LECTURE NOTES ON ANALYSIS 2

RICCARDO CRISTOFERI

Contents

1. Structures on spaces 3
1.1. Scalar product spaces 3
1.2. Normed spaces 7
1.3. Metric spaces 12
1.4. Topological spaces 15
2. Convergence of sequences and compactness 16
2.1. Convergence of sequences in metric spaces 16
2.2. Compactness in metric spaces 20
2.3. Comparison of metrics and norms 24
3. Continuous functions 26
3.1. Continuity in metric spaces 26
3.2. Pointwise and uniform convergence 32
4. The space of continuous functions 37
4.1. Completeness 37
4.2. Characterization of compact sets: the Ascoli-Arzelà Theorem 38
4.3. Separability: the Theorems of Weierstraß 43
4.4. Separability: the Theorem of Stone 47
4.5. Nowhere differentiable functions 49
5. Differentiation of functions of several variables 51
5.1. Differentiability in the one dimensional case $N = 1$ 51
5.2. Differentiability in the general case $N \geq 1$ 52
5.3. Partial derivatives 54
5.4. Tangent hyperplane 58
5.5. Differentiability of functions from $\mathbb{R}^N$ to $\mathbb{R}^M$ 64
6. Vector fields and gradients 66
6.1. Schwarz's Theorem 67
6.2. Differential forms 69
6.3. Poincarè Lemma 71
6.4. Helmholtz Decomposition Theorem 75
7. Inverse Function Theorem 77
7.1. The differential as a tangent application 77
7.2. When the differential is injective 79
7.3. When the differential is surjective 81
7.4. Diffeomorphisms 82
8. Implicit Function Theorem 87
9. Analysis on submanifolds 94
9.1. Submanifolds in $\mathbb{R}^N$ 94
9.2. Critical points on submanifolds: Lagrange multipliers 102
10. The Darboux-Riemann integral and the Peano-Jordan content 106
10.1. The problems of the Cauchy-Riemann integration 106
10.2. The Peano-Jordan content 109

10.3. The Darboux-Riemann integral 113
10.4. Lebesgue's characterization of Riemann integrability 114
11. Lebesgue measure 118
11.1. Definition and relation to the Peano-Jordan content 118
11.2. Measurable sets 122
11.3. Negligible sets 127
11.4. Operations on measurable sets 128
11.5. Non-measurable sets 131
11.6. Vitali's characterization of Riemann integrability 132
12. Lebesgue integration 134
12.1. Lebesgue measurable functions 136
12.2. The Lebesgue integral for $\mathcal{L}^N$-measurable functions 140
12.3. The Lebesgue integral for general functions 145
13. Limiting theorems 147
13.1. Fatou's Lemma 147
13.2. Lebesgue Monotone Convergence Theorem 149
13.3. Lebesgue Dominated Convergence Theorem 150
13.4. Lebesgue integral as extension of the Riemann integral 152

1. Structures on spaces

The space $\mathbb{R}^N$ has several nice properties that allow to talk about convergence of sequences, continuity of a function, and differentiation of a function, among others. In order to work on more general spaces and be sure that such properties still hold, it is important to extrapolate the structures that allow to define such notions and to have results of interest in force. The property we are interested in is continuity. We will not work on the more general structure that allows for a definition of continuity, namely topological spaces (since there will be next semester an entire course devoted to that!), but we will focus our attention to spaces with a richer structure, namely metric ad normed spaces. In order to get to such objects, we will unveil several structures of $\mathbb{R}^N$ and investigate their relations.

The notion of continuity you learned in *Analysis 1* is based on the convergence of sequences: a function $f : \mathbb{R} \to \mathbb{R}$ is said to be continuous at a point $\bar{x} \in \mathbb{R}$ if

$$\lim_{n\to\infty} f(x_n) = f(\bar{x}),$$

for each sequence $(x_n)_{n\in\mathbb{N}}$ converging to $\bar{x}$. That is f transforms converging sequences of the domain space into converging sequences in the target space. Therefore, in order to generalize the notion of continuity to more general spaces, we need to generalize the notion of convergence of sequences. In $\mathbb{R}^N$, there is a canonical notion of convergence for sequences. Indeed, we say that a sequence $(x_n)_{n\in\mathbb{N}} \subset \mathbb{R}^N$ converges to some $\bar{x} \in \mathbb{R}^N$, if

$$\lim_{n\to\infty} \|x_n - \bar{x}\| = 0.$$

The idea behind such definition is that the points x_n becomes closer and closer to the point $\bar{x}$. We translate this concept in a mathematical form by using the notion of *(Euclidean) distance* d between the points x_n and $\bar{x}$. Such distance is the *norm* of the vector $x_n - \bar{x}$, namely

$$\mathrm{d}(x_n, \bar{x}) \coloneqq \|x_n - \bar{x}\|.$$

In turn, this is defined by using the *scalar (or inner) product* on $\mathbb{R}^N$ given by

$$\|v\| \coloneqq \langle v, v\rangle^{\frac{1}{2}} \coloneqq \left(\sum_{i=1}^{N} v_i^2\right)^{\frac{1}{2}},$$

for $v = (v_1, \dots, v_N) \in \mathbb{R}^N$. Therefore, the notion of convergence in $\mathbb{R}^N$ is based on the scalar product, which induces a norm, which induces a distance, which we used to translate mathematically the heuristic concept of *coming closer* in a quantitative way. Next sections will give conditions for generalizing these structures.

1.1. Scalar product spaces.

In $\mathbb{R}^N$ it is possible to do geometry. What are the main ingredients that allow to do that? The answer is: the possibility to sum vectors and to multiply them by a scalar, and having a notion of angle. The first points require to work with a vector space, while the latter to have a function that, for each pair of vectors, determines a notion of angle between them. This will be done through the concept of scalar (or inner) product by mimicking what happens in $\mathbb{R}^2$. If we consider a general triangle (see Figure 1), by al-Kashi's Theorem (the generalization of Pythagorean Theorem also known as the Theorem of cosine) we have that

$$c^2 = a^2 + b^2 - 2bc \cdot \cos\alpha.$$

By viewing the sides of the triangle as the vectors v, w, and $v - w$ respectively, we get

$$\|v - w\|^2 = \|v\|^2 + \|w\|^2 - 2\|v\| \cdot \|w\| \cdot \cos\alpha.$$

By expanding the square on the left-hand side, and using the fact that

$$\|z\|^2 = \sum_{i=1}^{N} z_i^2, \tag{1.1}$$

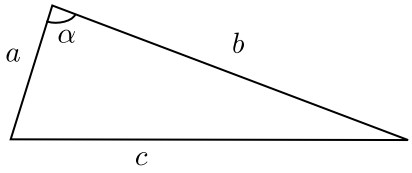

FIGURE 1. al-Kashi's Theorem

for all $z \in \mathbb{R}^N$, the above equality yields

$$\cos\alpha = \frac{1}{\|v\| \cdot \|w\|} \sum_{i=1}^{N} v_i w_i. \tag{1.2}$$

Therefore, taking also into consideration (1.1), we see from (1.2) that the function

$$(v, w) \mapsto \sum_{i=1}^{N} v_i w_i \tag{1.3}$$

is the crucial ingredient to define a notion of angle between the vectors v and w. Note that the definition of the cosine between v and w does not change if we multiply those vectors by any positive real numbers. The *Euclidean* notion of angle, namely that that you learned in high school, is based on the quantity (1.3). Therefore, a general notion of angle will depend on a generalization of this quantity that has to enjoy similar properties as (1.3). The essential ones are identified in the following definition.

Definition 1.1. Let $(X, +, \cdot)$ be a vector space over $\mathbb{R}$. A function $\langle \cdot, \cdot \rangle : X \times X \to \mathbb{R}$ is said to be a *(real) scalar* (or *inner) product* if:

(i) Linearity: $\langle \lambda v + \mu z, w \rangle = \lambda \langle v, w \rangle + \mu \langle z, w \rangle$, for every $v, w, z \in X$, and $\lambda, \mu \in \mathbb{R}$;
(ii) Symmetry: $\langle v, w \rangle = \langle w, v \rangle$, for every $v, w \in X$;
(iii) Definiteness: $\langle v, v \rangle \geq 0$ for every $v \in X$, where equality holds if and only if $v = 0$.

In case the vector space $(X, +, \cdot)$ is over $\mathbb{C}$, then a function $\langle \cdot, \cdot \rangle : X \times X \to \mathbb{C}$ is said to be a *(complex) scalar (or inner) product* if (i), (iii) hold, and property (ii) is replaced by

(ii') $\langle v, w \rangle = \overline{\langle w, v \rangle}$, for every $v, w \in X$,

where $\overline{\lambda}$ is the conjugate of $\lambda \in \mathbb{C}$. In either of the above cases, we say that $(X, \langle \cdot, \cdot \rangle)$ is a *scalar (or inner) product space*.

