^T$ . Doing so, we find that  $L = R_2SR_1$  where  $R_1$  is  $m \times m$  orthogonal  $S$  is  $m \times m$  diagonal with nonnegative entries. Now,  $R_2$  is and  $n \times d$  matrix satisfying  $R_2^TR_2 = I_n$ . In other words, the columns of  $R_2$  correspond to the first  $m$  vectors of an orthonormal basis of  $\mathbb{R}^n$ .

In other words any linear map  $L : \mathbb{R}^m \rightarrow \mathbb{R}^n$  can as a composition of:

1. A linear map  $R_1 : \mathbb{R}^m \rightarrow \mathbb{R}^m$  that preserves the lengths of vectors (i.e., an Euclidean isometry).
2. A linear map  $S : \mathbb{R}^m \rightarrow \mathbb{R}^m$  that is diagonal with nonnegative entries (it gives the 'stretching factors')
3. A linear map  $R_2 : \mathbb{R}^m \rightarrow \mathbb{R}^n$  that preserves the lengths of vectors (i.e., an Euclidean isometry).

In this setting, the Gram determinant is the determinant of the ‘stretching matrix’  $S$ .

**DEFINITION 13.57:  $m$ -VOLUME OF PARAMETRIZED  $m$ -SUBMANIFOLD**

Let  $\phi$  be a parametrized  $m$ -dimensional submanifold: more precisely suppose that  $V \subset \mathbb{R}^m$  open and  $\phi \in C^1(V, \mathbb{R}^n)$  is such that  $J\phi(x)$  has rank  $m$  for all  $x \in V$ — we define the  $m$ -volume of  $\phi$  as

$$\text{vol}_m(\phi) = \int_V \sqrt{\det(J\phi^T J\phi)(x)} dx$$

13.58. — Notice the integral  $\int_V \sqrt{\det(J\phi^T J\phi)(x)} dx$  is always well-defined, at least in the improper sense.

**LEMMA 13.59: REPARAMETRIZATION INVARIANCE OF  $m$ -VOLUME**

Assume that  $\psi : U \rightarrow V$  is an any  $C^1$  diffeomorphism. Then,

$$\text{vol}_m(\phi) = \text{vol}_m(\phi \circ \psi)$$

*Proof.* By the chain rule

$$J(\phi \circ \psi)(x) = J\phi(\psi(x))J\psi(x)$$

hence

$$\det(J\psi(x)^T J\phi(\psi(x))^T J\phi(\psi(x))J\psi(x)) = \det(J\psi(x)^T) \det(J\phi(\psi(x))^T J\phi(\psi(x))) \det(J\psi(x))$$

and the Gram determinant for  $J(\phi \circ \psi)(x)$  equals

$$\sqrt{\det(J\psi(x)^T J\phi(\psi(x))^T | \det J\psi(x)| dx}.$$

But by the change of variables formula we obtain (recall  $\psi(U) = V$ ):

$$\begin{aligned} \text{vol}_m(\phi \circ \psi) &= \int_U \sqrt{\det(J\phi(\psi(x))^T J\phi(\psi(x)) | \det \psi(x)|} dx \\ &= \int_{\psi(U)} \sqrt{\det(J\phi(y)^T J\phi(y))} dy = \text{vol}_m(\phi). \end{aligned}$$

In this proof we are assuming that the integrals are well-defined in Riemann sense (as it will always be the case in practice). However, it is not difficult to show that a small modification of this proofs applies when the integrals are defined in the improper sense.  $\square$

**EXERCISE 13.60.** — Prove that if  $F$  is a rigid motion (Euclidean isometry), then the  $d$ -dimensional volume is invariant under  $F$ , i.e.,  $\text{vol}_d(F \circ \phi) = \text{vol}_d(\phi)$ . Additionally, discuss how the additivity and normalization properties can be reinterpreted in the context of the  $d$ -dimensional volume.

EXERCISE 13.61. — Show that the length of a curve is nothing but its  $\text{vol}_d$  for  $d = 1$ .

INTERLUDE: VECTOR PRODUCT IN  $\mathbb{R}^3$

Recall that given two vectors  $a = (a_1, a_2, a_3)$  and  $b = (b_1, b_2, b_3)$  in  $\mathbb{R}^3$ , the cross product (also known as the vector product) is defined as

$$a \times b = \begin{pmatrix} a_2b_3 - a_3b_2 \\ a_3b_1 - a_1b_3 \\ a_1b_2 - a_2b_1 \end{pmatrix}$$

It is customary to compute the cross product as the formal determinant

$$a \times b = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \end{vmatrix},$$

where  $\mathbf{i} = e_1$ ,  $\mathbf{j} = e_2$ , and  $\mathbf{k} = e_3$ .

This determinant can be computed using the rule of Sarrus as follows:

$$a \times b = (a_2b_3 - a_3b_2)\mathbf{i} - (a_1b_3 - a_3b_1)\mathbf{j} + (a_1b_2 - a_2b_1)\mathbf{k}.$$

We also have the following identity:

$$|a \times b|^2 = |a|^2|b|^2 - (a \cdot b)^2. \quad (13.11)$$

This identity can be established by direct computation:

$$|a \times b|^2 = (a_2b_3 - a_3b_2)^2 + (a_3b_1 - a_1b_3)^2 + (a_1b_2 - a_2b_1)^2$$

$$|a|^2|b|^2 = (a_1^2 + a_2^2 + a_3^2)(b_1^2 + b_2^2 + b_3^2).$$

$$(a \cdot b)^2 = (a_1b_1 + a_2b_2 + a_3b_3)^2.$$

Another useful and well-known identity involving the cross product (which can also be established by direct computation) is:

$$(a \times b) \cdot c = \begin{vmatrix} a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \\ c_1 & c_2 & c_3 \end{vmatrix}.$$

This determinant gives the volume of the parallelepiped spanned by the vectors  $a$ ,  $b$ , and  $c$ , and it is zero if and only if the vectors are coplanar.

From now on, and whenever no confusion between the linear map  $Df_x$  and its matrix  $Jf(x)$  in the standard basis is possible, we will also use the (standard) notation  $Df(x)$  to refer to the Jacobi matrix.

LEMMA 13.62: SURFACE AREA

In the case  $d = 2$  and  $n = 3$ , if  $\phi : V \rightarrow \mathbb{R}^3$  then its **area** can be computed as:

$$\begin{aligned} A(\phi) &:= \text{vol}_2(\phi) = \int_U \sqrt{\langle \partial_1 \phi, \partial_1 \phi \rangle \langle \partial_2 \phi, \partial_2 \phi \rangle - \langle \partial_1 \phi, \partial_2 \phi \rangle^2}(x) dx_1 dx_2 \\ &= \int_U |\partial_1 \phi \times \partial_2 \phi|(x) dx_1 dx_2 \end{aligned}$$

where we integrate the Euclidean modulus of  $\partial_1 \phi \times \partial_2 \phi : U \rightarrow \mathbb{R}^3$ : the function that maps  $x$  to the cross (or vector) product of the two vector  $\partial_1 \phi(x)$  and  $\partial_2 \phi(x)$ .

*Proof.* The Jacobi matrix  $D\phi$  of  $\phi$  at a point  $x \in U$  is a  $3 \times 2$  matrix given by:

$$D\phi(x) = \begin{bmatrix} \partial_1 \phi_1 & \partial_2 \phi_1 \\ \partial_1 \phi_2 & \partial_2 \phi_2 \\ \partial_1 \phi_3 & \partial_2 \phi_3 \end{bmatrix}.$$

The transpose of  $D\phi(x)$ , denoted  $D\phi^T$ , is a  $2 \times 3$  matrix:

$$D\phi^T = \begin{bmatrix} \partial_1 \phi_1 & \partial_1 \phi_2 & \partial_1 \phi_3 \\ \partial_2 \phi_1 & \partial_2 \phi_2 & \partial_2 \phi_3 \end{bmatrix}.$$

The Gram matrix is the product of  $D\phi(x)^T$  and  $D\phi(x)$ :

$$D\phi^T D\phi = \begin{bmatrix} \partial_1 \phi \cdot \partial_1 \phi & \partial_1 \phi \cdot \partial_2 \phi \\ \partial_2 \phi \cdot \partial_1 \phi & \partial_2 \phi \cdot \partial_2 \phi \end{bmatrix}$$

which is a  $2 \times 2$  matrix representing the inner products of the partial derivatives.

Hence, the determinant of this matrix is:

$$\det(D\phi(x)^T D\phi(x)) = (\partial_1 \phi \cdot \partial_1 \phi)(\partial_2 \phi \cdot \partial_2 \phi) - (\partial_1 \phi \cdot \partial_2 \phi)^2 = |\partial_1 \phi|^2 |\partial_2 \phi|^2 - (\partial_1 \phi \cdot \partial_2 \phi)^2$$

But using the identity (13.11) the right hand side is equal to  $|\partial_1 \phi \times \partial_2 \phi|^2$ , and hence the lemma follows. □

EXAMPLE 13.63. — Consider the parameterization of a sphere of radius  $r$  in  $\mathbb{R}^3$  given by:

$$\psi(\theta, \varphi) = \begin{pmatrix} r \sin \theta \cos \varphi \\ r \sin \theta \sin \varphi \\ r \cos \theta \end{pmatrix},$$

where  $0 < \theta < \pi$  and  $0 \leq \varphi < 2\pi$ .

To compute the area of the sphere, we first find the partial derivatives of  $\psi$ :

$$\partial_\theta \psi = \begin{pmatrix} r \cos \theta \cos \varphi \\ r \cos \theta \sin \varphi \\ -r \sin \theta \end{pmatrix}$$

and

$$\partial_\varphi \psi = \begin{pmatrix} -r \sin \theta \sin \varphi \\ r \sin \theta \cos \varphi \\ 0 \end{pmatrix}.$$

The Gram matrix is given by

$$\begin{bmatrix} \partial_\theta \psi \cdot \partial_\theta \psi & \partial_\theta \psi \cdot \partial_\varphi \psi \\ \partial_\varphi \psi \cdot \partial_\theta \psi & \partial_\varphi \psi \cdot \partial_\varphi \psi \end{bmatrix} = \begin{bmatrix} r^2 & 0 \\ 0 & r^2 \sin^2 \theta \end{bmatrix}$$

Hence the gram determinant (the square root of the determinant of the matrix above) is

$$r^2 \sin \theta.$$

Finally, the area  $A$  of the sphere is given by:

$$A = \int_0^{2\pi} \int_0^\pi r^2 \sin \theta \, d\theta \, d\varphi = 4\pi r^2,$$

the expected result.

#### DEFINITION 13.64: SUPPORT OF A FUNCTION

Let  $U \subset \mathbb{R}^n$  an open set and  $f : U \rightarrow \mathbb{R}$  a continuous function. We define the **support** of  $f$ , denoted  $\text{spt}(f) \subset \mathbb{R}^n$  as the set

$$\text{spt}(f) = \overline{\{x \in U \mid f(x) \neq 0\}}.$$

If  $\text{spt}(f)$  is bounded and contained in  $U$  then we say that  $f$  is **compactly supported in  $U$** .

## LEMMA 13.65: EXTENSION OF COMPACTLY SUPPORTED FUNCTIONS

If  $U \subset \mathbb{R}^n$  is open  $f : U \rightarrow \mathbb{R}$  is of class  $C^k$  for some  $k \geq 0$  continuous and compactly supported in  $U$ , then the function  $\tilde{f} : \mathbb{R}^n \rightarrow \mathbb{R}$  defined as

$$\tilde{f}(x) = \begin{cases} f(x) & \text{for } x \in U \\ 0 & \text{for } x \in \mathbb{R}^n \setminus U \end{cases}$$

belongs to  $C^k(\mathbb{R}^n)$ .

*Proof.* It is clear that  $\tilde{f}$  has continuous  $k$ -th order partial derivatives inside  $U$ . And also in  $\mathbb{R}^n \setminus \bar{U}$ , since  $\tilde{f} \equiv 0$  in this open set.

Hence to prove that  $\tilde{f}$  is of class  $C^k$  we only need to show that it has continuous  $k$ -th order partial derivatives across the boundary  $\partial U$ . But since  $\text{spt}(f)$  is closed and contained in  $U$ , for every point in  $x_o \in \partial U$  there is  $r > 0$  such that  $\tilde{f} \equiv 0$   $B_r(x_o)$ . Hence, that  $\tilde{f}$  and all of its partial derivatives (all zero) are indeed continuous across  $\partial U$ .  $\square$

26

After we discuss the notion of  $d$ -volume for  $d$ -submanifolds we can accordingly define integrals of compactly supported continuous function with respect to the  $d$ -volume measure

DEFINITION 13.66: INTEGRATION WITH RESPECT TO THE  $m$ -VOLUME

Assume that  $M \subset \mathbb{R}^n$  a parametrized  $m$ -dimensional submanifold: that is, suppose that there are  $V \subset \mathbb{R}^m$  open, and  $\phi \in C^1(V, \mathbb{R}^n)$ , such that  $J\phi(x)$  has rank  $m$  for all  $x \in V$ , such that  $M = \phi(V)$ . Given  $f : M \rightarrow \mathbb{R}$  continuous (where  $M$  is endowed with the restriction of the Euclidean metric) and with  $\text{spt}(f \circ \phi) \subset V$  we define

$$\int_M f \, d\text{vol}_m = \int_M f(p) \, d\text{vol}_m(p) := \int_V f \circ \phi \sqrt{\det(J\phi^T J\phi)}(x) \, dx$$

13.67. — We notice  $f \circ \phi : V \rightarrow \mathbb{R}$  is continuous and  $\text{spt}(f \circ \phi) \subset \phi$ . This implies (exercise) that  $(f \circ \phi)$  is Riemann integrable over  $V$ .