Remark 1.2. Note that scalar products are structures on a *vector space* $(X, +, \cdot)$, namely in an environment where a notion of sum of vectors and of multiplication by a scalar is defined.

Moreover, note that (i) and (ii) imply that a scalar product is *bilinear*, namely, it is linear in both of its entries.

Remark 1.3. The terminology *scalar product* means 'product among vectors that gives a scalar', not to be confused with the product of a vector with a scalar (of the underlining field). Note that there is also a notion of *vector product*, which means 'product among vectors that gives a vector'. You talked about this latter notion in *Calculus B* when discussing Stokes' Theorem.

Remark 1.4. A scalar product can be seen as a way to define a notion of *similarity* between two vectors. Indeed, if $v, w \in \mathbb{R}^N$ both have unit norm, then

$$\langle v - w, v - w \rangle = 2(1 - \cos\alpha),$$

where α is the angle between v and w relative to the scalar product $\langle \cdot, \cdot \rangle$. We can see that we can interpret $\cos\alpha$ as a measure of how different v and w are. In particular, if $\cos\alpha = 1$, they are the same, if $\cos\alpha = 0$ they have nothing to do with each other (namely, they are *orthogonal*), while if $\cos\alpha = -1$ they are in completely opposite directions.

For instance, in statistics, the *covariance* is, roughly speaking, a measure of the tendency of two random variables to have a linear dependence. It is possible to see that the covariance is a scalar product in the space of probabilities.

We now present a series of examples of scalar products on different vector spaces.

Example 1.5 (Euclidean product)**.** The most common example of a scalar product is the canonical Euclidean scalar product on $\mathbb{R}^N$, given by

$$\langle v, w\rangle \coloneqq \sum_{i=1}^{N} v_i w_i,$$

for $v, w \in \mathbb{R}^N$.

Example 1.6 (Generalization of the Euclidean product)**.** Let A be an $N \times N$ matrix, and consider the function

$$\langle v, w\rangle_A \coloneqq \sum_{i,j=1}^{N} a_{ij} v_i w_j = \langle w, (Av)\rangle, \tag{1.4}$$

for $v, w \in \mathbb{R}^N$. Assume that A is symmetric, and positive definite. Namely $a_{ij} = a_{ji}$ for all $i, j = 1, \dots, N$, and $\langle v^T, (Av)\rangle \geq 0$ for all $v \in \mathbb{R}^N$, where $\langle \cdot, \cdot, \rangle$ is the Euclidean scalar product defined in the previous example. Then, the above function defined in (1.4) is a scalar product on $\mathbb{R}^N$ (prove it!). Note that the standard Euclidean product corresponds to taking A to be the identity matrix. When are two vectors orthogonal with respect to the scalar product $\langle \cdot, \cdot\rangle_A$?

Example 1.7 (The l^2 scalar product)**.** Consider the space $X = l^2$ of sequences $(a_n)_{n\in\mathbb{N}}$ such that

$$\sum_{n\in\mathbb{N}} |a_n|^2 < \infty.$$

By using the inequality (prove it!)

$$(a+b)^2 \leq 2(a^2 + b^2),$$

it is possible to see that X is a vector space. For $a \coloneqq (a_n)_{n\in\mathbb{N}}$, and $b \coloneqq (b_n)_{n\in\mathbb{N}} \in l^2$, define

$$\langle a, b\rangle_{l^2} \coloneqq \sum_{n\in\mathbb{N}} a_n b_n.$$

Then, this function defines a scalar product on X (check it!).

Example 1.8 (The L^2 scalar product)**.** Consider the space X of functions $f : (0, 1) \to \mathbb{R}$ that are continuous and such that

$$\int_0^1 |f(x)|^2 \, \mathrm{d}x < \infty.$$

It is possible to see that it is a vector space (actually for more general functions than just continuous, and in general dimension). For $f, g \in X$ define

$$\langle f, g\rangle_{L^2} \coloneqq \int_0^1 f(x) g(x) dx.$$

Then, this is a scalar product on X (check it!).

As we said at the beginning, scalar product spaces are environments where it is possible to do geometry like in $\mathbb{R}^N$. In particular, in a scalar product space $(X, \langle \cdot, \cdot\rangle)$, the following identities hold (see Figure 2):

- The *Pythagorean Theorem*:

$$\langle v - w, v - w\rangle = \langle v, v\rangle + \langle w, w\rangle, \tag{1.5}$$

  for all $v, w \in X$ such that $\langle v, w\rangle = 0$;

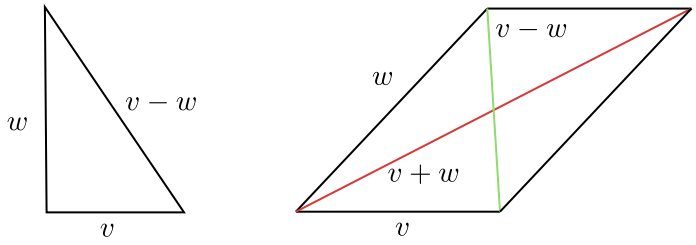

FIGURE 2. Two equalities that hold in scalar product spaces: the Pythagorean Theorem (on the left), and the parallelogram law (on the right).

- The *parallelogram law*:
$$2\langle v, v\rangle + 2\langle w, w\rangle = \langle v + w, v + w\rangle + \langle v - w, v - w\rangle, \tag{1.6}$$
for all $v, w \in X$.

It is also possible to define the notion of *orthogonal projection* of a vector w on a vector $v \neq 0$ as
$$\Pi_v(w) := \frac{\langle v, w\rangle}{\langle v, v\rangle} v. \tag{1.7}$$
In turns, this allows to introduce the notion of *orthogonal decomposition*:
$$w = \Pi_v(w) + z,$$
where $z \in X$ is defined by the above identity. Note that $\langle z, v\rangle = 0$. In particular, by using the Pythagorean Theorem (see (1.5)), we get that
$$\langle w, w\rangle = \langle \Pi_v(w), \Pi_v(w)\rangle + \langle z, z\rangle. \tag{1.8}$$
This orthogonal decomposition allows to prove an important property of inner products, the so called *Cauchy-Schwarz inequality*.

Proposition 1.9. *Let $(X, \langle\cdot,\cdot\rangle)$ be an scalar product space. Then, for all $v, w \in X$, the* Cauchy-Schwarz inequality *holds:*
$$|\langle v, w\rangle|^2 \leq \langle v, v\rangle \cdot \langle w, w\rangle,$$
and equality holds if and only if $w = \lambda v$, for some $\lambda \in \mathbb{R}$.

Proof. If $v = 0$ or $w = 0$, the inequality is trivial. Therefore, assume $v, w \neq 0$, and write (see (1.7))
$$w = \Pi_v(w) + z.$$
Using (1.8), we get
$$\langle w, w\rangle = \langle \Pi_v(w), \Pi_v(w)\rangle + \langle z, z\rangle = \frac{\langle v, w\rangle^2}{\langle v, v\rangle} + \langle z, z\rangle \geq \frac{\langle v, w\rangle^2}{\langle v, v\rangle},$$
where the last inequality follows from the fact that $\langle z, z\rangle \geq 0$. This proves the desired inequality. Note that $z = 0$ if and only if $w = \lambda v$ for some $\lambda \in \mathbb{R}$. This concludes the proof. □

Remark 1.10. In infinite dimensional vector spaces, the usual notion of a basis is basically useless. In order to construct a good substitute, the idea is to find a dense family of vectors that are orthonormal with respect to a given scalar product. In the case of the Lebesgue space L^2 (see Example 2.13), the scalar product introduced in Example 1.8 is the most commonly used. The study of such topic lies within the realm of Fourier Series, and it will be covered in the course *Introduction to Fourier series*.

1.2. **Normed spaces.** We now turn our attention to a more general structure on vector spaces. We want to define a notion of *length* of a vector. Such a notion has to be consistent with the idea we have in mind of length: it has to be non-negative, zero only for the null vector, positively homogeneous, and it has to satisfy the triangle inequality. Namely, the length of a composite path is no more than the sum of the lengths of each part of the path.

Definition 1.11. Let $(X, +, \cdot)$ be a vector space over $\mathbb{R}$. A function $\|\cdot\| : X \to \mathbb{R}$ is called a *norm* on X if:

(i) Triangle inequality: $\|v + w\| \leq \|v\| + \|w\|$ for all $v, w \in X$;
(ii) Homogeneity: $\|\lambda v\| = |\lambda| \, \|v\|$, for all $v \in X$, and $\lambda \in \mathbb{R}$;
(iii) Definiteness: $\|v\| \geq 0$ for all $v \in X$, and equality holds if and only if $v = 0$.

We now present a series of examples of norms on different vector spaces.

Example 1.12 (p-Minkowski norms in $\mathbb{R}^N$)**.** Let $X = \mathbb{R}^N$, and consider, for $p \in [1, \infty)$, the function

$$\|v\|_p \coloneqq \left(\sum_{i=1}^{N} v_i^p \right)^{\frac{1}{p}}.$$

Then, $\|\cdot\|_p$ is a norm on $\mathbb{R}^N$, called the *p-Minkowski norm*. For $p = 1$ it is also known as the *taxicab norm*, or *Manhattan distance*. To understand why, draw the unit ball of the norm $\|\cdot\|_1$. For $p = 2$, it is the usual Euclidean norm. Question: is the above function a norm, when $p < 1$? If yes, prove it, if not show what property fails. [Hint: draw the set $\{x \in \mathbb{R}^2 : \|x\|_p \leq 1\}$].

There is a natural relation between these norms (prove it!): if $1 \leq p < q < \infty$, then

$$\|v\|_q \leq \|v\|_p,$$

for all $v \in \mathbb{R}^N$. Moreover, it is possible to define a similar norm also for $p = \infty$. Indeed, for $v \in \mathbb{R}^N$, define

$$\|v\|_\infty \coloneqq \max\{v_1, \ldots, v_N\}.$$

Then, $\|\cdot\|_\infty$ is a norm on $\mathbb{R}^N$ (check it!) called the *infinite (or maximum) norm*. Moreover, it holds that (prove it!)

$$\lim_{p \to \infty} \|v\|_p = \|v\|_\infty$$

for all $v \in \mathbb{R}^N$.

Example 1.13 (p-Minkowski norms for sequences)**.** For $p \in [1, \infty)$, let $X = l^p$ of sequences $(a_n)_{n \in \mathbb{N}}$ such that

$$\sum_{n \in \mathbb{N}} |a_n|^p < \infty.$$

By using the inequality (prove it!)

$$(|a| + |b|)^p \leq 2^{p-1}(|a|^p + |b|^p), \tag{1.9}$$

it is possible to see that l^p is a vector space. Moreover, for $a = (a_n)_{n \in \mathbb{N}}$ the function

$$\|a\|_{l^p} \coloneqq \left(\sum_{n \in \mathbb{N}} |a_n|^p \right)^{\frac{1}{p}}$$

is a norm on l^p. This fact is not trivial to prove. The difficult property to check, for $p \in (1, \infty)$, is the triangle inequality, also known as the *Minkowski inequality*.

Example 1.14 (p-Minkowski norms for functions)**.** For $p \in [1, \infty)$, let X_p be the set of continuous functions $f : (0, 1) \to \mathbb{R}$ such that

$$\int_0^1 |f(x)|^p \, \mathrm{d}x < \infty.$$

By using (1.9), it is possible to see that X_p is a vector space. Define the function

$$\|f\|_{L^p} \coloneqq \left(\int_0^1 |f(x)|^p \, \mathrm{d}x\right)^{\frac{1}{p}},$$

for $f \in X_p$. Then, $\|\cdot\|_{L^p}$ is a norm on X_p. As for the case of sequences, the difficult property to check, for $p \in (1, \infty)$, is the triangle inequality, known as the *Minkowski inequality*. The reason why it has the same name as the one in the previous example, is because, by using notions that you will learn in *Measure Theory*, it is possible to see a series as the integration of the sequence $(a_n)_{n\in\mathbb{N}}$, which is nothing but a function $a : \mathbb{N} \to \mathbb{R}$, with respect to a certain measure. The Minkowski inequality holds indeed for more general measures than the one used to compute Riemann integrals.

Finally, for $p = \infty$, it is possible to define a similar norm. Let X_∞ be the space of continuous functions $f : (0, 1) \to \mathbb{R}$ such that

$$\|f\|_\infty \coloneqq \inf\{c \geq 0 \, : \, |f(x)| \leq c, \forall x \in (0, 1)\} < \infty.$$

Then, it is possible to see that X_∞ is a vector space, and that $\|\cdot\|_\infty$ is a norm on it. Moreover, $X_\infty \subset X_p$ for all $p \in [1, \infty)$, and

$$\lim_{p\to\infty} \|f\|_p = \|f\|_\infty,$$

for all $f \in X_\infty$.

Example 1.15 (Supremum (or uniform) norm)**.** Let $X \subset \mathbb{R}^N$ be a set. Consider the space $B(X)$ of bounded scalar functions on X. Then, the function

$$\|f\|_{C^0(X)} \coloneqq \sup_X |f|$$

for $f \in B(X)$ is a norm on $B(X)$, called the *supremum (or uniform) norm*. It is also denoted by $\|\cdot\|_\infty$, since it coincides with the norm on L^∞ defined in the previous example. Note that if X is compact, then the supremum is attained (see Theorem 3.15).

Example 1.16 (Operator norm)**.** Let $\mathcal{L}(\mathbb{R}^N; \mathbb{R}^M)$ be the space of linear maps between $\mathbb{R}^N$ and $\mathbb{R}^M$. It is easy to see that it is a vector space under the natural notion of sum of operators, and multiplication with a scalar. For $L \in \mathcal{L}(\mathbb{R}^N; \mathbb{R}^M)$, we define the function

$$\|L\|_{\mathcal{L}(\mathbb{R}^N;\mathbb{R}^M)} \coloneqq \sup\{\|L(v)\|_{\mathbb{R}^M} \, : \, \|v\| \leq 1\} = \sup\left\{\frac{\|L(v)\|_{\mathbb{R}^M}}{\|v\|_{\mathbb{R}^N}} \, : \, v \neq 0\right\},$$

where the last equality follows by the linearity of L. Then, $\|\cdot\|_{\mathcal{L}(\mathbb{R}^N;\mathbb{R}^M)}$ is a norm on $\mathcal{L}(\mathbb{R}^N; \mathbb{R}^M)$.

The study of linear operators in infinite dimensional vector spaces has a lot of interest and applications. For instance, Quantum Mechanics can be introduced by using the notions of linear operators between a certain class of infinite dimensional vectors spaces (called Hilbert spaces), and the notion of projections. Moreover, in the 20th century, mathematicians started looking at linear Partial Differential Equations (you can get acquaintance with PDEs in *Introduction to Partial Differential Equations*), as linear operators, rather than pointwise equalities (this modern view will be presented in a forthcoming Master course on *Sobolev spaces and PDEs*). The basics for the study of infinite dimensional vector spaces and linear operators between them will be covered in *Introduction to Functional Analysis*.

Remark 1.17. It seems quite natural that the *norm* of a vector is non-negative, and that the only vector with zero norm is the null vector. However, generalization of the notion of norm are needed in many applications. In particular, when the triangle inequality is generalized to

$$\|v + w\| \leq K(\|v\| + \|w\|)$$

or all $v, w \in X$, for some $K > 0$, we talk of a *quasi-norm*. An example is the generalization of the p-Minkowski norm to the case $p \in (0, 1)$.

Moreover, when the definitenss is weakened to

$$\|v\| \geq 0$$

for all $v \in X$, we talk about a *semi-norm*. An example is given by

$$f \mapsto \|f'\|_{C^0((0,1))},$$

for functions $f \in C^1((0,1))$.

What is the relation between scalar product spaces and normed spaces? Next result shows that a scalar product induces a norm.

Lemma 1.18. *Let $(X, \langle \cdot, \cdot \rangle)$ be an scalar product space. Then, the function*

$$\|v\| := \langle v, v \rangle^{\frac{1}{2}} \tag{1.10}$$

is a norm on X.

The proof is left as an exercise for the reader.

Remark 1.19. In an inner product space X, the Cauchy-Schwarz inequality (see Proposition 1.9) writes as

$$|\langle v, w \rangle| \leq \|v\| \cdot \|w\|,$$

for all $v, w \in X$.

Do all norms come from a scalar product, by using the natural relation (1.10)? The answer is no. Indeed, for $p \neq 2$, the p-Minkowski norm on $\mathbb{R}^N$ does not come from a scalar product (see (1.6)), since it does not satisfy the parallelogram law (see (1.6)) (Prove it!).

Now the questions is if it is possible to give a characterization of norms that come from a scalar product. It turns out that the parallelogram law is both a necessary and a sufficient condition for a norm to come from a scalar product.