LEMMA 13.68: INVARIANCES OF INTEGRAL WITH RESPECT TO  $m$ -VOLUME

The integral  $\int_M f \, d\text{vol}_m$  as in Definition 13.66 is invariant under reparametrization of  $M$  and Euclidean isometries.

More precisely: let  $M = \phi(V)$  be  $C^1$  parametrized  $m$ -submanifold, where  $V \subset \mathbb{R}^m$  open, and for  $f : M \rightarrow \mathbb{R}$  continuous such that  $\text{spt}(f \circ \phi) \subset V$ ,

- If  $\Psi : U \rightarrow V$  is a  $C^1$  diffeomorphism then

$$\begin{aligned} \int_M f \, d\text{vol}_m &= \int_V f \circ \phi \sqrt{\det(J\phi^T J\phi)}(x) \, dx \\ &= \int_U f \circ \phi \circ \psi \sqrt{\det(J(\phi \circ \psi)^T J(\phi \circ \psi)}(y) \, dy \end{aligned}$$

- If  $F : \mathbb{R}^n \rightarrow \mathbb{R}^n$  is an Euclidean isometry then

$$\int_{F(M)} f \circ F^{-1} \, d\text{vol}_m = \int_M f \, d\text{vol}_m.$$

*Proof.* The invariance under reparametrizations is proven exactly as in Lemma 13.59.

For the invariance under Euclidean isometry, let  $F(p) = Rp + a$  where  $R$  is orthogonal and  $a \in \mathbb{R}^n$ . Notice that  $F \circ \phi$  is a parametrization of  $F(M)$  and  $F \circ \phi$  is a parametrization of  $F(M)$  and  $(f \circ F^{-1}) : F(M) \rightarrow \mathbb{R}$  is continuous (being the composition of continuous maps). Then,

$$\int_{F(M)} f \circ F^{-1} \, d\text{vol}_m = \int_V f \circ F^{-1} \circ F \circ \phi \sqrt{\det(J(F \circ \phi)^T J(F \circ \phi))}(x) \, dx.$$

But since  $F(p) = Rp + a$ , we have  $JF(p) = R$  for all  $p$ . Therefore, by the chain rule:

$$J(F \circ \phi)(x) = RJ\phi(x).$$

Thus

$$J(F \circ \phi)(x)^T J(F \circ \phi)(x) = J\phi(x)^T R^T R J\phi(x) = J\phi(x)^T J\phi(x)$$

and the lemma follows.  $\square$

We finish the section showing how to compute  $(n - 1)$ -volumes and integrals in the important case of  $(n - 1)$ -dimensional graphs in  $\mathbb{R}^n$ .

LEMMA 13.69: VOLUME ELEMENT ON GRAPHICAL  $(n - 1)$ -SUBMANIFOLDS

Let  $V \subset \mathbb{R}^{n-1}$  open and let  $g \in C^1(V)$ . Consider the graphical  $(n - 1)$ -submanifold  $M = \{x \in \mathbb{R}^n \times \mathbb{R} \mid x_n = g(x_1, \dots, x_{n-1})\}$  with graphical parametrization

$$\phi(x) = (x_1, \dots, x_{n-1}, g(x_1, \dots, x_{n-1}))^T.$$

Then,

$$\begin{aligned} \int_M f \, d\text{vol}_{n-1} &= \int_M f(p) \, d\text{vol}_{n-1}(p) = \\ &= \int_V (f \circ \phi)(x_1, \dots, x_{n-1}) \sqrt{1 + |\nabla g(x_1, \dots, x_{n-1})|^2} \, dx_1 \cdots dx_{n-1}. \end{aligned}$$

In particular, for  $f \equiv 1$  we obtain

$$\text{vol}_{n-1}(M) = \int_V \sqrt{1 + |\nabla g(x_1, \dots, x_{n-1})|^2} \, dx.$$

*Proof.* The Jacobian matrix of the parametrization  $\phi : \mathbb{R}^{n-1} \rightarrow M$  defined by  $\phi(x) = (x, g(x))$  is:

$$J\phi(x) = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \\ \frac{\partial g}{\partial x_1}(x) & \frac{\partial g}{\partial x_2}(x) & \cdots & \frac{\partial g}{\partial x_{n-1}}(x) \end{bmatrix}.$$

Computing the Gram matrix  $G = J\phi(x)^T J\phi(x)$  we obtain:

$$G = \begin{bmatrix} 1 + \left(\frac{\partial g}{\partial x_1}(x)\right)^2 & \frac{\partial g}{\partial x_1}(x) \frac{\partial g}{\partial x_2}(x) & \cdots & \frac{\partial g}{\partial x_1}(x) \frac{\partial g}{\partial x_{n-1}}(x) \\ \frac{\partial g}{\partial x_2}(x) \frac{\partial g}{\partial x_1}(x) & 1 + \left(\frac{\partial g}{\partial x_2}(x)\right)^2 & \cdots & \frac{\partial g}{\partial x_2}(x) \frac{\partial g}{\partial x_{n-1}}(x) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial g}{\partial x_{n-1}}(x) \frac{\partial g}{\partial x_1}(x) & \frac{\partial g}{\partial x_{n-1}}(x) \frac{\partial g}{\partial x_2}(x) & \cdots & 1 + \left(\frac{\partial g}{\partial x_{n-1}}(x)\right)^2 \end{bmatrix}.$$

This may seem involved, but we can easily compute the determinant of the Gram matrix  $G$  by using the following trick.

For  $x$  fixed, put

$$w := \left[ \frac{\partial g}{\partial x_1}(x) \quad \frac{\partial g}{\partial x_2}(x) \quad \cdots \quad \frac{\partial g}{\partial x_{n-1}}(x) \right]^T$$

and choose some orthogonal  $(n - 1) \times (n - 1)$  matrix  $R$  such that  $Rw = |w|e_1$  (for example, any rotation sending  $w/|w|$  to  $e_1$ )

Notice that, since  $w^T R^T = |w|e_1^T$ , we have

$$J\phi(x)R^T = \begin{bmatrix} \text{Id} \\ w^T \end{bmatrix} R^T = \begin{bmatrix} R^T \\ |w|e_1^T \end{bmatrix} = \begin{bmatrix} R^T \\ |w| & 0 & \cdots & 0 \end{bmatrix}.$$

Hence,  $RGR^T = RJ\phi(x)^T J\phi(x)R^T$  is computed as:

$$\begin{bmatrix} R & \begin{bmatrix} |w| \\ 0 \\ \vdots \\ 0 \end{bmatrix} \\ \hline & \begin{bmatrix} R^T \\ |w| & 0 & \cdots & 0 \end{bmatrix} \end{bmatrix} = \begin{bmatrix} 1 + |w|^2 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix},$$

where we used  $RR^T = I_{n-1}$ .

Since orthogonal matrices have determinant one,  $\det(R) = 1$ , we have:

$$\det(G) = \det(RGR^T) = 1 + |w|^2 = 1 + |\nabla g(x)|^2.$$

Hence, the ‘volume element’ on the manifold  $M$  is:

$$d\text{vol}_{n-1} = \sqrt{\det(G)} dx = \sqrt{1 + |\nabla g(x)|^2} dx,$$

completing the proof. □

## 13.7 From local to global: partition of unity on compact submanifolds

In the previous section we discussed how to compute the  $d$ -volume (or the integral of a function) on ‘pieces’ of  $d$ -dimensional submanifold that can be described using a single parametrization. Here, we will see how can we define integrals over arbitrary compact submanifolds  $M \subset \mathbb{R}^n$ , which may not admit one single global parametrization.

To do so, it is convenient to introduce the so-called partitions of unity.

## LEMMA 13.70: PARTITION OF UNITY

Let  $K \subset \mathbb{R}^n$  be a compact set  $\mathcal{B} = \{B_{r_1}(p_1), B_{r_2}(p_2), \dots, B_{r_N}(p_N)\}$  some finite collection of open balls (of  $\mathbb{R}^n$ ) covering  $K$ , i.e.,  $K \subset \bigcup_{\ell=1}^N B_{r_\ell}(p_\ell)$ .

Then there exists a collection of functions  $\eta_\ell : \mathbb{R}^n \rightarrow [0, \infty)$  of class  $C^\infty$  such that  $\text{spt}(\eta_\ell) \subset B_{r_\ell}(p_\ell)$  for each  $\ell$  and

$$\sum_{\ell=1}^N \eta_\ell(x) = 1 \quad \text{for all } x \in U,$$

where  $U$  is some open set containing  $K$ .

*Proof.* We first show that for some  $\theta \in (0, 1)$   $B_{\theta r_\ell}(p_\ell)$  stills a covers of  $K$ . Indeed, for  $k \geq 1$  integer  $\theta_k = 1 - 2^{-k}$  and let  $\Omega_k := \bigcup_{\ell=1}^N B_{\theta_k r_\ell}(p_\ell)$ . Observe that  $\Omega_1 \subset \Omega_2 \subset \Omega_3 \subset \dots$  and that

$$\bigcup_{k=1}^{\infty} \Omega_k = \bigcup_{\ell=1}^N B_{r_\ell}(p_\ell)$$

is an open cover of  $K$ . Hence, it has a finite subcover, implying that there exists  $k_\circ$  such that

$$\Omega_{k_\circ} \supset K.$$

26

We can then take  $\theta := \theta_{k_\circ}$ .

Now, define, for  $x \in \mathbb{R}^n$ :

$$\Phi(x) = \begin{cases} \exp\left(\frac{-1}{1-|x|^2}\right) & \text{if } |x| < 1 \\ 0 & \text{if } |x| \geq 1, \end{cases}$$

$$\tilde{\Phi}(x) = \begin{cases} \exp\left(\frac{-1}{|x|^2-1}\right) & \text{if } |x| > 1 \\ 0 & \text{if } |x| \leq 1, \end{cases}$$

and notice that  $\Phi, \tilde{\Phi} \in C^\infty(\mathbb{R}^n)$  and  $\text{spt}(\Phi) = \overline{B_1}$  and  $\text{spt}(\tilde{\Phi}) = \mathbb{R}^n \setminus B_1$ .

Define

$$\xi_\ell(x) := \Phi\left(\frac{x - p_\ell}{\theta^{1/3} r_\ell}\right)$$

$$\xi_\ell(x) = \tilde{\Phi}\left(\frac{x - p_\ell}{\theta r_\ell}\right) \quad \text{and} \quad \tilde{\xi}(x) := \prod_{i=1}^N \xi_\ell(x).$$

Notice that  $\theta < \theta^{2/3} < \theta^{1/3} < 1$ ; that for  $x \in \overline{B_{\theta^{2/3} r_\ell}(p_\ell)}$  we have

$$\xi_\ell(x) \geq \exp\left(\frac{-1}{1 - (\theta^{-1/3} + 2/3)^2}\right) > 0;$$

and that for  $x \in \mathbb{R}^n \setminus (\cup_{1 \leq \ell \leq N} B_{\theta^{2/3}r_\ell}(p_\ell))$

$$\tilde{\xi}(x) \geq \exp\left(\frac{-N}{(\theta^{-1+2/3})^2 - 1}\right) > 0.$$

On the other hand putting

$$U := \cup_{\ell=1}^N B_{\theta r_\ell}(p_\ell) \supset K$$

we have

$$\tilde{\xi}_\ell = 0 \in B_{\theta r_\ell}(p_\ell)$$

and hence their product vanishes in the union of the balls:

$$\tilde{\xi} = 0 \in U.$$

Therefore,

$$\eta_\ell(x) := \frac{\xi_\ell(x)}{\sum_{\ell=1}^N \xi_\ell(x) + \tilde{\xi}(x)}$$

are  $C^\infty$  functions (the denominator is always positive), and by construction  $\eta_\ell \geq 0$ ,  $\text{spt}(\eta_i) = \overline{B_{\theta^{1/3}r_\ell}(x_i)} \subset B_{r_\ell}(x_i)$  and  $\sum_{\ell=1}^N \eta_i = 1$  in  $\cup_{1 \leq \ell \leq N} B_{\theta r_\ell}(p_\ell) \supset K$  since  $\tilde{\xi}$  vanishes in this set.  $\square$

#### DEFINITION 13.71: PARTITION OF UNITY (FOR COMPACT SUBMANIFOLDS OF $\mathbb{R}^n$ )

Let  $M \subset \mathbb{R}^n$  be a compact  $C^k$  submanifold of dimension  $m$ , where  $1 \leq m < n$ , and  $\mathcal{B} = \{B_{r_1}(p_1), B_{r_2}(p_2), \dots, B_{r_N}(p_N)\}$ , with  $p_\ell \in M$ , some collection of open balls (of  $\mathbb{R}^n$ ) covering  $M$ , i.e., with  $M \subset \cup_{\ell=1}^N B_{r_\ell}(p_\ell)$ .

We call any collection of  $N$  functions  $\eta_\ell : \mathbb{R}^n \rightarrow [0, \infty)$  as in Lemma 13.70 (for  $K = M$ ) a **partition of unity on  $M$  subordinated to  $\mathcal{B}$** .