Proposition 1.20. *Let $(X, \|\cdot\|)$ be a normed space. If the parallelogram law (see (1.6)) holds, then there exists a scalar product that induces that norm. In particular, the scalar product is given by*

$$\langle v, w \rangle := \frac{\|v+w\|^2 - \|v-w\|^2}{4}, \tag{1.11}$$

for all $v, w \in X$.

Proof. Step 1. We claim that, for each $x, y \in X$, it holds that

$$\langle x+y, z \rangle = \langle x, z \rangle + \langle y, z \rangle.$$

Indeed, from the parallelogram law (see (1.6)) we get that

$$2\|x+z\|^2 + 2\|y\|^2 = \|x+y+z\|^2 + \|x-y+z\|^2,$$

from which we get

$$\|x+y+z\|^2 = 2\|x+z\|^2 + 2\|y\|^2 - \|x-y+z\|^2 = 2\|y+z\|^2 + 2\|x\|^2 - \|y-x+z\|^2,$$

where the last inequality follows from interchanging the role of x and y. By summing the above two equalities, we get

$$2\|x+y+z\|^2 = 2\|x+z\|^2 + 2\|y+z\|^2 + 2\|y\|^2 + 2\|x\|^2 - \|y-x+z\|^2 - \|x-y+z\|^2, \tag{1.12}$$

and, by using $-z$ instead of z, also

$$2\|x+y-z\|^2 = 2\|x-z\|^2 + 2\|y-z\|^2 + 2\|y\|^2 + 2\|x\|^2 - \|y-x-z\|^2 - \|x-y-z\|^2. \tag{1.13}$$

Thus, from (1.12) and (1.13) we get

$$\begin{aligned}\langle x+y, z \rangle &= \frac{\|x+y+z\|^2 - \|x+y-z\|^2}{4} \\ &= \frac{1}{4}\left(\|x+z\|^2 - \|x-z\|^2\right) + \frac{1}{4}\left(\|y+z\|^2 - \|y-z\|^2\right)\end{aligned}$$

$$= \langle x, z\rangle + \langle y, z\rangle,$$

as desired.

Step 2. From step 1 and using induction, we get that

$$\langle \lambda x, y\rangle = \lambda\langle x, y\rangle, \tag{1.14}$$

for all $x, y \in X$, and $\lambda \in \mathbb{N}$. Moreover, since

$$\langle -x, y\rangle = -\langle x, y\rangle,$$

we get that (1.14) holds also for all $\lambda \in \mathbb{Z}$. Now, let $\lambda = \frac{p}{q}$, for $p, q \in \mathbb{Z} \setminus \{0\}$. Consider the vector $v := \frac{x}{q}$. Then,

$$\langle \frac{p}{q}x, y\rangle = p\langle \frac{x}{q}, y\rangle = p\langle v, y\rangle = \frac{p}{q}q\langle v, y\rangle = \frac{p}{q}\langle qv, y\rangle = \frac{p}{q}\langle x, y\rangle,$$

which gives $\langle \lambda x, y\rangle = \lambda\langle x, y\rangle$. Finally, for $\lambda \in \mathbb{R}$, we argue by continuity. Namely, since the continuous map

$$t \mapsto \frac{1}{t}\langle tx, y\rangle$$

coincides with the constant map $\langle x, y\rangle$ for all $t \in \mathbb{Q}$, we conclude since two continuous maps that are equal on a dense set, are equal everywhere. □

Finally, we would like to give a geometric characterization of a norm. We start with the following question: can two different norms have the same unit ball? The answer is no: a norm is completely determined by its *unit ball*. Indeed, let $(X, \|\cdot\|)$ be a normed space, and define its unit ball

$$B_{\|\cdot\|} := \{x \in X \,:\, \|x\| < 1\}.$$

It is easy to see that if two norms have the same unit ball, then the two norms coincide (prove it!). Therefore, defining a norm is equivalent to specifying its unit ball, namely a set. This raises the question: what sets can be the unit ball of a norm? First of all, we identify some properties of unit balls of norms.

Lemma 1.21. *Let $(X, \|\cdot\|)$ be a normed space. Then $B_{\|\cdot\|}$ is convex, and symmetric with respect to the origin. That is, if*

$$\lambda x + (1-\lambda)y \in B_1,$$

for all $x, y \in B_1$ and $\lambda \in [0, 1]$, and

$$x \in B_1 \quad \Leftrightarrow \quad -x \in B_1,$$

respectively.

Proof. Exercise for the reader. Note that convexity is related to the triangle inequality, while symmetry to homogeneity and definiteness, which imply

$$\|\lambda v\| = |\lambda|\|v\|,$$

for all $v \in X$, and $\lambda \in \mathbb{R}$. □

The question is now: are the above properties also sufficient for a set $E \subset X$ to be the unit ball of some norm? We will give an answer to this question only for $X = \mathbb{R}^N$. In particular, in $\mathbb{R}^N$ we know what a closed set is.

Proposition 1.22. *Let $E \subset \mathbb{R}^N$ be a bounded convex closed set (with respect to the Euclidean topology) that is symmetric with respect to the origin, and that is not contained in any k-dimensional linear space, with $k \leq N-1$. Then, there exists a unique norm on $\mathbb{R}^N$ having E as the closure of its unit ball.*

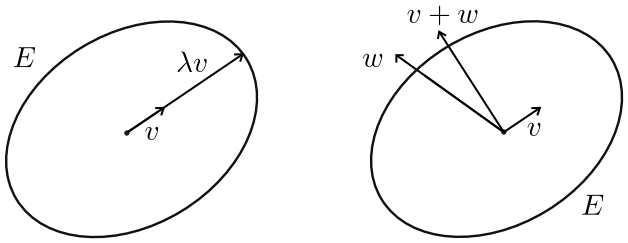

FIGURE 3. The construction of the norm of the vector v on the left. The proof of the triangle inequality on the right.

Proof. The idea of the proof is the following: let $v \in \mathbb{R}^N$. Consider, $\mathbb{R}^+ v$, the half-line in the direction of v, namely

$$\mathbb{R}^+ v := \{tv \ : t \geq 0\}.$$

The assumption that E is not contained in any k-dimensional linear space with $k \leq N-1$, together with the convexity of E and the fact that it is closed, ensures that (see Figure 3)

$$\mathbb{R}^+ v \cap E = \{tv : t \in [0, a]\}, \tag{1.15}$$

for some $a > 0$. In particular, if $\|\cdot\|$ is the norm whose unit ball is E, then we must have $\|av\| = 1$. Thus, defining

$$\|v\| = \frac{1}{a},$$

is the only way for the function $\|\cdot\|$ (that we need to prove to be a norm), to have E as its unit ball.

We formalize this idea mathematically as follows: for $v \in \mathbb{R}^N$ define

$$\|v\| := \min\left\{t > 0 : \frac{v}{t} \in E\right\}. \tag{1.16}$$

Note that, thanks to (1.15), we have that

$$\mathbb{R}^+ v \cap E = \left\{\frac{v}{t} : t \in \left[\frac{1}{a}, \infty\right]\right\},$$

and our definition of the norm of v is $1/a$.

Step 1: Homogeneity. Let $v \in \mathbb{R}^N$ and $\lambda > 0$. Then

$$\|\lambda v\| = \min\left\{t > 0 : \frac{\lambda v}{t} \in E\right\} = \lambda \min\left\{s > 0 : \frac{v}{s} \in E\right\} = \lambda \|v\|,$$

where in the second equality we used the change of variable $s = \frac{t}{\lambda}$. This proves that $\|\lambda v\| = \lambda \|v\|$. By symmetry of E, this is true also for $\lambda < 0$.

Step 2: Definiteness. It is clear that $\|v\| \geq 0$ for all $v \in \mathbb{R}^N$. Moreover, $\|v\| < \infty$, since E is not contained in any k-dimensional linear space with $k \leq N-1$. Moreover, since E is bounded, we have that, if $v \neq 0$, there exists $t > 0$ such that $\frac{v}{t} \notin E$, and thus $\|v\| > 0$.

Step 3: Triangle inequality. Let $v, w \in \mathbb{R}^N \setminus \{0\}$. Then, by the definition of the norm (see (1.16)), and the fact that E is closed, we have that

$$\frac{v}{\|v\|} \in E, \qquad \frac{w}{\|w\|} \in E.$$

By the convexity of E, we get that

$$\frac{v+w}{\|v\|+\|w\|} = \frac{\|v\|}{\|v\|+\|w\|}\frac{v}{\|v\|} + \frac{\|w\|}{\|v\|+\|w\|}\frac{w}{\|w\|} \in E,$$

which gives the triangle inequality.

Step 4: Compatibility. Finally, we prove that E is the unit ball of the norm $\|\cdot\|$. We start by proving that $E \subset B_{\|\cdot\|}$. Let $v \in E$. Then

$$\min\left\{t > 0 : \frac{v}{t} \in E\right\} \leq 1,$$

and thus $\|v\| \leq 1$. To prove the opposite inclusion, let $w \in \mathbb{R}^N$ such that $\|w\| \leq 1$. By using the definition of the norm, we get that

$$\min\left\{t > 0 : \frac{w}{t} \in E\right\} \leq 1,$$

which implies that $w \in E$.

Step 5: Uniqueness. Uniqueness of the norm having E has the unit ball follows from homogeneity. □

Example 1.23 (Crystalline norms)**.** A class of sets that satisfy the assumptions of Proposition 1.22 is that of convex polygons in $\mathbb{R}^2$. The norms they generate are called *crystalline norms*, and they have applications in materials science, for instance, in relation of the formation of crystals. The reason why such norms are important is because they favor certain directions more than other, and this dependence is piecewise constant.

1.3. **Metric spaces.** Finally, we consider an even more general structure on *sets*, one that allows to define the notion of *distance* between two points.

Definition 1.24. Let X be a set. A function $\mathrm{d} : X \times X \to [0, \infty)$ satisfying

(i) Symmetry: $\mathrm{d}(x, y) = \mathrm{d}(y, x)$, for all $x, y \in X$;
(ii) Triangle inequality: $\mathrm{d}(x, y) \leq \mathrm{d}(x, z) + \mathrm{d}(z, y)$, for all $x, y, z \in X$
(iii) Definiteness: $\mathrm{d}(x, y) = 0$ if and only if $x = y$,

is said to be a *distance* (or a *metric*) on X.

Remark 1.25. Note that a distance can be defined on a *set*! Indeed, we do not need any underlining linear structure on X.

Since a distance is defined on a set, this allows to define it on more general ambient spaces than those seen in the previous sections. We now present some interesting examples.

Example 1.26 (Hamming distance)**.** For $X = \mathbb{R}^N$, consider the function $\mathrm{d} : \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{N}$ given by

$$\mathrm{d}(x, y) \coloneqq \#\{i \in \{1, \dots, N\} \,:\, x_i \neq y_i\},$$

namely the number of different entries of the vectors $x, y \in \mathbb{R}^N$. Then, it is a distance (prove it!) called *Hamming distance*, used in information theory to determine how many substitutions one has to do to transform one string of characters into another.

Example 1.27 (Distance on graphs)**.** Let X be a connected graph, namely a set of vertexes $\{p_1, \dots, p_k\}$ and edges between them. We assume that for each pair of vertexes there is a *path* of edges connecting them. We define the distance between two vertices as the length of a shortest path between them.

Example 1.28 (2-Wasserstein distance on empirical measures)**.** The objects we are interested in here are k indistinguishable particles in $\mathbb{R}^N$. We want to define a notion of distance from two of such objects. One way to do it, is to consider the best labeling of the particles that gives the smallest sum of the Euclidean distance between particles with the same label. Namely,

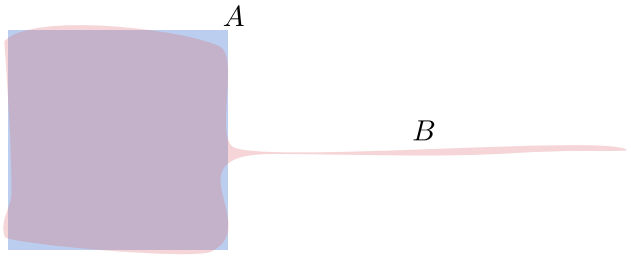

FIGURE 4. An example of a set, B, with small L^1, but large Hausdorff distance, from a set A.

we consider the family S_k of all possible permutations of k elements, and, for two sets $X \coloneqq \{x_1, \dots, x_k\}$, and $Y \coloneqq \{y_1, \dots, y_k\}$ of k indistinguishable particles, we define

$$\mathrm{d}(X, Y) \coloneqq \min \left\{ \sum_{i=1}^{k} |x_i - y_{\sigma(i)}|^2 : \sigma \in S_k \right\}^{\frac{1}{2}}.$$

Such distance is called the 2-Wasserstein distance, and it is used in the field of Optimal Transport.

How to quantify how two sets are *far apart* from each other? We will see two ways to do it.

Example 1.29 (L^1 distance on sets)**.** The idea is to identify a set $E \subset \mathbb{R}^N$ with its characteristic function $\mathbb{1}_E$ defined as

$$\mathbb{1}_E(x) \coloneqq \begin{cases} 1 & \text{if } x \in E, \\ 0 & \text{else.} \end{cases}$$

The L^1 norm between two sets $E, F \subset \mathbb{R}^N$ is defined as

$$\|A - B\|_{L^1} \coloneqq \|\mathbb{1}_A - \mathbb{1}_B\|_{L^1(\mathbb{R}^N)} = |A \triangle B|,$$

where $A \triangle B \coloneqq (A \setminus B) \cup (B \setminus A)$ is the symmetric difference between E and F. Namely, the L^1 norm measures the *volume* of the non-overlapping region of E and F.

Suppose we are now interested in how two shapes are different, but up to rigid motions. It is possible to modify such a norm to obtain a distance that is invariant under rigid motions as follows:

$$\mathrm{d}(A, B) \coloneqq \inf\{\|R(A) - B\|_{L^1} : R : \mathbb{R}^N \to \mathbb{R}^N \text{ is a rigid motion } \}.$$

Example 1.30 (Hausdorff distance)**.** Let X be the family of all compact sets of $\mathbb{R}^N$. Define the function

$$(A, B) \mapsto \max\left\{ \sup\{ \mathrm{d}(x, Y) \,:\, x \in X\}, \sup\{ \mathrm{d}(X, y) \,:\, y \in Y\} \right\},$$

for all $A, B \in X$, where, given a point $p \in \mathbb{R}^N$ and a set $E \subset \mathbb{R}^N$, we define

$$\mathrm{d}(p, E) \coloneqq \inf\{\|p - e\| \,:\, e \in E\}.$$

Then, it is a distance known as the *Hausdorff distance*, and it measures the maximum distance needed to go from the point of A which is farthest away from B to B itself (or viceversa).

This distance is stronger than the L^1 distance. Indeed, two sets have arbitrarily small L^1 distance and arbitrarily large Hausdorff distance. For instance, the set B in Figure 4 has small L^1 distance from A, but a large Hausdorff distance from A.

Next result shows that a norm induces a distance.

Lemma 1.31. *Let $(X, \|\cdot\|)$ be an normed space. Then, the function $\mathrm{d}_{\|\cdot\|} : X \times X \to \mathbb{R}$ given by*

$$\mathrm{d}_{\|\cdot\|}(x, y) \coloneqq \|x - y\|$$

is a distance on X.

The proof is left as an exercise for the reader.

Remark 1.32. Note that the above result implies that all of the example of scalar product spaces, and metric spaces given in the previous sections are also examples of metric spaces.

In particular, the metric on the space of bounded functions induced by the supremum norm (see Example 1.15) is called the *uniform metric*.

Do all distances on vector spaces come from a norm? The answer is no. Indeed, consider the function $\mathrm{d} : \mathbb{R} \times \mathbb{R} \to [0, \infty)$ given by

$$\mathrm{d}(x, y) \coloneqq \begin{cases} 0 & \text{if } x = y, \\ 1 & \text{else.} \end{cases} \tag{1.17}$$

Then, d is a distance (prove it!), but it cannot come from a norm. Indeed, if by absurd there was a norm $\|\cdot\|$ on $\mathbb{R}$ inducing the above distance, namely such that

$$\mathrm{d}(x, y) = \|x - y\|.$$

Then, by taking $|\lambda| \neq 1$ and $x \neq y$, we would have

$$1 = \mathrm{d}(\lambda x, \lambda y) = \|\lambda x - \lambda y\| = |\lambda| \|x - y\| = |\lambda| \, \mathrm{d}(x, y) = |\lambda|,$$

which is a contradiction with the choice of λ.

There is a simple way to characterize distances on a vector space that come from a norm.

Proposition 1.33. *Let (X, d) be a metric space, and assume that X is a vector space. Then, the distance d comes from a norm if and only if it is homogeneous and translation invariant. Namely, if for all $x, y, z \in X$ and $\lambda \in \mathbb{R}$ it holds*

$$\mathrm{d}(\lambda x, \lambda y) = |\lambda| \, \mathrm{d}(x, y),$$

and

$$\mathrm{d}(x + z, y + z) = \mathrm{d}(x, y),$$

respectively.