**DEFINITION 13.72: GRAPHICAL COVER OF A COMPACT  $m$ -SUBMANIFOLD**

Let  $M \subset \mathbb{R}^n$  be a compact  $m$ -dimensional submanifold of class  $C^k$ , with  $k \geq 1$ . We call collection of  $N$  maps  $\phi_\ell : V_\ell \rightarrow B_{r_\ell}(p_\ell) \subset \mathbb{R}^n$ ,  $1 \leq \ell \leq N$ , of class  $C^k$ , a **graphical cover of  $M$**  if the following properties are satisfied:

1. For each  $\ell$ ,  $V_\ell$  is an open set of  $\mathbb{R}^m$ ,  $r_\ell > 0$ , and  $p_\ell \in M$ .
2. For each  $\ell$ , there is an orthogonal  $n \times n$  matrix  $R_\ell$  such that  $\phi_\ell : V_\ell \rightarrow \mathbb{R}^n$  is of the form

$$\phi_\ell(x_1, \dots, x_m) = R_\ell(x_1, \dots, x_m, g_\ell(x_1, \dots, x_m))^T,$$

for some  $g_\ell \in C^k(V_\ell, \mathbb{R}^{n-m})$ .

3. For all  $\ell$ , we have  $\phi_\ell(V_\ell) = M \cap B_{r_\ell}(p_\ell)$ .
4. The balls  $B_{r_\ell}(p_\ell)$  cover  $M$ : i.e.,  $M \subset \bigcup_{\ell=1}^N B_{r_\ell}(p_\ell)$ .

**PROPOSITION 13.73: EXISTENCE OF GRAPHICAL COVERS**

*Any compact  $m$ -dimensional  $C^k$  submanifold admits graphical covers.*

*Proof.* Using Proposition 12.12 (4), for any given point  $p_o \in M$  exists  $V_{p_o} \subset \mathbb{R}^m$  open and  $g_{p_o} \in C^1(V_{p_o}, \mathbb{R}^{n-m})$  and  $R_{p_o}$ , a permutation of the coordinates (which is in particular an orthogonal linear map), such that, for some open neighborhood  $U$  of  $p_o$  (i.e., for some open set  $U \subset \mathbb{R}^n$  containing  $p_o$ ) we have

$$M \cap U_{p_o} = \phi_{p_o}(V_{p_o}),$$

where

$$\phi_{p_o}(x_1, \dots, x_m) = R_{p_o}(x_1, \dots, x_m, g_\ell(x_1, \dots, x_m))^T$$

For each  $p_o$ , let  $r_{p_o} > 0$  be such that  $B_{r_{p_o}}(p_o) \subset U_{p_o}$ .

Now, since  $M$  is compact and the collection of open balls  $\{B_{r_{p_o}}(p_o)\}_{p_o \in M}$  covers  $M$ , there is a finite subcover. Let  $p_1, p_2, \dots, p_N$  be the centers of the balls in the subcover. Then defining  $r_\ell := r_{p_\ell}$ ,  $V_\ell := \phi_{p_\ell}^{-1}(B_{r_\ell}(p_\ell))$  (which is open subset of  $\mathbb{R}^m$ ), and  $\phi_\ell := \phi_{p_\ell}|_{V_\ell}$  we obtain the wished graphical cover of  $M$ .  $\square$

**DEFINITION 13.74: INTEGRAL OVER A COMPACT SUBMANIFOLD**

Let  $M \subset \mathbb{R}^n$  be an  $m$ -dimensional compact  $C^1$  submanifold and  $f : M \rightarrow \mathbb{R}$  a continuous function (where  $M$  is endowed with the restriction of the Euclidean metric).

Given a graphical cover  $\phi_\ell : V_\ell \rightarrow B_{r_\ell}(p_\ell)$ ,  $1 \leq \ell \leq N$ , let  $\eta_\ell : \mathbb{R}^n \rightarrow [0, \infty)$  be a partition of unity on  $M$  subordinated to  $\{B_{r_\ell}(p_\ell)\}_{1 \leq \ell \leq N}$ , we define:

$$\int_M f \, d\text{vol}_d = \sum_{\ell=1}^N \int_{M \cap B_\ell(x_\ell)} f \eta_\ell \, d\text{vol}_d = \sum_{\ell=1}^N \int_{V_\ell} f(\phi_\ell(y)) \eta_\ell(\phi_\ell(y)) \sqrt{J\phi_\ell^T J\phi_\ell}(y) \, dy_1 \dots dy_d.$$

**LEMMA 13.75: INTEGRAL IS INDEPENDENT OF COVER AND PARTITION OF UNITY**

Let  $M \subset \mathbb{R}^n$  be an  $m$ -dimensional compact  $C^k$  submanifold and  $f : M \rightarrow \mathbb{R}$  a continuous function.

Given two graphical covers  $\phi_\ell : V_\ell \rightarrow B_{r_\ell}(x_\ell)$ ,  $1 \leq \ell \leq N$ , and  $\tilde{\phi}_k : W_k \rightarrow B_{s_k}(y_k)$  of  $M$ , where  $1 \leq \ell \leq N$  and  $1 \leq m \leq \tilde{N}$ ; and corresponding partitions of unity  $\eta_\ell, \tilde{\eta}_k$ .

Then:

$$\sum_{\ell=1}^N \int_{V_\ell} (f\eta_\ell) \circ \phi_\ell \sqrt{J\phi_\ell^T J\phi_\ell} = \sum_{m=1}^{\tilde{N}} \int_{W_k} (f\tilde{\eta}_k) \circ \tilde{\phi}_k \sqrt{J\tilde{\phi}_k^T J\tilde{\phi}_k}$$

27

*Proof.* (Extra material) By definition of graphical cover we have

$$\phi_\ell(x_1, \dots, x_d) = R_\ell(x_1, \dots, x_d, g_\ell(x_1, \dots, x_d))^T,$$

$$\tilde{\phi}_k(x_1, \dots, x_d) = \tilde{R}_k(x_1, \dots, x_d, \tilde{g}_k(x_1, \dots, x_d))^T,$$

for suitable maps  $g_\ell : V_\ell \rightarrow \mathbb{R}$ ,  $\tilde{g}_k : W_k \rightarrow \mathbb{R}$  and orthogonal matrices  $R_\ell, \tilde{R}_k$ .

Let  $\pi_d : \mathbb{R}^n \rightarrow \mathbb{R}^d$  denote the standard projection onto the first  $d$  components:

$$(x_1, \dots, x_d, x_{d+1}, \dots, x_n) \mapsto (x_1, \dots, x_d)$$

The key observation to prove this lemma is that, for all  $\ell$ ,

$$\phi_\ell^{-1}(x) = \pi_d(R_\ell^T x) \quad \text{for all } x \in \phi_\ell(\overline{V}_\ell);$$

(recall that for orthogonal matrices  $R^{-1} = R^T$ ), and similarly, for all  $m$ ,

$$\tilde{\phi}_k^{-1}(x) = \pi_d(\tilde{R}_k^T x) \quad \text{for all } x \in \tilde{\phi}_k(\overline{W}_k);$$

Notice that for every pair  $\ell, k$ , define the open sets

$$V_{\ell,k} := \phi_\ell^{-1}(B_{s_k}(y_k)) \subset V_\ell \quad \text{and} \quad W_{k,\ell} := \tilde{\phi}_k^{-1}(B_{r_\ell}(x_\ell)) \subset W_k$$

and the ‘transition’ maps

$$\psi_{\ell,k} : V_{\ell,k} \rightarrow W_{k,\ell} \quad \text{and} \quad \psi_{k,\ell} : W_{k,\ell} \rightarrow V_{\ell,k}$$

respectively defined as

$$\psi_{k,\ell}(x) = \pi_d(\tilde{R}_k^T \phi_\ell(x)) \quad \text{and} \quad \psi_{\ell,k}(x) = \pi_d(R_\ell^T \tilde{\phi}_k(x)).$$

Notice that they are  $C^k$  diffeomorphisms. Indeed, on the one hand  $\psi_{\ell,k}$  and  $\psi_{k,\ell}$  are inverse to each other. On the other hand they are both  $C^k$  maps being the composition of a  $C^k$  maps, and a linear map.

Now, notice that

$$\sum_{\ell=1}^N \sum_{k=1}^{\tilde{N}} (\eta_\ell \tilde{\eta}_k)(x) = \left( \sum_{\ell=1}^N \eta_\ell(x) \right) \left( \sum_{m=1}^{\tilde{N}} \tilde{\eta}_m(x) \right) = 1 \quad \text{for all } x \in M$$

where each function  $(\eta_\ell \tilde{\eta}_k) : \mathbb{R}^n \rightarrow [0, \infty)$  has compact support in  $B_{r_\ell}(x_\ell) \cap B_{s_k}(y_k)$ . As a consequence,

$$\text{spt}((\eta_\ell \tilde{\eta}_k) \circ \phi_\ell) \subset V_{\ell,k}.$$

By these observations we can put:

$$\begin{aligned} \sum_{\ell=1}^N \int_{V_\ell} (f \eta_\ell) \circ \phi_\ell \sqrt{J \phi_\ell^T J \phi_\ell} &= \sum_{\ell=1}^N \int_{V_\ell} \left( \sum_{k=1}^{\tilde{N}} (f \eta_\ell \tilde{\eta}_k) \circ \phi_\ell \sqrt{J \phi_\ell^T J \phi_\ell} \right). \\ &= \sum_{\ell=1}^N \sum_{k=1}^{\tilde{N}} \int_{V_{\ell,k}} (f \eta_\ell \tilde{\eta}_k) \circ \phi_\ell \sqrt{J \phi_\ell^T J \phi_\ell}. \end{aligned}$$

But since

$$\phi_\ell = \tilde{\phi}_k \circ \psi_{\ell,k} \quad \text{in } V_{\ell,k},$$

and  $\psi_{\ell,k}$  is a diffeomorphism, the invariance of the integral with respect to  $d\text{vol}_d$  under reparametrization (Lemmas 13.59 and 13.68) yields

$$\int_{V_{\ell,k}} (f \eta_\ell \tilde{\eta}_k) \circ \phi_\ell \sqrt{J \phi_\ell^T J \phi_\ell} = \int_{W_{k,\ell}} (f \eta_\ell \tilde{\eta}_k) \circ \tilde{\phi}_k \sqrt{J \tilde{\phi}_k^T J \tilde{\phi}_k}.$$

Then summing over all  $\ell$  and  $k$  the lemma follows.  $\square$



---

# Chapter 14

## Global integral theorems

In this chapter, we deal with integral theorems for vector fields in  $\mathbb{R}^n$ , which are in some sense multidimensional generalizations of the fundamental theorem of integral and differential calculus.

### 14.1 The integration by parts formula

#### 14.1.1 Bounded, smooth domains, exterior unit normal

##### DEFINITION 14.1: TANGENT AND NORMAL VECTORS

Let  $M \subset \mathbb{R}^n$  be a  $m$ -dimensional  $C^1$  submanifold. For a given point  $p \in M$ , we say that a vector  $\tau \in \mathbb{R}^n$  is **tangent** to  $M$  at  $p$  if there exists a sequence  $p_k \in M$  with  $p_k \rightarrow p$  and  $p_k \neq p$ , and a sequence of positive number  $r_k$  converging to zero such that

$$\lim_k \frac{p_k - p}{r_k} \rightarrow \tau.$$

We say that a vector  $\nu \in \mathbb{R}^n$  is **normal** (i.e., perpendicular) to  $M$  at  $p$  if

$$\tau \cdot \nu = 0 \quad \text{for all } \tau \text{ tangent to } M \text{ at } p.$$

##### LEMMA 14.2: THE TANGENT SPACE: CHARACTERIZATION FOR GRAPHS

Let  $M$  be a  $m$ -dimensional submanifold of  $\mathbb{R}^n$ .

Assume that  $\phi : V_\ell \rightarrow B_r(p_0)$  is a  $C^1$  graphical parametrization of  $M \cap B_r(p_0)$ , for some  $p_0 \in M$  and let  $p$  be any point belongs to  $M \cap B_r(p_0)$ .

Then,  $\tau \in \mathbb{R}^n$  is tangent to  $M$  at  $p$  if and only if  $\tau$  belongs to the linear space  $D\phi_p(\mathbb{R}^m) \subset \mathbb{R}^n$  (the range of the map  $D\phi_p$ ).

In particular, the collection of all tangent vectors to  $M$  at given point  $p \in M$  is a linear subspace of  $\mathbb{R}^n$ : the so-called the **tangent space of  $M$  at  $p$** .

*Proof.* We have

$$\phi(y) = R(y, g(y))^T,$$

for some  $g : V \rightarrow \mathbb{R}^{m-n}$  of class  $C^1$ .

Let  $y \in V$  be the (unique) point mapped to  $p$ : that is  $p = \phi(y)$

Let us show that first that if  $\tau \in D\phi_p(\mathbb{R}^m)$  then  $\tau$  is tangent to  $M$  at  $p$ . Indeed, let  $z$  be such that  $D\phi_p(z) = \tau$ . Then since  $\phi$  is  $C^1$ , for any sequence  $r_k > 0$  converging to 0 we have

$$\lim_{r_k \rightarrow 0} \frac{\phi(y + r_k z) - \phi(y)}{r_k} = D\phi_p(z)$$

Hence defining  $p_k := \phi(y + r_k z) \rightarrow p$  we have shown

$$\lim_{r_k \rightarrow 0} \frac{p_k - p}{r_k} = D\phi_p(z) = \tau,$$

hence  $\tau$  is tangent. This proves that every vector belonging to the image of  $D\phi_p$  is tangent to  $M$  at  $p$ .