The proof is left as an exercise for the reader.

Remark 1.34. If (X, d) is a metric space, and $A \subset X$ is any subset, then (A, d) is also a metric space. This is something that you cannot do with a norm, since you need to ask the subset A to have a linear structure in order to obtain a metric. As an example, consider $X = \mathbb{R}^3$, and $A = \mathbb{S}^2$, the unit sphere. Then, the Euclidean norm restricted to $\mathbb{S}^2$ is not a norm on $\mathbb{S}^2$. On the other hand, if we consider the distance induced by the Euclidean metric, then its restriction to $\mathbb{S}^2$ is a metric on $\mathbb{S}^2$.

Example 1.35. Let $X = \mathbb{S}^2$ be the unit sphere in $\mathbb{R}^3$. Given two points $p, q \in \mathbb{S}^2$, we define

$$\mathrm{d}(p, q) \coloneqq \inf \left\{ \int_0^1 |\gamma'(t)| \, dt \, : \, \gamma : [0, 1] \to \mathbb{S}^2 \text{ } \textit{differentiable and s.t.} \text{ } \gamma(0) = p, \gamma(1) = q \right\}.$$

Then, d is a distance on $\mathbb{S}^2$. Note that this is an example of a distance that is induced by a notion of *length* of curves.

Remark 1.36. We would like to stress again that a metric, and therefore a norm, is a way to *understand* a particular relation between objects in a set. Different metrics, or norms, on the same space correspond to different points of view on the same class of objects. For instance, on the space $X = C^1((0, 1))$ we can consider the norms

$$\|f\|_{C^0}, \qquad \|f\|_{C^1} \coloneqq \|f\|_{C^0} + \|f'\|_{C^0}.$$

Take, for $n \in \mathbb{N} \setminus \{0\}$, the function $f_n : [0,1] \to \mathbb{R}$ given by

$$f_n(x) \coloneqq \begin{cases} nx & \text{if } x \in \left[0, \frac{1}{n^2}\right], \\ -nx + \frac{2}{n} & \text{if } x \in \left[\frac{1}{n^2}, \frac{2}{n^2}\right], \\ 0 & \text{if } x \in \left[\frac{2}{n^2}, 1\right], \end{cases}$$

Then

$$\|f_n\|_{C^0} = \frac{1}{n}, \qquad \|f_n\|_{C^1} = n.$$

Thus, for increasing values of n, the C^0 norm of f_n vanishes, while its C^1 norm blows up.

1.4. **Topological spaces.** In this section, we briefly mention the idea of the notion of topological spaces. Deeper investigations on this topic will be undertaken in the next semester's course *Topology*. We started by trying to understand what are the key ingredients needed to talk about convergence of sequences, in order to extend such notion to more general spaces, and, in turn, also the notion of continuity of functions. In a metric space, it is possible to define a notion of convergence based on the distance d. Such notion of convergence is *quantitative*, namely we want the number $\mathrm{d}(x_n, \bar{x})$ to go to zero, in order to say that x_n is closer and closer to $\bar{x}$. If we drop the requirement of having a quantitative knowledge of how close two points are, we can generalize by noting that $\mathrm{d}(x_n, \bar{x}) < r$, for some $r > 0$, is equivalent to say that $x_n \in B(\bar{x}, r)$. In particular, the condition

$$\lim_{n\to\infty} \mathrm{d}(x_n, \bar{x}) = 0$$

is equivalent to require that

$$x_n \in B(\bar{x}, r),$$

for all $n \geq \bar{n}$, for some $\bar{n} \in \mathbb{N}$. Therefore, it is possible to use *balls* to define the notion of being arbitrarily close to a point $\bar{x}$. Note that, since open sets are unions of balls, all of the above are equivalent to require that, for each open set $U \subset \mathbb{R}^N$ that contains $\bar{x}$, there exists $\bar{n} \in \mathbb{N}$ such that

$$x_n \in U$$

for all $n \geq \bar{n}$. This allows to generalize the notion of convergence by considering a set X, and a family of *open sets* $(O_i)_i \in I$ satisfying similar properties to those that open sets in $\mathbb{R}^N$ enjoy: the family $(O_i)_i \in I$ is closed under arbitrary union, finite intersection, it contains the empty set and the entire space X. This is the notion of *topological space*. Fine properties will be studied in the course Topology, next semester. For what concerns us, we will define several notions that we will need in two ways: a *sequential way*, and a *topological* way. In metric spaces, we will prove that the two notions coincide, but sometimes it will be useful to use one or the other. The reason why in metric spaces (or, more in general, in nice topological spaces) the sequential and the topological notion coincide, is because in a metric spaces (X, d) we have, for each $x \in X$, a *countable* sequence of balls $(B(x, 1/i))_{i\in\mathbb{N}\setminus\{0\}}$. This allows to consider such family instead of the uncountable family of open sets containing the point x. In particular, since a sequence is a *countable* object, we can relate sequences with this family of balls.

Finally, a comment on topological spaces. You might think that, if you are interested in Analysis, you won't need such abstract and complicated notion of topological spaces, since the good old $\mathbb{R}^N$ is good enough. Unfortunately (or, better, fortunately!), you couldn't be more wrong! Modern (and even the not so modern) directions of research and applications in (applied) analysis, engineering, machine learning, use topological spaces, where the topology used does not come from a norm! And this is not because mathematicians like to make things complicated: it's the way we understand/investigate the world around us that requires us to work with such mathematical structures.

2. Convergence of sequences and compactness

We have now introduced the structures that allow us to talk about convergence of sequences. This notion will depend on the metric we choose: different metrics will give different notions of convergence.

2.1. **Convergence of sequences in metric spaces.** When talking about sequences, there are two main notations in use: $\{a_n\}_{n\in\mathbb{N}}$, and $(a_n)_{n\in\mathbb{N}}$. From the technical point of view, the latter is the correct one, since a sequence in X is a function $\mathbb{N}\to X$, and therefore we can identify it with a *vector* $(a_n)_{n\in\mathbb{N}} = (a_1, a_2, \dots, a_n, \dots)$ with countably many entries. On the other hand, the notation $\{a_n\}_{n\in\mathbb{N}}$ refers to a *set* whose elements are $a_1, a_2, \dots, a_n, \dots$. In this case, the *order* of the elements is not taken into account, and that is the issue with that notation. The advantage of the latter notation, is that it is possible to write $\{a_n\}_{n\in\mathbb{N}} \subset X$ instead of writing 'let $(a_n)_{n\in\mathbb{N}}$ be a sequence in X'. In (modern, or classical) literature, both notations are in use.

Definition 2.1. Let (X, d) be a metric space, $(a_n)_{n\in\mathbb{N}}$ a sequence, and $a \in X$. If

$$\lim_{n\to\infty} \mathrm{d}(a_n, a) = 0, \tag{2.1}$$

we say that the sequence *converges to a with respect to the metric* d, or *in the metric* d. In this case, we write $a_n \to a$ with respect to (w.r.t.) d, or $a_n \xrightarrow{\mathrm{d}} a$, or

$$\lim_{n\to\infty} a_n = a,$$

if the metric has been specified.

Remark 2.2. In a normed space, condition (2.1) writes as

$$\lim_{n\to\infty} \|a_n - a\| = 0,$$

while in a scalar product space

$$\lim_{n\to\infty} \langle a_n - a, a_n - a\rangle = 0.$$

Note that this latter expression expands as

$$\lim_{n\to\infty} \big[\, \|a_n\|^2 + \|a\|^2 - 2\langle a_n, a\rangle \,\big] = 0.$$

A first property of a limit of a sequence in a metric space is that it is unique.

Lemma 2.3. *Let $(a_n)_{n\in\mathbb{N}}$ be a sequence on a metric space converging to some $a \in X$. If the same sequence converges to $b \in X$, then $a = b$.*

The proof is left as an exercise for the reader.

Remark 2.4. The above result seems trivial, but it is, in general, not true on topological spaces.

We now investigate the relation of sequences and subsequences with respect to convergence. A subsequence is an increasing selection of elements of the original subsequence. Namely, a subsequence of $(a_n)_{n\in\mathbb{N}}$ is $(a_{n_i})_{i\in\mathbb{N}}$, where $i \mapsto n_i$ is increasing.

Lemma 2.5. *Let $(a_n)_{n\in\mathbb{N}}$ be a sequence on a metric space (X, d) converging to some $a \in X$ in the metric* d*. Then, every subsequence $(a_{n_i})_{i\in\mathbb{N}}$ converges to a.*

The proof is left as an exercise for the reader.

The question is now: is it true that if a sequence $(a_n)_{n\in\mathbb{N}}$ is such that there exists $a \in X$ for which *every* subsequence converges to a, then the entire sequence converges to a as well? The answer is yes, and the proof is easy. A more useful property is a version of the above claim with a weaker assumption.

Proposition 2.6 (Urysohn property)**.** *Let $(a_n)_{n\in\mathbb{N}}$ be a sequence in a metric space satisfying the following property: there exists $a \in X$ such that every subsequence $(a_{n_k})_{k\in\mathbb{N}}$ has a further subsequence converging to a. Then, the entire sequence converges to a.*

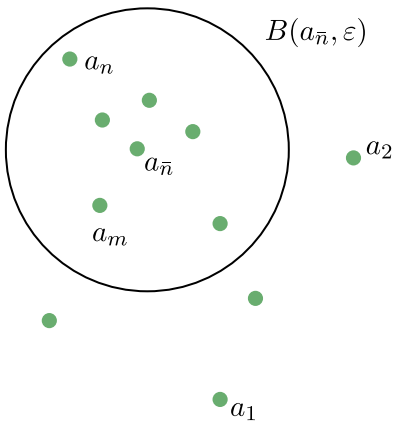

FIGURE 5. The idea of the definition of a Cauchy sequence: elements gets closer and closer to each other.

Proof. Assume by contradiction that the sequence $(a_n)_{n\in\mathbb{N}}$ does not converge to $a \in X$. Then, there exists $\varepsilon > 0$ and a subsequence $(a_{n_k})_{k\in\mathbb{N}}$ such that

$$\mathrm{d}(a_{n_k}, a) > \varepsilon,$$

for all $k \in \mathbb{N}$. This is a contradiction, since from such a subsequence we cannot extract any further subsequence converging to a. □

We would like to understand the behaviour of converging sequences. Heuristically, if a sequence converges to a limit, then all of its elements become closer and closer to each other. We formalize this idea as follows (see Figure 5).

Definition 2.7. A sequence $(a_n)_{n\in\mathbb{N}}$ in a metric space is said to be a *Cauchy sequence* if for every $\varepsilon > 0$ there exists $\bar{n} \in \mathbb{N}$ such that

$$\mathrm{d}(a_n, a_m) < \varepsilon,$$

for all $n, m \geq \bar{n}$.

Remark 2.8. Note that the index $\bar{n} \in \mathbb{N}$ *depends* on ε. Smaller ε's will give larger $\bar{n}$'s.

As expected, converging sequences are Cauchy sequences.

Lemma 2.9. *Let $(a_n)_{n\in\mathbb{N}}$ be a converging sequence in a metric space. Then it is a Cauchy sequence.*

The proof is left as an exercise for the reader.

For Cauchy sequences, it is sufficient to check the convergence to a limit of a subsequence in order to deduce the convergence of the entire sequence.

Lemma 2.10. *Let $(a_n)_{n\in\mathbb{N}}$ be a Cauchy sequence in a metric space such that there exists a converging subsequence. Then, the entire sequence is converging.*

Proof. Let $(a_{n_i})_{i\in\mathbb{N}}$ be the converging subsequence, and let $a \in X$ be its limit. We want to prove that

$$\lim_{n\to\infty} \mathrm{d}(a_n, a) = 0.$$

For, we will show that for every given $\varepsilon > 0$, we can find $\overline{n} \in \mathbb{N}$ such that

$$\mathrm{d}(a_n, a) < \varepsilon, \tag{2.2}$$

for all $n \geq \overline{n}$. Fix $\varepsilon > 0$. Since $(a_{n_i})_{i\in\mathbb{N}}$ converges to a, there exists $i_0 \in \mathbb{N}$ such that

$$\mathrm{d}(a_{n_i}, a) < \frac{\varepsilon}{2}, \tag{2.3}$$

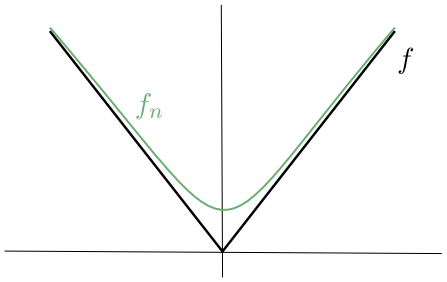

FIGURE 6. The C^1 approximation of the absolute value.

for all $i \geq i_0$. Moreover, since $(a_n)_{n\in\mathbb{N}}$ is a Cauchy sequence, let $n_0 \in \mathbb{N}$ be such that

$$\mathrm{d}(a_n, a_m) < \frac{\varepsilon}{2},$$

for all $n, m \geq n_0$. Let $\overline{n} \coloneqq \max\{n_{i_o}, n_0\}$. Therefore, if $n \geq \overline{n}$, by the triangle inequality and using (2.2) and (2.3) we get

$$\mathrm{d}(a_n, a) \leq \mathrm{d}(a_n, a_{n_i}) + \mathrm{d}(a_{n_i}, a) < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

This concludes the proof. □

Intuitively, we would expect also the opposite to hold, since if all elements of a sequence become closer and closer, then the sequence must converge to some limiting point. This poses the question: do all Cauchy sequences converge to a limit? The answer is no! For instance, consider the set $X \coloneqq \mathbb{R}^2 \setminus \{0\}$ endowed with the Euclidean metric Then, the sequence $(a_n)_{n\in\mathbb{N}}$ defined as

$$a_n \coloneqq \left(\frac{1}{n+1}, 0\right)$$

is a Cauchy sequence, but does not have a limit in X. The problem is that the expected limiting point does not belong to X. This suggests to give a name to those metric spaces in which Cauchy sequences admit a limit.

Definition 2.11. A metric space where all Cauchy sequences admit a limit is called *complete*. A normed space where all Cauchy sequences admit a limit is called *a Banach space*. An inner product space where all Cauchy sequences admits a limit is called *an Hilbert space*.

How does a metric space which is not complete look like? Well, we might think about it as having *holes*. This can be seen by considering the example above, the punctured $\mathbb{R}^2$, where the *missing point* is evident, but it can be more subtle, especially in infinite dimensional vector spaces. For instance, consider the space $C^1([-1,1])$ endowed with the C^0 norm (see Example 1.15). Then, the sequence $(f_n)_{n\in\mathbb{N}}$ given by (see Figure 6)

$$f_n(x) \coloneqq \sqrt{x^2 + \frac{1}{n}}$$

is a Cauchy sequence, but does not admit a limit in $C^1([-1,1])$. This is because the expected limiting object, namely $f(x) \coloneqq |x|$, is not in the space X. Even more dramatically, *each* continuous function on $[-1,1]$ can be approximated uniformly by a sequence of smooth functions (by using, for instance, the technique of *convolution*).

In this case, it is possible to *fill the holes*, by simply considering the space $C^0([-1,1])$. The same idea is behind the notion of *closure* of a space with respect to a distance. Note that if we took the C^1 norm (see Remark 1.36), the space $C^1([-1,1])$ would have no holes. This is another example that illustrates how a metric on a space determines its properties, namely how we look at the space itself.

Definition 2.12. Let (X, d) be a metric space, and let $Y \subset X$. We define the *closure* of Y in X with respect to the metric d, by

$$\overline{Y}^{\mathrm{d}} := \{a \in X \,:\, \exists\, (a_n)_{n\in\mathbb{N}} \subset Y \text{ s.t. } a_n \xrightarrow{\mathrm{d}} a\}.$$

Example 2.13 (L^p spaces)**.** For $p \in [1,\infty)$, consider the space X_p and the norm $\|\cdot\|_{L^p}$ defined in Example 1.14. Is the space $(X_p, \|\cdot\|_{L^p})$ a Banach space? The answer is no. Indeed, consider the sequence $(f_n)_{n\in\mathbb{N}}$ in X_p given by

$$f_n(x) := \begin{cases} 1 & \text{if } x \in \left[0, \frac{1}{2}\right], \\ -nx + \frac{n}{2} + 1 & \text{if } x \in \left[\frac{1}{2}, \frac{1}{2} + \frac{1}{n}\right], \\ 0 & \text{if } x \in \left[\frac{1}{2} + \frac{1}{n}, 1\right]. \end{cases}$$

Then, for $n, m \in \mathbb{N}$ (without loss of generality we can assume $m > n$) we have

$$\|f_n - f_m\|^p = \int_{1/2}^{\frac{1}{2}+\frac{1}{n}} |f_n(x) - f_m(x)|^p \,\mathrm{d}x \le \int_{1/2}^{\frac{1}{2}+\frac{1}{n}} |f_n(x)|^p \,\mathrm{d}x \le \frac{1}{n},$$

where in the first inequality we used the fact that $f_m \le f_n$, where the second follows from $|f_n(x)| \le 1$ for all $n \in \mathbb{N}$ and $x \in [0,1]$. Therefore, $(f_n)_{n\in\mathbb{N}}$ is a Cauchy sequence, but $(f_n)_{n\in\mathbb{N}}$ does not have a limit in X_p. Indeed, it should be the function $f : [0,1] \to \mathbb{R}$ given by

$$f(x) := \begin{cases} 1 & \text{if } x \in \left[0, \frac{1}{2}\right], \\ 0 & \text{if } x \in \left[\frac{1}{2}, 1\right], \end{cases} \tag{2.4}$$

Since f is not continuous, it does not belong to X_p. This raises the question of what is the closure of X_p in the L^p-norm? The resulting space is called *Lebesgue space* L^p, and it is the space of measurable[1] functions with finite p-moment. We will encounter this space later in the course.