To prove the opposite implication, let  $\pi_m : \mathbb{R}^n \rightarrow \mathbb{R}^m$  be the projection onto the first  $m$ -components  $(x_1, \dots, x_m, x_{m+1}, \dots, x_n) \mapsto (x_1, \dots, x_m)$

Notice that

$$(\pi_m \circ R^T)(\phi(y)) = y \quad \text{for all } y \in V.$$

Given any sequence  $p_k \rightarrow p$  in  $M \cap U$  such that

$$\lim_k \frac{p_k - p}{r_k} \rightarrow \tau,$$

for some  $r_k > 0$  put  $y_k := (\pi_m \circ R^T)(p_k)$ .

Notice that  $(\pi_m \circ R^T)$  is 1-Lipschitz, since it is the composition of orthogonal map and a projection. Therefore, the sequence

$$z_k := \frac{y_k - y}{r_k}$$

must be bounded, since

$$|z_k| = \frac{|y_k - y|}{r_k} \leq \frac{|p_k - p|}{r_k} \rightarrow |\tau| < \infty.$$

Hence  $z_k$  has accumulation points in  $\mathbb{R}^n$ . Let  $z_{k_\ell} \rightarrow z \in \mathbb{R}^m$  as  $\ell \rightarrow \infty$  be one of the accumulation points.

Then, using again that  $\phi$  is  $C^1$

$$\frac{p_k - p}{r_k} = \frac{\phi(y + r_k z_k) - \phi(y)}{r_k} = D\phi_y(z_k) + \frac{o(|r_k z_k|)}{r_k}.$$

Using this for  $k = k_\ell$  and sending  $\ell \rightarrow \infty$  we obtain

$$\tau = \lim_\ell \frac{p_{k_\ell} - p}{r_{k_\ell}} = \lim_\ell D_y \phi(z_{k_\ell}) = D\phi_y(z).$$

This proves that every tangent vector at  $p$  belongs to  $D\phi_p(\mathbb{R}^m)$ .  $\square$

**DEFINITION 14.3: BOUNDED  $C^k$  DOMAIN**

Given  $k \geq 1$ , bounded open set  $\Omega \subset \mathbb{R}^n$  is called **bounded  $C^k$  domain** if  $M := \partial\Omega$  is a  $(n-1)$ -dimensional submanifold of class  $C^k$ .

**DEFINITION 14.4: EXTERIOR AND INTERIOR NORMAL VECTORS**

Suppose that  $\Omega$  is a bounded  $C^1$  domain. A normal vector  $\nu$  to  $\partial\Omega$  at  $x \in \partial\Omega$  is called **exterior** if

$$p + h\nu \in \mathbb{R}^n \setminus \bar{\Omega} \quad \text{for all } h > 0 \text{ sufficiently small.}$$

And it is called **interior** if

$$p + h\nu \in \mathbb{R}^n \subset \Omega \quad \text{for all } h > 0 \text{ sufficiently small.}$$

27

**PROPOSITION 14.5: EXTERIOR UNIT NORMAL FOR GRAPHS**

Suppose that  $\Omega$  is a bounded  $C^1$  domain. Put  $y := (x_1, \dots, x_{n-1})$  and assume that

$$\phi(y) = (y, g(y))^T,$$

where  $V \subset \mathbb{R}^{n-1}$  is open and  $g : V \rightarrow \mathbb{R}$ , is a graphical  $C^1$  parametrization of  $M \cap B_r(p_0)$ .

Suppose also that:

$$\Omega \cap B_r(p_0) = \{x = (y, x_n) \in \mathbb{R}^n \mid x_n < g(y)\} \cap B_r(p_0)$$

Then, for  $p = (y, g(y))$ , the vector

$$\nu(p) = \nu(y, g(y)) := \frac{(-\nabla g(y), 1)}{\sqrt{1 + |\nabla g(y)|^2}}$$

is normal, exterior, and has norm one.

By Lemma 14.2 tangent vectors at  $p = (y, g(y))$  belong to the image of  $D\phi_y$ . In other words, are linear combinations of the column vectors in the matrix

28

$$J\phi(y) = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \\ \partial_1 g(y) & \partial_2 g(y)(x) & \cdots & \partial_{n-1} g(y) \end{bmatrix}.$$

*Proof.* But

$$(-\partial_1 g(y), -\partial_2 g(y), \dots, -\partial_{n-1} g(y), 1)^T$$

is has zero scalar product with each of this column vectors, so it is perpendicular.

Also, it is clear that  $\nu$  has norm one, because we are dividing by the square root, that is precisely the norm of the vector in the numerator.

Finally to show that  $\nu$  is points outwards notice that, for  $h > 0$  small the point

$$(y - h\nabla g(y), g(y) + h)$$

does not belong to  $\Omega$  because:

$$g(y - h\nabla g(y)) = g(y) + Dg_y(-\nabla g(y)h) + o(h) = g(y) - |\nabla g(y)|^2 h + o(h) < g(y) + h.$$

□

#### DEFINITION 14.6: EXTERIOR UNIT NORMAL MAP

If  $\Omega \subset \mathbb{R}^n$  is a bounded  $C^1$  domain, the continuous map  $\nu : \partial\Omega \rightarrow \mathbb{S}^{n-1}$  (where  $\mathbb{S}^{n-1}$  is the unit  $(n-1)$ -sphere, i.e.,  $\partial B_1(0) \subset \mathbb{R}^n$ ) assigning to each point  $p \in \partial\Omega$  the exterior normal vector at  $p$  is called **exterior unit normal map**.

28

14.7. — Notice that the fact that the exterior and interior unit normal are uniquely defined and are continuous follows from Proposition 14.5, since we showed that every bounded  $C^1$  domain is locally graphical around each of its boundary points.

### 14.1.2 Integration by parts formula: special cases

An important special case of the integration by parts formula is the following:

#### PROPOSITION 14.8: SPECIAL LOCAL INTEGRATION BY PARTS FOR GRAPHS

Let  $\Omega \subset \mathbb{R}^n$  be a bounded  $C^1$  domain. Suppose that  $\Omega \cap B_r(p_o)$  is graphical: that is, there exists  $R : \mathbb{R}^n \rightarrow \mathbb{R}^n$  orthogonal  $V \subset \mathbb{R}^{n-1}$  open, and  $g \in C^1(V, \mathbb{R})$  such that:

$$\Omega \cap B_r(p_o) = R \{(y, x_n) \in \mathbb{R}^{n-1} \times \mathbb{R} \mid y \in V, x_n < g(y)\} \cap B_r(p_o).$$

Then for all  $f \in C^1(\overline{\Omega \cap B_r(p_o)})$  with  $\text{spt}(f) \subset B_r(p_o)$  the following formula holds true for  $w := Re_n$

$$\int_{\Omega \cap B_r(p_o)} \partial_w f \, d\text{vol}_n = \int_{\partial\Omega \cap B_r(p_o)} f(p) (\nu(p) \cdot w) \, d\text{vol}_{n-1}(p),$$

where  $\nu : \partial\Omega \rightarrow \mathbb{S}^{n-1}$  is the exterior unit normal map.

In the proof of Proposition 14.8 we will need the following variant of Fubini's theorem:

## LEMMA 14.9: YET ANOTHER VARIANT OF FUBINI

Suppose that  $\Omega \subset \mathbb{R}^n$  satisfies

$$\Omega \cap B_r(p_o) = \{(y, x_n) \in \mathbb{R}^{n-1} \times \mathbb{R} \mid y \in V, x_n < g(y)\} \cap B_r(p_o).$$

Then for all  $h \in C^0(\overline{\Omega \cap B_r(p_o)})$  with  $\text{spt}(f) \subset B_r(p_o)$  we have

$$\int_{\Omega \cap B_r(p_o)} h(x) dx = \int_V \int_{-\infty}^{g(y)} h(y, x_n) dx_n dy,$$

where to define the integral on the right hand side we are considering  $h$  to be extended by zero in the complement of its support.

*Proof.* (extra material) Notice first that since the support of  $h$  is contained in the open ball  $B_r(p_o)$  we  $f$  be zero and obtain a continuous function, still denoted  $h$ , defined in all the ‘epigraph’ of  $g$ , namely the set

$$\{(y, x_n) \in \mathbb{R}^n \mid x_n \leq g(y)\}.$$

Now let us fix  $C \in \mathbb{N}$  large enough so that

$$\text{spt}(h) \subset A := \{(y, x_n) \in \mathbb{R}^n \mid y \in V, -C \leq x_n \leq g(y)\}.$$

By definition of  $\int_A h$ , for each  $\epsilon > 0$  there are dyadic step functions  $h_{\text{low}} : \mathbb{R}^n \rightarrow \mathbb{R}$  and  $h_{\text{up}} : \mathbb{R}^n \rightarrow \mathbb{R}$  such that

$$h_{\text{low}} \leq h \mathbf{1}_A \leq h_{\text{up}} \tag{14.1}$$

and

$$\int h_{\text{low}} \leq \int_A h \leq \int h_{\text{up}} \leq \int h_{\text{low}} + \epsilon.$$

Consider  $H_{\text{low}}, H_{\text{up}} : \mathbb{R}^{n-1} \rightarrow \mathbb{R}$ , defined as

$$H_{\text{low}}(y) := \int_{\mathbb{R}} h_{\text{low}}(y, x_n) dx_n \quad \text{and} \quad H_{\text{up}}(y) := \int_{\mathbb{R}} h_{\text{up}}(y, x_n) dx_n$$

are notice that they are step functions in  $\mathbb{R}^{n-1}$  (they are constant in every dyadic cube of the same pixel size as the one of  $h_{\text{down}}$  and  $h_{\text{up}}$ ).

Also, from integrating (14.1) we obtain:

$$H_{\text{low}} \leq H \mathbf{1}_V \leq H_{\text{up}}$$

for  $H : V \rightarrow \mathbb{R}$  defined as

$$H(y) := \int_{-C}^{g(y)} h(y, x_n) dx_n.$$

But notice also that, since  $h_{\text{low}}$  and  $h_{\text{up}}$  we have dyadic step functions

$$\int_{\mathbb{R}^{n-1}} H_{\text{low}} = \int_{\mathbb{R}^n} h_{\text{low}} \quad \text{and} \quad \int_{\mathbb{R}^{n-1}} H_{\text{up}} = \int_{\mathbb{R}^n} h_{\text{up}}$$

(because the integrals reduce to discrete sums).

Therefore we have shown that

$$\int_{\mathbb{R}^n} h_{\text{low}} = \int_{\mathbb{R}^{n-1}} H_{\text{low}} \leq \int_V H(y) dy \leq \int_{\mathbb{R}^{n-1}} H_{\text{up}} = \int_{\mathbb{R}^n} h_{\text{up}} \leq \int_{\mathbb{R}^n} h_{\text{low}} + \epsilon.$$

But then sending  $\epsilon$  to zero we prove the lemma.  $\square$

We can now give the

*Proof of Proposition 14.8.* Notice first that it is enough to do the proof in the ‘special case’  $R = \text{Id}$ . Indeed, the general case simply follows from the ‘special’ one applied in ‘a rotated Euclidean coordinate frame’.

More precisely we can apply the special case to  $\tilde{\Omega} = R^T(\Omega)$ ,  $\tilde{p}_o = R^T(p_o)$ ,  $\tilde{f} = f \circ R$ , observing that :

- $\tilde{\Omega} \cap B_r(\tilde{p}_o) = R^T(\Omega \cap B_r(p_o))$ .
- $\partial_n \tilde{f}(q) = D(f \circ R)_q(e_n) = (Df_{R(q)} \circ R)(e_n) = Df_{R(q)}(w) = \partial_w f(R(q))$ , holds for all  $q \in R^T(\Omega \cap B_r(p_o))$ .
- If  $\nu : \partial\Omega \rightarrow \mathbb{S}^{n-1}$  is the exterior normal vector for the domain  $\Omega$ , then  $\tilde{\nu} = R^T \circ \nu \circ R$  is the exterior unit normal for  $\tilde{\Omega}$ . Hence,  $\tilde{\nu}(q) \cdot e_n = R^T \nu(R(q)) \cdot e_n = \nu(R(q)) \cdot R e_n = \nu(R(q)) \cdot w$ .

Hence since the integrals are invariant under Euclidean isometries (by the Change of Variables Formula and Lemma 13.68), and  $R$  is an isometry, we obtain:

$$\int_{\Omega \cap B_r(p_o)} \partial_w f(p) d\text{vol}_n(q) = \int_{\tilde{\Omega} \cap B_r(\tilde{p}_o)} \partial_n \tilde{f}(q) d\text{vol}_n(q)$$

and

$$\int_{\partial\Omega \cap B_r(p_o)} f(p)(\nu(p) \cdot w) d\text{vol}_{n-1}(p) = \int_{\partial\tilde{\Omega} \cap B_r(\tilde{p}_o)} \tilde{f}(q)(\tilde{\nu}(q)) d\text{vol}_{n-1}(q)$$

So from now on let us assume without loss of generality that  $R = \text{Id}$ . Applying Lemma 14.9 to the function  $h = \partial_n f$  (conveniently extended by zero to the whole hypograph of  $g$ ,  $\{(x_n, y) \in \mathbb{R}^n \mid x_n \leq g(y), y \in V\}$ ), we obtain:

$$\int_{\Omega \cap B_r(p_o)} \partial_n f(x) dx = \int_V \int_{-\infty}^{g(y)} \partial_n f(y, x_n) dx_n dy = \int_V f(y, g(y)) dy,$$

where for the last equality we used the Fundamental Theorem of Calculus and that  $f(y, -C) = 0$  for  $C$  large.

Hence, it only remains to show that

$$\int_V f(y, g(y)) dy = \int_{\partial\Omega \cap B_r(p_o)} f(p) \nu_n(p) d\text{vol}_{n-1}(p),$$

where  $\nu_n = \nu \cdot e_n$  is the  $n$ -th component of the exterior normal vector.

But this follows from combining Lemma 13.69 and Proposition 14.5: indeed, using by Lemma 13.69 we have submanifold we have:

$$\int_{\partial\Omega \cap B_r(p_o)} f(p) \nu_n(p) d\text{vol}_{n-1}(p) = \int_V f(y, g(y)) \nu_n(y, g(y)) \sqrt{1 + |\nabla g(y)|^2} dy$$

but by Proposition 14.5

$$\nu_n(y, g(y)) \sqrt{1 + |\nabla g(y)|^2} = 1,$$

so the proposition follows □

**COROLLARY 14.10: LOCAL INTEGRATION BY PARTS FOR GRAPHS**

Let  $\Omega \subset \mathbb{R}^n$  be a bounded  $C^1$  domain. Suppose that  $\Omega \cap B_r(p_o)$  is graphical: that is, there exists  $R : \mathbb{R}^n \rightarrow \mathbb{R}^n$  orthogonal  $V \subset \mathbb{R}^{n-1}$  open, and  $g \in C^1(V, \mathbb{R})$  such that:

$$\Omega \cap B_r(p_o) = R \{(y, x_n) \in \mathbb{R}^{n-1} \times \mathbb{R} \mid y \in V, x_n < g(y)\} \cap B_r(p_o).$$

Then for all  $f \in C^1(\overline{\Omega \cap B_r(p_o)})$  with  $\text{spt}(f) \subset B_r(p_o)$  and  $w \in \mathbb{R}^n$  the following formula holds true:

$$\int_{\Omega \cap B_r(p_o)} \partial_w f d\text{vol}_n = \int_{\partial\Omega \cap B_r(p_o)} f(p) (\nu(p) \cdot w) d\text{vol}_{n-1}(p) \quad (14.2)$$

where  $\nu : \partial\Omega \rightarrow \mathbb{S}^{n-1}$  is the exterior unit normal map.

*Proof.* The difference between Proposition 14.8 and Corollary 14.10 is that can take  $w$  to be any vector in  $\mathbb{R}^n$ , and not just  $Re_n$ .

However, we can deduce this from Proposition 14.8 with a ‘smart trick’.

Pbserve that the formula (14.2) is linear in  $w$ : if holds for  $w = w_1$  and for  $w = w_2$ , then it also holds for  $w = t_1 w_1 + t_2 w_2$  where  $t_1, t_2$  are real numbers. Hence, to establish (14.2) for all  $w$  it is enough to prove the formula for  $n$  linearly independent vectors.

Now fix  $s \in (0, r)$  such that  $\text{spt}(f) \subset B_s(p_o)$  and let us show that there exists  $\epsilon > 0$  such that for any orthogonal linear transformation  $O : \mathbb{R}^n \rightarrow \mathbb{R}^n$  satisfying  $\|O - \text{Id}\|_2 < \epsilon$  (i.e.,  $O$  is sufficiently close to the identity, where we measure the difference with the Hilbert-Schmidt norm) we have

$$\Omega \cap B_s(p_o) = OR(\{(y, x_n) \in \mathbb{R}^n \mid y \in W_O, x_n < g_O(y)\}) \cap B_s(p_o).$$

where  $W_O \subset \mathbb{R}^{n-1}$  is a suitable open set and  $g_O : W_O \rightarrow \mathbb{R}$  is some  $C^1$  map. Indeed, the key observation is that  $\partial\Omega \cap B_s(p_o)$  can be expressed also as a graph in slightly Euclidean

coordinate frame. Indeed, let  $\pi_{n-1} : \mathbb{R}^n \rightarrow \mathbb{R}^{n-1}$  be the projection onto the first  $n - 1$  components, i.e.  $\pi_{n-1} : (y, x_n) \mapsto y$ .

Since  $\phi(y) = R(y, g(y))$  is the parametrization of  $\partial\Omega \cap B_r(p_o)$  then  $\pi_{n-1} \circ R^T$  is an inverse of  $\phi$  in the sense that

$$(\pi_{n-1} \circ R^T \circ \phi)(y) = y \quad \text{for all } y \in V.$$

Now, let  $W := \phi^{-1}(B_s(p_o))$  and notice that, since  $s < r$  we have  $\overline{W} \subset V$ . Hence, for  $O$  sufficiently close to the identity, the map

$$\psi_O = (\pi_{n-1} \circ O^T \circ R^T \circ \phi) : W \rightarrow \mathbb{R}^{n-1}$$

is a diffeomorphism from  $W$  onto its image  $W_O := \psi_O(W)$ . To check this we only need to compute  $D\psi_O$  at a given  $y \in \overline{W}$ : by the chain rule we have

$$D(\psi_O)_y = \pi_{n-1} \circ O^T \circ R^T \circ D\phi_y$$

Now since for  $O = \text{Id}$  we have  $\psi_O = \text{Id}$  and thus  $D(\psi_O)_y$ , by continuity there is  $\epsilon > 0$  such that  $D(\psi_O)_y$  is invertible, provided  $\|O - \text{Id}\| < \epsilon$  with  $\epsilon > 0$  small enough. The fact that we can take  $\epsilon > 0$  that works for all  $y \in \overline{W}$  follows from the compactness of this set.

Thanks to this observation, and since

$$\int_{\Omega \cap B_r(p_o)} \partial_w f \, d\text{vol}_n = \int_{\Omega \cap B_s(p_o)} \partial_w f$$

and

$$\int_{\partial\Omega \cap B_r(p_o)} f(\nu \cdot w) \, d\text{vol}_{n-1} = \int_{\partial\Omega \cap B_s(p_o)} f(\nu \cdot w) \, d\text{vol}_{n-1}$$

Proposition 14.8 gives the validity of (??) for all  $w = ORe_n$ , whenever  $O$  is orthogonal with  $\|O - \text{Id}\| < \epsilon$ . But such set of  $w$  spans all of  $\mathbb{R}^n$  and hence, by linearity, 14.8 holds for all  $w \in \mathbb{R}^n$ . This proves the corollary.  $\square$

### 14.1.3 The integration by parts formula and the divergence theorem

We can now give the following important result:

#### THEOREM 14.11: INTEGRATION BY PARTS

Let  $\Omega \subset \mathbb{R}^n$  be a bounded  $C^1$  domain and  $f \in C^1(\overline{\Omega})$ . Then, the following formula holds true for every vector  $w \in \mathbb{R}^n$ :

$$\int_{\Omega} \partial_w f \, d\text{vol}_n = \int_{\partial\Omega} f(p)(\nu(p) \cdot w) \, d\text{vol}_{n-1}(p),$$

where  $\nu : \partial\Omega \rightarrow \mathbb{S}^{n-1}$  is the exterior unit normal map.



Before giving the proof we need a simple preliminary result

**LEMMA 14.12: INTEGRATION BY PARTS FOR  $C^1$  AND COMPACTLY SUPPORTED**

Assume that  $h \in C^1(\mathbb{R}^n)$  has compact support and let  $(-C, C)^n$  be a box containing the support of  $h$ .

Then for all  $w \in \mathbb{R}^n$  we have

$$\int_{\mathbb{R}^n} \partial_w h = \int_{(-C, C)^n} \partial_w h = 0$$

*Proof.* Notice first that the support of  $\partial_w h$  must be contained in the support of  $h$  (as if  $h = 0$  in some open set then  $\partial_w = 0$  in the same open set). Also, using for example the change of variable formula we easily obtain:

$$\int_{\mathbb{R}^n} h(x + \tau) dx = \int_{\mathbb{R}^n} h(x) dx.$$

Now for if we pick  $C$  sufficiently large so that the support of  $h$  is contained in  $(-C, C)^n$ , we will have, for  $|\tau|$  sufficiently small so that the support of  $x \mapsto h(x + \tau)$  is still contained in  $(-C, C)^n$ ,

$$\int_{(-C, C)^n} h(x + \tau) dx = \int_{(-C, C)^n} h(x) dx.$$

29 But then taking  $\tau = \delta w$  with  $\delta > 0$  small we obtain

$$0 = \int_{(-C, C)^n} \frac{h(x + \delta w) - h(x)}{\delta} dx \rightarrow \int_{(-C, C)^n} \partial_w h(x) dx,$$

as  $\delta \rightarrow 0$ . □

*Proof.* Proof of Theorem 14.11 Let  $\phi_\ell : V_\ell \rightarrow B_{r_\ell}(p_\ell)$ ,  $1 \leq \ell \leq N$ , be a graphical cover of  $\partial\Omega$  and  $\eta_\ell : \mathbb{R}^n \rightarrow [0, \infty)$  be a partition of unity on  $M$  subordinated to  $\{B_{r_\ell}(p_\ell)\}_{1 \leq \ell \leq N}$ .

Recall that by definition  $V_\ell \subset \mathbb{R}^n$  are open and we have

$$M \cap B_{r_\ell}(p_\ell) = \phi_\ell(V_\ell)$$

and

$$\phi_\ell(y) = R_\ell \left( (y, g(y))^T \right),$$

where  $R_\ell; \mathbb{R}^n \rightarrow \mathbb{R}^n$  are orthogonal transformations and  $g_\ell : V_\ell \rightarrow \mathbb{R}$  are  $C^1$  maps.

Also by definition we have:

$$\int_{\partial\Omega} h \, d\text{vol}_{n-1} = \sum_{\ell=1}^N \int_{M \cap B_\ell(x_\ell)} h \eta_\ell \, d\text{vol}_{n-1},$$

for any continuous function  $h : \partial\Omega \rightarrow \mathbb{R}$ .

Define now  $\tilde{\eta} : \mathbb{R}^n \rightarrow \mathbb{R}$  as

$$\tilde{\eta}(x) = \begin{cases} 1 - \sum_{\ell=1}^N \eta_\ell(x) & \text{for } x \in \Omega \\ 0 & \text{for } x \in \mathbb{R}^n \setminus \Omega. \end{cases}$$

Notice that since  $\eta_\ell$  is a partition of unity of  $\partial\Omega$  we have  $\sum_{\ell=1}^N \eta_\ell(x)$  in some open neighborhood of  $\partial\Omega$  (in other words, in some open set containing  $\partial\Omega$ ).

Then, it follows that  $\tilde{\eta}$  is of class  $C^\infty$  in all of  $\mathbb{R}^n$ , that the support of  $\tilde{\eta}$  is contained in  $\Omega$  and that  $\sum_{\ell=1}^N \eta_\ell + \tilde{\eta} = 1$  in  $\Omega$ . Hence,

$$\begin{aligned} \int_{\Omega} \partial_w f \, d\text{vol}_n &= \sum_{\ell=1}^N \int_{\Omega} \partial_w (f \eta_\ell) \, d\text{vol}_n + \int_{\Omega} \partial_w (f \tilde{\eta}) \, d\text{vol}_n \\ &= \sum_{\ell=1}^N \int_{\Omega \cap B_{r_\ell}(p_\ell)} \partial_w (f \eta_\ell) \, d\text{vol}_n + \int_{\Omega} \partial_w (f \tilde{\eta}) \, d\text{vol}_n. \end{aligned}$$

Now, on the one hand, since the support of  $(f \tilde{\eta})$  a compact subset of  $\Omega$  we can extend this function by zero outside of  $\Omega$  obtaining a compactly supported  $C^1(\mathbb{R}^n)$  function. But then Lemma 14.12 gives:

$$\int_{\Omega} \partial_w (f \tilde{\eta}) \, d\text{vol}_n$$

On the other hand, for each  $\ell$  we can apply Corollary 14.10 and obtain

$$\int_{\Omega \cap B_{r_\ell}(p_\ell)} \partial_w (f \eta_\ell) \, d\text{vol}_n = \int_{\partial\Omega \cap B_{r_\ell}(p_\ell)} f \eta_\ell (\nu \cdot w) \, d\text{vol}_{n-1}.$$

Hence, after summing in  $\ell$  we obtain the desired formula.  $\square$

We can now give the divergence theorem as a corollary

#### DEFINITION 14.13: VECTOR FIELD IN $\mathbb{R}^n$ , DIVERGENCE

Let  $U \subset \mathbb{R}^n$  be open. A  $C^k$  **vector field** on  $U$  is a function  $F : U \rightarrow \mathbb{R}^n$  of class  $C^k$ . If we write

$$F(x) = (F_1(x), \dots, F_n(x)),$$

then functions  $\{F_i\}$ , from  $U$  to  $\mathbb{R}$ , are called the **components** of the vector field.

Given a  $C^1$  vector field  $F$  the **divergence** of  $F$  is defined as the map  $\text{div}(F) : U \rightarrow \mathbb{R}$  given by

$$\text{div}F(x) := \sum_{i=1}^n \partial_i F_i(x).$$

We now prove an important corollary of the integration by parts formula:

**THEOREM 14.14: DIVERGENCE THEOREM**

Let  $\Omega$  be a bounded  $C^1$  domain and  $F : \bar{\Omega} \rightarrow \mathbb{R}^n$  some  $C^1$  vector field.

Then,

$$\int_{\Omega} \operatorname{div} F \, d\operatorname{vol}_n = \int_{\partial\Omega} F \cdot \nu \, d\operatorname{vol}_{n-1}$$

*Proof.* Applying Theorem 14.11 to  $w = e_i$  and  $f = F_i$  we obtain:

$$\int_{\Omega} \partial_i F_i \, d\operatorname{vol}_n = \int_{\partial\Omega} F_i(p) (\nu(p) \cdot e_i) \, d\operatorname{vol}_{n-1}(p).$$

Hence, summing over  $i = 1, \dots, n$  and noticing  $\sum_i F_i(p) (\nu(p) \cdot e_i) = F(p) \cdot \nu(p)$  we conclude.  $\square$

**14.1.4 Examples and application of the divergence theorem**

**EXAMPLE 14.15.** — Consider a compact region  $B$  in  $\mathbb{R}^2$  and the smooth vector field  $F : \mathbb{R}^2 \rightarrow \mathbb{R}^2$  given by  $F(x) = x$ . Then,  $\operatorname{div}(F) = 2$ , and the divergence theorem yields

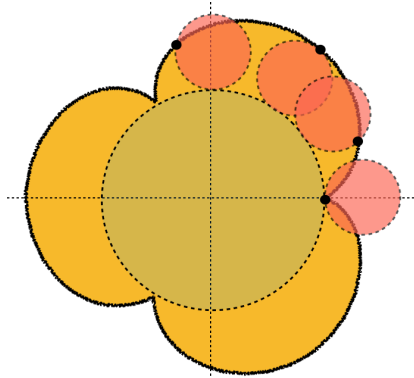
$$2 \operatorname{vol}(B) = \int_B \operatorname{div}(F) \, dx = \int_{\partial B} F \cdot \nu \, dL.$$

29

if we can apply it to the region  $B$ . So, for example, if  $\partial B$  is parametrized by a suitable curve  $\gamma : [a, b] \rightarrow \partial B$ , then

$$\operatorname{vol}(B) = \frac{1}{2} \int_a^b \langle \gamma(t), \nu \circ \gamma(t) \rangle dt = \frac{1}{2} \int_a^b \gamma_1(t) \gamma_2'(t) - \gamma_2(t) \gamma_1'(t) dt.$$

As a concrete example, we calculate the area of the region  $B$  inside the curve obtained by unrolling a circle with radius  $\frac{1}{m}$  on a circle with radius 1 for an integer  $m \geq 1$ , while tracking a point on the moving circle. This curve is called an **epicycloid**, and in the special case  $m = 1$ , it is also called a **cardioid**. The area enclosed by the epicycloid, denoted by  $B$ , is not smoothly bounded, but as explained in ??, the divergence theorem holds for the area  $B$ .



We represent elements of  $\mathbb{R}^2$  as column vectors. The epicycloid is described by the positively oriented closed path  $\gamma : [0, 2\pi] \rightarrow \mathbb{R}^2$

$$\gamma(t) = \frac{m+1}{m} \begin{pmatrix} \cos(t) \\ \sin(t) \end{pmatrix} + \frac{1}{m} \begin{pmatrix} \cos((m+1)t) \\ \sin((m+1)t) \end{pmatrix}$$

The outward normal to  $\gamma$  is given by

$$n_\gamma(t) = \frac{m+1}{m} \begin{pmatrix} \cos(t) \\ \sin(t) \end{pmatrix} + \frac{m+1}{m} \begin{pmatrix} \cos((m+1)t) \\ \sin((m+1)t) \end{pmatrix}$$

The scalar product  $\langle \gamma(t), n_\gamma(t) \rangle$  is

$$\frac{(m+1)^2}{m^2} + \frac{m+1}{m^2} + a \cos(t) \cos((m+1)t) + a \sin(t) \sin((m+1)t)$$

for a constant  $a$  that we do not worry about. We obtain

$$\text{vol}(B) = \frac{1}{2} \int_0^{2\pi} \langle \gamma(t), n_\gamma(t) \rangle dt = \pi \frac{(m+1)(m+2)}{m^2}$$

since the integral over  $[0, 2\pi]$  of  $\cos(t) \cos((m+1)t)$  and also  $\sin(t) \sin((m+1)t)$  is zero.

29

### 14.1.5 Rotation and the Green's Theorem

**DEFINITION 14.16.** — Let  $F : U \rightarrow \mathbb{R}^2$  be a continuously differentiable vector field on an open set  $U \subseteq \mathbb{R}^2$ . The **vorticity** or **rotation** of  $F$  is the real-valued function defined by

$$\text{rot}(F)(x) = \partial_1 F_2(x) - \partial_2 F_1(x)$$

on  $U$ .

**THEOREM 14.17 (Green's Theorem).** — Let  $F : U \rightarrow \mathbb{R}^2$  be a continuously differentiable vector field on an open set  $U \subseteq \mathbb{R}^2$ . Then, for any smoothly bounded, compact region  $B \subseteq U$ ,

$$\int_B \text{rot}(F) dx = \int_{\partial B} F \cdot \tau dL$$

, where  $\tau = \nu^\perp = (-\nu_2, \nu_1)$

*Proof.* We define the vector field  $F^\perp : U \rightarrow \mathbb{R}^2$  by  $F^\perp(x) = (-F_2(x), F_1(x))$ . Then,

$$\text{div}(F^\perp) = \partial_1 F_1^\perp + \partial_2 F_2^\perp = -\partial_1 F_2 + \partial_2 F_1 = \text{rot}(F)$$

and for a path  $\gamma$  in  $U$ ,

$$\langle F^\perp(\gamma(t)), \nu \circ \gamma(t) \rangle = \langle (F_2(\gamma(t)), -F_1(\gamma(t))), (\gamma'_2(t), -\gamma'_1(t)) \rangle = \langle F(\gamma(t)), \gamma'(t) \rangle.$$

Applying the divergence theorem 14.14 to  $F^\perp$ , we obtain Green's theorem for  $F$ .  $\square$

**Applet 14.18 (Divergence and Rotation).** *This applet illustrates the concepts of irrotational and divergence-free, as well as the theorems of this section.*

14.19. — As an application of the divergence theorem in the plane, we discuss in the following exercises the **Jordan Curve Theorem**. The question that this theorem answers is essentially the following. Suppose you have drawn a *complicated* closed curve without self-intersections in the plane. Does this curve then divide the plane into an *inside* and an *outside*? How do you even define whether a point is in the inside or the outside? Although this is intuitively clear, proving it is surprisingly difficult. This problem was recognized by Camille Jordan (1838-1922). Jordan's original proof of the theorem (around 1882) uses a kind of polygonal approximation to the curve. It is problematic in several places and was doubted for several decades but is now considered essentially correct. The problem can also be formulated in higher dimensions. For example, one can ask whether a complicatedly deformed sphere in  $\mathbb{R}^3$  divides space  $\mathbb{R}^3$  into a well-defined inside and outside. A positive answer to the general problem is provided by the **Jordan-Brouwer Separation Theorem**. A modern proof of this, which works in any dimension and does not require analysis, can be found in [Hat02] Proposition 2B.1, or in [Rot88], Theorem 6.35. Here, we formulate and prove the theorem in the case of smooth curves in  $\mathbb{R}^2$ .

29

**THEOREM 14.20 (Jordan Curve Theorem).** — *Let  $\gamma : [0, 1] \rightarrow \mathbb{R}^2$  be a smooth, regular, simple, closed path. Then, one can uniquely write the complement of  $\Gamma = \gamma([0, 1])$  as the disjoint union*

$$\mathbb{R}^2 \setminus \Gamma = \text{Inn}(\Gamma) \cup \text{Out}(\Gamma), \quad \text{Inn}(\Gamma) \cap \text{Out}(\Gamma) = \emptyset$$

where the interior  $\text{Inn}(\Gamma)$  is an open, bounded, connected subset and the exterior  $\text{Out}(\Gamma)$  is an open, unbounded, connected subset. Furthermore,  $\partial \text{Inn}(\gamma) = \partial \text{Out}(\gamma) = \Gamma$ .

14.21. — A path  $\gamma$  as in the theorem  $\gamma : [0, 1] \rightarrow \mathbb{R}^2$  is injective and smooth except for the equation  $\gamma(0) = \gamma(1)$ , and  $\gamma'(t) \neq 0$  for all  $t \in [0, 1]$ . In the following, we additionally assume that  $\gamma'(0) = \gamma'(1)$ , meaning the image  $\Gamma = \gamma([0, 1])$  has no kink at the point  $\gamma(0) = \gamma(1)$ . Thus,  $\gamma$  could be extended to a periodic, continuously differentiable function on  $\mathbb{R}$ . We define, for  $u \in \mathbb{R}^2 \setminus \Gamma$ , the **rotation number** of  $\gamma$  around  $u$  as

$$I_\gamma(u) = \frac{1}{2\pi} \int_0^1 \frac{\langle \gamma(t) - u, \nu \circ \gamma(t) \rangle}{|\gamma(t) - u|^2} dt.$$

EXERCISE 14.22. — Consider the closed curve  $\gamma : [0, 1] \rightarrow \mathbb{R}^2$  for  $r > 0$

$$\gamma(t) = (r \cos(2\pi t), r \sin(2\pi t))$$

parametrizing the boundary of the circular disk  $B(0, r)$ . Realize  $I_\gamma(u)$  for  $u \notin \partial B(0, r)$  as the flux integral of a vector field over  $\partial B(0, r)$ , and show that for all  $u \notin \partial B(0, r)$ , the rotation number is given by

$$I_\gamma(u) = \begin{cases} 1 & \text{if } u \in B(0, r) \\ 0 & \text{if } u \notin B(0, r) \end{cases}$$

For circles, the rotation number is thus able to decide whether a point is inside or outside the circle. Does it work for a rectangle? For an ellipse?

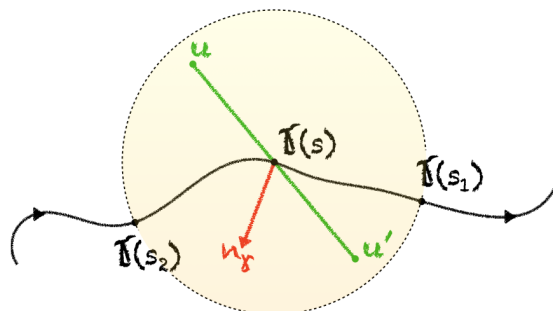
EXERCISE 14.23. — Let  $\gamma : [a, b] \rightarrow \mathbb{R}^2$  be a smooth, regular, simple, closed path. Show that  $I_\gamma$  is locally constant.

14.24. — Exercise 14.23 shows, in particular, that  $I_\gamma$  vanishes outside a sufficiently large ball. In fact,  $I_\gamma(u)$  approaches zero as  $|u| \rightarrow \infty$ , since in the definition of  $I_\gamma(u)$ , the norm of  $u$  appears to the power of 2 in the denominator and to the power of 1 in the numerator. If  $B(0, r)$  is a ball containing  $\Gamma$ , then  $I_\gamma$  must be constant on  $\mathbb{R}^2 \setminus B(0, r)$  due to Exercise 14.23, and at the same time, it must go to zero for increasing radii, proving the claim. The essential idea to prove Jordan's Curve Theorem is now to examine what happens to the rotation number  $I_\gamma(u)$  when  $u$  crosses the curve  $\Gamma$ . Using the divergence theorem, one can show that the rotation number changes by  $\pm 1$  when crossing  $\Gamma$ , depending on the direction in which the normal vector of  $\gamma$  points.

EXERCISE 14.25. — Let  $u, u' \in \mathbb{R}^2 \setminus \gamma([a, b])$ , such that the line segment from  $u$  to  $u'$  intersects the trace of  $\gamma$  at exactly one point  $\gamma(s)$ . Show that

$$I_\gamma(u') - I_\gamma(u) = \text{sgn}(\langle u - u', \nu(\gamma(s)) \rangle)$$

Note that the trace of  $\gamma$  is intersected at exactly one point. Thus, the statement can be reduced to the following illustration:



Now, replace the path segment of  $\gamma$  between the times  $s_1, s_2$  chosen sufficiently close to  $s$  with either of the circular segments as shown in the image.

EXERCISE 14.26. — Show that there is at least one point  $u_0 \in \mathbb{R}^2 \setminus \Gamma$  with  $I_\gamma(u_0) = 1$  or  $I_\gamma(u_0) = -1$ .

29

EXERCISE 14.27. — Show that the sets

$$\text{Inn}(\gamma) = \{u \in \mathbb{R}^2 \setminus \Gamma \mid I_\gamma(u) \neq 0\} \quad \text{and} \quad \text{Out}(\gamma) = \{u \in \mathbb{R}^2 \setminus \Gamma \mid I_\gamma(u) = 0\}$$

are open and path-connected.

EXERCISE 14.28. — Prove Jordan's Curve Theorem 14.20.

## 14.2 Line integrals and the Poincaré Lemma

### 14.2.1 The work of a Vector Field along a path

14.29. — We often visualize vector fields by drawing the vector  $F(x)$  with the point  $x \in U$ . In physics, vector fields are often force fields or indicate the flow velocity of a medium.

#### DEFINITION 14.30: WORK OF A FIELD ALONG A PATH (LINE INTEGRAL)

Let  $U \subset \mathbb{R}^n$  be an open subset, and let  $F : U \rightarrow \mathbb{R}^n$  be a continuous vector field and  $\gamma : [a, b] \rightarrow U$  a  $C^1$  map (also called 'path'). We define the **work of  $F$  along  $\gamma$**  as

$$\int_\gamma F := \int_a^b F(\gamma(t)) \cdot \gamma'(t) dt.$$

If  $\gamma$  is piecewise continuously differentiable with respect to a partition  $a = t_0 < t_1 < \dots < t_N = b$ , the integral is interpreted as the sum of integrals over intervals  $[t_{k-1}, t_k]$ .

In Physics literature, one often finds notations like:

$$\int_\gamma F = \int_\gamma \vec{F} \cdot d\vec{r}.$$

where  $\vec{r}$  is the position vector.

30

## LEMMA 14.31: REPARAMETRIZATION INVARIANCE OF LINE INTEGRALS

Let  $U \subset \mathbb{R}^n$  be an open subset,  $f : U \rightarrow \mathbb{R}^n$  be a continuous vector field, let  $\gamma : [a, b] \rightarrow \mathbb{R}^d$  be a continuously differentiable path and let  $\psi : [0, 1] \rightarrow [a, b]$  be a  $C^1$  function such that  $\psi(0) = a, \psi(1) = b$ . Then

$$\int_{\gamma} F = \int_{\gamma \circ \psi} F.$$

*Proof.* This is a consequence of the change of variable formula (the one for one variable seen in Analysis I) and the chain rule:

$$\begin{aligned} \int_{\gamma \circ \psi} F &= \int_0^1 F(\gamma(\psi(t))) \cdot (\gamma \circ \psi)'(t) dt = \int_0^1 F(\gamma(\psi(t))) \cdot \gamma'(\psi(t)) \psi'(t) dt \\ &= \int_a^b F(\gamma(s)) \cdot \gamma'(s) ds = \int_{\gamma} F. \end{aligned}$$

□

## DEFINITION 14.32: POTENTIAL OF A VECTOR FIELD

Let  $U \subset \mathbb{R}^n$  be open, and  $F : U \rightarrow \mathbb{R}^n$  be a continuous vector field. A continuously differentiable function  $f : U \rightarrow \mathbb{R}$  is called a **potential** for  $F$ , if  $\nabla f = F$  holds. That is to say

$$\partial_i f(x) = F_i(x) \text{ for all } x \in U, 1 \leq i \leq n.$$

If  $F$  admits a potential  $f$  in  $U$ , we say that  $F$  is **conservative in  $U$** .

In the next two propositions we show that a vector field  $F$  admits a potential if and only if the value of  $\int_{\gamma} F$  depends only on the endpoints of  $\gamma$ .

## PROPOSITION 14.33: WORK ALONG A PATH AS DIFFERENCE OF POTENTIAL

Let  $U \subset \mathbb{R}^n$  be open, and let  $F : U \rightarrow \mathbb{R}^n$  be a continuous vector field. Suppose there exists a potential  $f : U \rightarrow \mathbb{R}$  for  $F$ . Then

$$\int_{\gamma} F = f(\gamma(1)) - f(\gamma(0))$$

for any  $C^1$  path  $\gamma : [0, 1] \rightarrow U$ .

*Proof.* If  $\gamma : [0, 1] \rightarrow U$  is a continuously differentiable path, then for  $t \in [0, 1]$ ,  $F(\gamma(t)) = \text{grad } f(\gamma(t)) = Df(\gamma(t))$ , and thus

$$\int_{\gamma} F = \int_0^1 \langle F(\gamma(t)), \gamma'(t) \rangle dt = \int_0^1 Df(\gamma(t))(\gamma'(t)) dt = \int_0^1 (f \circ \gamma)'(t) dt = f(\gamma(1)) - f(\gamma(0))$$



by the chain rule. If  $\gamma$  is only piecewise continuously differentiable with respect to a partition  $0 = s_0 < s_1 < \dots < s_N = 1$ , the calculation can be applied to the subintervals  $[s_{k-1}, s_k]$ . This leads to a telescoping sum where all terms except  $f(\gamma(1)) - f(\gamma(0))$  cancel.  $\square$

**DEFINITION 14.34: PIECEWISE  $C^k$  PATH (EXTRA MATERIAL)**

Let  $U \subset \mathbb{R}^n$  be an open set and  $k \geq 1$  an integer. A map  $\gamma : [0, 1] \rightarrow U \subset \mathbb{R}^n$  is called **piecewise  $C^k$  path** if there exist finitely many points  $0 = t_0 < t_1 < t_2 < \dots < t_N = 1$  in  $[0, 1]$  such that for all  $k = 1, 2, \dots, N$ , the restriction  $\gamma|_{[t_{k-1}, t_k]}$  belongs to  $C^k([t_{k-1}, t_k], \mathbb{R}^n)$ .

**EXERCISE 14.35.** — Prove that Proposition 14.33 extend to piecewise  $C^1$  maps

**EXAMPLE 14.36.** — Consider the vector field  $F$  on  $\mathbb{R}^2$  defined by  $F(x, y) = (-y, x)$  and calculate the integral of  $F$  along different paths from  $(0, 0)$  to  $(1, 1)$ .

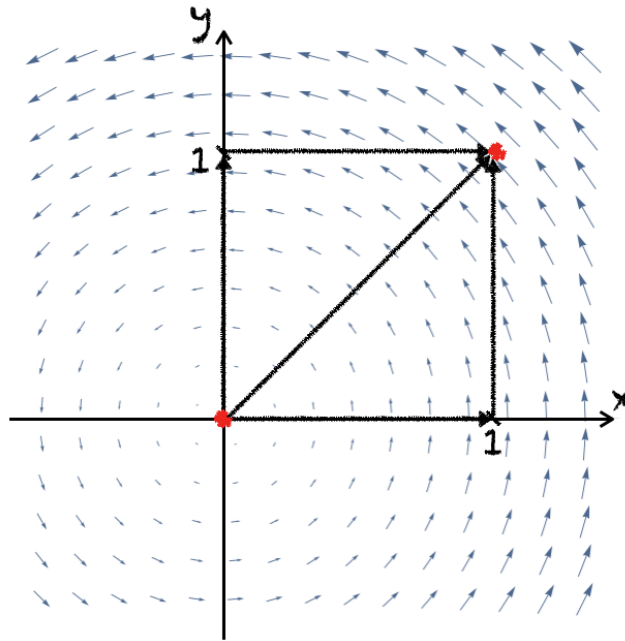


Figure 14.1: The vector field  $F$  - vector lengths are scaled by a factor of 0.1.

Let  $\gamma_0, \gamma_1$ , and  $\gamma_2 : [0, 1] \rightarrow \mathbb{R}^2$  be the paths from  $(0, 0)$  to  $(1, 1)$  given by  $\gamma_0(t) = (t, t)$  and

$$\gamma_1(t) = \begin{cases} (2t, 0) & \text{if } t \in [0, \frac{1}{2}] \\ (1, 2t - 1) & \text{if } t \in [\frac{1}{2}, 1] \end{cases} \quad \gamma_2(t) = \begin{cases} (0, 2t) & \text{if } t \in [0, \frac{1}{2}] \\ (2t - 1, 1) & \text{if } t \in [\frac{1}{2}, 1] \end{cases}$$

Then,

$$\int_{\gamma_0} F = \int_0^1 \langle F(\gamma_0(t)), \gamma_0'(t) \rangle dt = \int_0^1 \langle (-t, t), (1, 1) \rangle dt = 0$$

$$\int_{\gamma_1} F = \int_0^{\frac{1}{2}} \langle (0, 2t), (2, 0) \rangle dt + \int_{\frac{1}{2}}^1 \langle (1 - 2t, 1), (0, 2) \rangle dt = 1$$

$$\int_{\gamma_2} Fd = \int_0^{\frac{1}{2}} \langle (-2t, 0), (0, 2) \rangle dt + \int_{\frac{1}{2}}^1 \langle (-1, 2t - 1), (2, 0) \rangle dt = -1$$

We see that the work performed  $\int_{\gamma} F$  depends on the chosen path  $\gamma$ . If one moves perpendicular to the vector field, no work is done. If one moves with the vector field, positive work is done, and if one moves against the vector field, negative work is done.

From these calculations, it follows in particular that the vector field  $F$  does not possess a potential.

**EXERCISE 14.37.** — A **loop** in an open subset  $U \subset \mathbb{R}^n$  is a path  $\gamma : [0, 1] \rightarrow U$  with  $\gamma(0) = \gamma(1)$ . Show that a continuous vector field  $F : U \rightarrow \mathbb{R}^n$  is conservative if and only if, for every piecewise continuously differentiable loop  $\gamma$  in  $U$ ,

$$\int_{\gamma} F = 0$$

holds.

**LEMMA 14.38: INTEGRABILITY CONDITIONS**

Let  $U \subset \mathbb{R}^n$  be open, and let  $F$  be a  $C^1$  conservative vector field on  $U$ , with components  $F_1, \dots, F_n$ . Then we necessarily have the **integrability conditions**:

$$\partial_j F_k = \partial_k F_j$$

for all pairs  $j, k \in \{1, \dots, n\}$ .

*Proof.* If  $\text{grad } f = F$ . Then, for  $j, k \in \{1, \dots, n\}$ , we have

$$\partial_j F_k = \partial_j \partial_k f = \partial_k \partial_j f = \partial_k F_j. \quad (14.3)$$

We are allowed to commute second derivatives because  $f$  is  $C^2(U)$ , since  $\nabla f = F \in C^1(U, \mathbb{R}^n)$ , hence Schwarz Lemma applies.  $\square$

For a continuously differentiable vector field  $F$  on  $U \subset \mathbb{R}^n$  to satisfy the partial differential equations (14.3) established in Corollary 14.38 is a necessary condition for the existence of a potential for  $F$ . We call these differential equations **integrability conditions** and aim to investigate to what extent they are also sufficient for the existence of a potential for  $F$ . The following example shows that the equations 14.38 do not generally imply that  $F$  has a potential.

EXAMPLE 14.39. — Let  $U = \mathbb{R}^2 \setminus \{0\}$ , and consider the vector field  $F : U \rightarrow \mathbb{R}^2$  given by

$$F(x, y) = \left( \frac{-y}{x^2 + y^2}, \frac{x}{x^2 + y^2} \right)$$

for  $(x, y) \in U$ . A direct calculation shows

$$\partial_1 F_2(x, y) = \partial_x \left( \frac{x}{x^2 + y^2} \right) = \frac{-x^2 + y^2}{(x^2 + y^2)^2} = \partial_y \left( \frac{-y}{x^2 + y^2} \right) = \partial_2 F_1(x, y)$$

thus satisfying the integrability conditions (14.3) throughout  $U$ . However,  $F$  is not conservative. Let  $\gamma : [0, 1] \rightarrow U$  be the continuously differentiable loop defined by

30

$$\gamma(t) = (\cos(2\pi t), \sin(2\pi t))$$

which rotates once counterclockwise around the unit circle. Then,

$$\int_{\gamma} F dt = 2\pi \int_0^1 \langle (-\sin(2\pi t), \cos(2\pi t)), (-\sin(2\pi t), \cos(2\pi t)) \rangle dt = 2\pi,$$

even though  $\gamma$  is a closed path with  $\gamma(0) = \gamma(1) = (1, 0)$ .

**Applet 14.40 (Integrability Conditions).** *What different values for the path integral can you obtain when considering closed paths? Why does the value of the path integral usually not change, but sometimes does when you move the middle three points?*

### 14.2.2 The Poincaré Lemma

In order to state the main theorem of this section we need to introduce the concept of *homotopy of curves*.

DEFINITION 14.41: HOMOTOPY, SIMPLE CONNECTED SUBSET OF  $\mathbb{R}^n$

Let  $U \subset \mathbb{R}^n$  set  $\gamma_0$  and  $\gamma_1$  be two continuous maps from  $[0, 1] \rightarrow U$  with the same initial point  $x_0 = \gamma_0(0) = \gamma_1(0)$  and the same endpoint  $x_1 = \gamma_0(1) = \gamma_1(1)$ .

A **homotopy** from  $\gamma_0$  to  $\gamma_1$  within  $U$  is a continuous function  $H : [0, 1] \times [0, 1] \rightarrow U$  with the following properties:

$$H(0, t) = \gamma_0(t), \quad H(1, t) = \gamma_1(t) \quad \text{and} \quad H(s, 0) = x_0, \quad H(s, 1) = x_1$$

for all  $t \in [0, 1]$  and all  $s \in [0, 1]$ . We say  $\gamma_1$  is **homotopic to**  $\gamma_0$  if there exists a homotopy from  $\gamma_0$  to  $\gamma_1$ .

A subset  $U$  of  $\mathbb{R}^n$  is called **simply connected** if for any two given paths in  $U$  with the same end points are homotopic.

31

14.42. — Let  $H$  be a homotopy from  $\gamma_0$  to  $\gamma_1$  as in the definition. For each fixed  $s \in [0, 1]$ , the function  $\gamma_s : t \mapsto H(s, t)$  is a path from  $x_0$  to  $x_1$ . For  $s = 0$  and  $s = 1$ , we obtain the given paths  $\gamma_0$  and  $\gamma_1$ . This way, we can view the homotopy  $H$  as a parametrized family of paths depending continuously on the parameter  $s \in [0, 1]$ .

14.43. — Heuristically, a connected topological space  $X$  is called **simply connected** if every path  $\gamma$  from  $x_0$  to  $x_1$  in  $X$  can be continuously deformed into any other given path from  $x_0$  to  $x_1$ .

**THEOREM 14.44: POINCARÉ LEMMA**

Let  $U \subset \mathbb{R}^n$  be open, and let  $F : U \rightarrow \mathbb{R}^n$  be a continuously differentiable vector field that satisfies the integrability conditions

$$\partial_k F_j = \partial_j F_k \quad (14.4)$$

for all  $j, k \in \{1, \dots, n\}$ . Let  $\gamma_0 : [0, 1] \rightarrow U$  and  $\gamma_1 : [0, 1] \rightarrow U$  be (piecewise)  $C^1$  paths with the same initial point  $x_0$  and the same endpoint  $x_1$ . If  $\gamma_0$  and  $\gamma_1$  are homotopic, then

$$\int_{\gamma_0} F = \int_{\gamma_1} F.$$

Theorem 14.44 is an example of a so-called **global integration theorem** because, it has something to do with the global nature of the domain  $U$ . This is evident in the following important

**COROLLARY 14.45: CONSERVATIVE VECTOR FIELDS IN SIMPLY CONNECTED DOMAINS**

Let  $U \subset \mathbb{R}^n$  be open and simply connected. A continuously differentiable vector field on  $U$  is conservative if and only if it satisfies the integrability conditions (14.4).

*Proof.* This directly follows from Theorem 14.44, Proposition ?? and Definition ?.  $\square$

Before proving Theorem 14.44 we show a simpler proof under the additional assumption that  $U$  is convex.

**LEMMA 14.46: POINCARÉ LEMMA IN CONVEX DOMAINS**

Let  $U \subset \mathbb{R}^n$  be open and convex, and let  $F : U \rightarrow \mathbb{R}^n$  be a continuously differentiable vector field that satisfies the integrability conditions (14.4). Then  $F$  is conservative in  $U$ .

*Proof.* The necessity of the integrability conditions was already proven in Corollary 14.38. For the converse, assume without loss of generality that  $0 \in U$ . We use the path integral of

$F$  along the straight line from 0 to  $x \in U$  to define a function  $f : U \rightarrow \mathbb{R}$  by

$$f(x) := \int_0^1 \langle F(tx), x \rangle dt$$

for  $x \in U$ , and let us prove that  $f$  is a potential of  $F$ .

Indeed, fix  $j \in \{1, \dots, n\}$  and consider, as a preparation for the computation of  $\partial_j f$ , for  $h \in \mathbb{R}^n$

$$\partial_h F_j = \sum_{k=1}^n h_k \partial_k F_j = \sum_{k=1}^n h_k \partial_j F_k \quad (14.5)$$

by the assumed integrability conditions. According to the Theorem ??,  $\partial_j f$  exists, and for  $x \in U$  it holds

$$\partial_j f(x) = \partial_j \int_0^1 \left( \sum_{k=1}^n F_k(tx) x_k \right) dt = \int_0^1 \left( \sum_{k=1}^n (\partial_j F_k)(tx) t x_k + F_j(tx) \right) dt, \quad (14.6)$$

since only the term with  $k = j$  requires the product rule, and the partial derivative of  $x \mapsto F_k(tx)$  with respect to  $x_j$  is given by  $t(\partial_j F_k)(tx)$  for  $x \in U$ , following from the chain rule. We use (14.5) for  $h = x$  in (14.6) and obtain, by partial integration,

$$\partial_j f(x) = \int_0^1 \underbrace{t \partial_x F_j(tx)} dt + \int_0^1 F_j(tx) dt = [t F_j(tx)]_0^1 - \int_0^1 F_j(tx) dt + \int_0^1 F_j(tx) dt = F_j(x)$$

where we have recognized the derivative with respect to  $t$  of  $F_j(tx)$  in the underbraced term. Thus,  $F = \nabla f$ , and  $f$  is continuously differentiable (by Theorem 10.12).  $\square$

EXERCISE 14.47. — For which values of  $\lambda \in \mathbb{R}$  is the vector field  $F : \mathbb{R}^2 \rightarrow \mathbb{R}^2$  defined by

$$F(x, y) = (\lambda x \exp(y), (y + 1 + x^2) \exp(y))$$

conservative? Determine a potential for  $f$  for these values.

*Proof of Theorem 14.44.* Suppose that  $\gamma_0$  and  $\gamma_1$  are two given piecewise  $C^1$  homotopic paths joining  $x_0$  and  $x_1$  and let  $H : [0, 1]^2 \rightarrow U$  be an homotopy, i.e.  $H(s, t)$  is continuous and  $H(0, t) = \gamma_0(t)$ ,  $H(1, t) = \gamma_1(t)$ , for all  $t \in [0, 1]$ .

The map  $H$  is continuous in the compact set  $[0, 1]^2$  so it is uniformly continuous. Also the image of  $H$ ,  $K := H([0, 1]^2)$  is a compact subset of  $U$  and hence the 1-Lipchitz (exercise) function  $x \mapsto d_{\mathbb{R}^n \setminus U}(x) = \inf \{|x - y| \mid y \in \mathbb{R}^n \setminus U\}$  attains its positive minimum in  $K$ .

By the previous considerations there exist  $\delta > 0$  and  $\epsilon > 0$  such that for all  $x = H(s, t) \in K$ , we have  $B_\delta(x) \subset U$  and  $\bar{x} = H(\bar{s}, \bar{t}) \in B_\delta(x)$  whenever  $|\bar{s} - s| < \epsilon$  and  $|\bar{t} - t| < \epsilon$ .

Fix now  $K > 10/\epsilon$  and consider the ‘mesh’

$$s_l := \epsilon l, \quad t_m := \epsilon m, \quad 0 \leq l, m \leq K.$$

For every  $0 \leq l, m \leq K$  let

$$x_{l,m} := H(s_l, t_m).$$

and let  $f_{l,m} : B_\delta(x_{l,m}) \rightarrow \mathbb{R}$  be the a potential of  $F$  restricted to  $B_\delta(x_{l,m})$ , which exists thanks to Lemma 14.46. (We can choose for concreteness the potential satisfying  $f_{l,m}(x_{l,m}) = 0$ , although this will not play any role in the proof).

Let us show that for all  $0 \leq \ell \leq K - 1$  and  $0 \leq l, m \leq K - 1$  we have

$$\begin{aligned} f_{l+1,m}(x_{l+1,m+1}) - f_{l+1,m}(x_{l+1,m}) &= + f_{l,m+1}(x_{l+1,m+1}) - f_{l,m+1}(x_{l,m+1}) \\ &\quad + f_{l,m}(x_{l,m+1}) - f_{l,m}(x_{l,m}) \\ &\quad - f_{l,m}(x_{l+1,m}) + f_{l,m}(x_{l,m}) \end{aligned} \quad (14.7)$$

Indeed, it follows from noticing first that

$$f_{l+1,m}(x_{l+1,m+1}) - f_{l+1,m}(x_{l+1,m}) = f_{l,m}(x_{l+1,m+1}) - f_{l,m}(x_{l+1,m})$$

because the functions  $f_{l+1,m}$ ,  $f_{l,m}$  are both potentials for  $X$  in  $B_\delta(x_{l+1,m}) \cap B_\delta(x_{l,m})$  and the two points  $x_{l+1,m+1}$  and  $x_{l+1,m}$  belong two the intersection of the two balls.

31

Similarly,

$$f_{l,m+1}(x_{l+1,m+1}) - f_{l,m+1}(x_{l,m+1}) = f_{l,m}(x_{l+1,m+1}) - f_{l,m}(x_{l,m+1}).$$

But then (14.7) follows from the trivial identity

$$\begin{aligned} f_{l,m}(x_{l+1,m+1}) - f_{l,m}(x_{l+1,m}) &= + f_{l,m}(x_{l+1,m+1}) - f_{l,m+1}(x_{l,m}) \\ &\quad + f_{l,m}(x_{l,m+1}) - f_{l,m}(x_{l,m}) \\ &\quad - f_{l,m}(x_{l+1,m}) + f_{l,m}(x_{l,m}) \end{aligned}$$

Finally notice that

$$\int_{\gamma_0} F = \sum_{m=0}^{K-1} (f_{0,m}(x_{0,m+1}) - f_{0,m}(x_{0,m}))$$

and

$$\int_{\gamma_1} F = \sum_{m=0}^{K-1} (f_{K,m}(x_{K,m+1}) - f_{K,m}(x_{K,m}))$$

Therefore, summing (14.7) over all  $0 \leq l \leq K - 1$  and all  $0 \leq m \leq K - 1$  and using that  $x_{l,0} = x_0$  and  $x_{l,K} = x_K$  for all  $l$ , so

$$f_{l,0}(x_{l+1,0}) + f_{l,0}(x_{l,0}) = f_{l,K}(x_{l+1,K}) + f_{l,K}(x_{l,K}) \quad \text{for all } 0 \leq l \leq K - 1,$$

31 we obtain  $\int_{\gamma_1} F - \int_{\gamma_0} F = 0$ .

□

---

# Index

- 1-norm, 31
- Cardioid, 144
- Cauchy sequence, 10
- closed, 14
- closure, 16
- compact, 21
- complete, 10
- completion, 11
- connected component, 28
- contraction, 20
- converges, 7
- critical point, 63
  
- diffeomorphism, 80
- differentiable, 40
- differential, 40
- discrete metric, 6
- distance, 5
- dyadic cube, 89
- dyadic set, 89
  
- Epicycloid, 144
- Equivalence
  - Norms, 36
- Euclidean isometry, 118
- Euclidean norm, 35
  
- fixed point, 20
  
- gradient, 63
  
- Hessian matrix, 69
  
- infinity norm, 31
- integrability condition, 151
- integrability conditions, 151
- interior, 16
  
- Jacobian matrix, 45
  
- Laplacian, 69
- limit, 7
- Lipschitz constant, 19
- Lipschitz continuous, 19
- loop, 151
  
- Manhattan metric, 6
- maximum norm, 31
- metric, 5
- metric space, 5
  
- Norm
  - Euclidean, 35
  - induced, 35
- norm, 31
  
- open, 14
  
- permutation of the coordinates, 82
- points, 5
  
- Riemman integrable, 102
- rigid motion, 118
- rotation, 145
- rotation number, 146
  
- Schwarz, Theorem of-, 53
- sequence, 7
- simply connected, 152
- space, 5
- standard metric, 6
- subsequence, 8
  
- term, 7
- topological boundary, 16
- topology, 14



uniformly continuous, 19

vector field, 143

vorticity, 145

---

# Bibliography

- [ACa2003] N. A'Campo, *A natural construction for the real numbers* arXiv preprint 0301015, (2003)
- [Apo1983] T. Apostol, *A proof that Euler missed: Evaluating  $\zeta(2)$  the easy way* The Mathematical Intelligencer **5** no.3, p. 59–60 (1983)
- [Aig2014] M. Aigner and G. M. Ziegler, *Das BUCH der Beweise* Springer, (2014)
- [Amm2006] H. Amann und J. Escher, *Analysis I*, 3. Auflage, Grundstudium Mathematik, Birkhäuser Basel, (2006)
- [Bla2003] C. Blatter, *Analysis I* ETH Skript, <https://people.math.ethz.ch/~blatter/dlp.html> (2003)
- [Bol1817] B. Bolzano, *Rein analytischer Beweis des Lehrsatzes, daß zwischen je zwei Werthen, die ein entgegengesetztes Resultat gewähren, wenigstens eine reelle Wurzel der Gleichung liege*, Haase Verl. Prag (1817)
- [Boo1847] G. Boole, *The mathematical analysis of logic* Philosophical library, (1847)
- [Can1895] G. Cantor, *Beiträge zur Begründung der transfiniten Mengenlehre* Mathematische Annalen **46** no.4, 481–512 (1895)
- [Cau1821] A.L. Cauchy, *Cours d'analyse de l'école royale polytechnique* L'Imprimerie Royale, Debure frères, Libraires du Roi et de la Bibliothèque du Roi. Paris, (1821)
- [Ded1872] R. Dedekind, *Stetigkeit und irrationale Zahlen* Friedrich Vieweg und Sohn, Braunschweig (1872)
- [Die1990] J. Dieudonné, *Elements d'analyse* Editions Jacques Gabay (1990)
- [Hat02] A. Hatcher, *Algebraic Topology* Cambridge University Press (2002)
- [Hil1893] D. Hilbert, *Über die Transzendenz der Zahlen  $e$  und  $\pi$*  Mathematische Annalen **43**, 216–219 (1893)
- [Hos1715] G.F.A. Marquis de l'Hôpital, *Analyse des Infiniment Petits pour l'Intelligence des Lignes Courbes* 2nde Edition, F. Montalant, Paris (1715)

- [Lin1894] E. Lindelöf, *Sur l'application des méthodes d'approximations successives à l'étude des intégrales réelles des équations différentielles ordinaires* Journal de mathématiques pures et appliquées **10** no.4, 117–128 (1894)
- [Rus1903] B. Russell, *The principles of mathematics* WW Norton & Company, (1903)
- [Rot88] J. J. Rotman, *An introduction to Algebraic Topology* Graduate Texts in Mathematics **119** Springer 1988
- [Smu1978] R. Smullyan, *What is the name of this book?* Prentice-Hall, (1978)
- [Zag1990] D. Zagier, *A one-sentence proof that every prime  $p \equiv 1 \pmod{4}$  is a sum of two squares.* Amer. Math. Monthly **97**, no.2, p. 144 (1990)