Example 2.14 (Sobolev spaces)**.** For $p \in [1,\infty)$, let Z_p be the space of C^1 functions $f : (-1,1) \to \mathbb{R}$ such that

$$\|f\|_{W^{1,p}} := \left[\int_{-1}^{1} |f(x)|^p \,\mathrm{d}x + \int_{-1}^{1} |f'(x)|^p \,\mathrm{d}x\right]^{\frac{1}{p}} < \infty.$$

The function $\|\cdot\|_{W^{1,p}}$ is a norm, called the *Sobolev norm*. Is Z_p complete in this norm? The answer is no. Indeed, consider the sequence of functions $(f_n)_{n\in\mathbb{N}}$ given by

$$f_n(x) := \sqrt{x^2 + \frac{1}{n}}.$$

Then, it is easy to see that

$$\lim_{n\to\infty} \|f_n - f\|_{W^{1,p}} = 0,$$

where $f(x) := |x|$, but f does not belong to Z_p. The completion of Z_p is the Sobolev norm is denoted by $W^{1,p}((0,1))$, and it is called the *Sobolev space*. More interesting examples of functions that are in higher dimension, where everywhere unbounded functions belong to Sobolev spaces. Such spaces are nowadays used in several branches of Analysis and Engineering. In particular, they are the modern space used to study PDEs, both from an analytical and numerical point of view. You can discover more in the Master course on *Sobolev spaces and PDEs*, and in the Bachelor course *Numerical Methods for PDEs*, respectively.

[1] Here we are a bit unprecise in the meaning of *measurability* of a function. The correct notion to consider is that of Lebesgue measurability, that will be introduced later in the course.

Example 2.15 (Functions vanishing at infinity)**.** Consider the space $C_c(\mathbb{R}^N)$ of continuous functions $f : \mathbb{R}^N \to \mathbb{R}$ such that, for each of them, there exists $r > 0$ (not the same for all of functions in $C_c(\mathbb{R}^N)$) for which $f \equiv 0$ outside $B(0, r)$. The space $C_c(\mathbb{R}^N)$ is called the space of functions with *compact support*. We endow the space $C_c(\mathbb{R}^N)$ with the supremum norm. Is the space complete? The answer is no! The closure of the space $C_c(\mathbb{R}^N)$ in the supremum norm is denoted by $C_0(\mathbb{R}^N)$ and is called the space of functions *vanishing at infinity*. It is possible to see that a function $f : \mathbb{R}^N \to \mathbb{R}$ belongs to $C_0(\mathbb{R}^N)$ if and only if, for each $\varepsilon > 0$, there exists $r > 0$ such that $|f(x)| < \varepsilon$ for all $x \in \mathbb{R}^N$ with $|x| > r$.

Remark 2.16. Note that the closure of the same space with respect to different metrics will give different objects. For instance, the closure of $C^0([0, 1])$ with respect to the C^0 norm will be $C^0([0, 1])$ itself (we will prove it in a couple of classes), while its closure with respect to the L^p norm $W^{1,p}$, for some $p \in [1, \infty)$ will be a larger space. For instance, the function f defined in (2.4) is in $L^p([0, 1]) \setminus C^0([0, 1])$.

It is indeed the combination of the space X and the metric d that determines whether or not the space is complete.

2.2. **Compactness in metric spaces.** Not all sequences converge to a limit. Not even all sequences admit a converging subsequence. A question is whether there are sufficient conditions ensuring the existence of a converging subsequence. This will depend on a property of the *region* of the space the sequence is in.

Definition 2.17. Let (X, d) be a metric space. We say that a set $K \subset X$ is *sequentially compact* if every sequence $(a_n)_{n\in\mathbb{N}}$ with $a_n \in K$ has a subsequence converging to some $a \in K$.

We now want to investigate properties of compact sets. We will prove that, in a metric space, sequentially compact sets are bounded, closed, and complete.

Definition 2.18. Let (X, d) be a metric space, and $E \subset X$. We say that E is *bounded* if

$$\operatorname{diam}(E) \coloneqq \sup\{\mathrm{d}(x, y) : x, y \in E\} < \infty.$$

The above quantity is called the *diameter* of E.

Lemma 2.19. *A sequentially compact set in a metric space is bounded.*

Proof. Let $E \subset X$ be sequentially compact. Assume by contradiction that it is not bounded. We will construct a sequence $(a_n)_{n\in\mathbb{N}}$ with no converging subsequence. This sequence will be constructed inductively. Let $a_1 \in E$. Then, since we are assuming E to be not bounded, there exists $a_2 \in E$ with

$$\mathrm{d}(a_1, a_2) \geq 1.$$

Now, suppose we have constructed $a_1, \dots, a_n$. We choose $a_{n+1} \in E$ such that

$$\mathrm{d}(a_{n+1}, a_i) \geq 1, \qquad \text{for all } i \in \{1, \dots, n\}.$$

We repeat this process for every $n \in \mathbb{N}$. Note that

$$\mathrm{d}(a_i, a_j) \geq 1, \qquad \text{for all } i \neq j \in \mathbb{N}. \tag{2.5}$$

Now, the sequence $(a_n)_{n\in\mathbb{N}}$ does not have any converging subsequence. Indeed, if $(a_{n_i})_{i\in\mathbb{N}}$ was a converging subsequence, then by Lemma 2.9 it would be a Cauchy sequence. This is in contradiction with (2.5). Therefore, the assumption that E is not bounded is absurd. □

Lemma 2.20. *A sequentially compact set in a metric space is complete.*

The proof is left as an exercise for the reader.

Definition 2.21. A set $C \subset X$ is said to be *sequentially closed* if a sequence $(a_n)_{n\in\mathbb{N}}$, with $a_n \in C$, converges to some $a \in X$, then $a \in C$.

Lemma 2.22. *A sequentially compact set in a metric space is sequentially closed.*

The proof is left as an exercise for the reader.

We now want to understand, if the sequentially closed bounded sets are sequentially compact, since the former two are usually easier to verify than the latter. It turns out that this is the case for *finite* dimensional vector spaces.

Theorem 2.23 (Bolzano-Weierstraß Theorem)**.** *A set $K \subset \mathbb{R}^N$ is sequentially compact if and only if it is bounded and sequentially closed.*

Proof. Step 1. Let us prove that a sequentially compact set $K \subset \mathbb{R}^N$ is bounded and sequentially closed. Thanks to Lemma 2.19 we get that K is bounded. To prove that it is sequentially closed, let $(a_n)_{n\in\mathbb{N}}$ with $a_n \in K$ converging to some $a \in X$. Since K is sequentially compact, we get that there exists a subsequence $(a_{n_i})_{i\in\mathbb{N}}$ converging to some $b \in K$. Therefore, by Lemma 2.3 $a = b$, and $a \in K$.

Step 2. Let $K \subset \mathbb{R}^N$ be a sequentially closed bounded set. Let $(a_n)_{n\in\mathbb{N}}$ be a sequence in K. The idea is the following. You know from Analysis 1, that a sequence $(x_n)_{n\in\mathbb{N}}$ of elements of $\mathbb{R}$ has a converging subsequence (since you can extract a monotone sequence). Thus, by arguing componentwise, we can extract a finite number of subsequences ensuring the desired convergence.

We will construct a subsequence of indexes $(n_i)_{i\in\mathbb{N}}$ such that

$$\lim_{i\to\infty} a_{n_i} = a,$$

for some $a \in K$, as follows. Let $(n_i^1)_{i\in\mathbb{N}}$ be a subsequence of indexes such that there exist $a^1 \in \mathbb{R}$ with

$$\lim_{i\to\infty} a^1_{n_i^1} = a^1, \tag{2.6}$$

where $a^1_{n_i^1}$ is the first component of the vector $a_{n_i^1}$. Then, from the sequence of indexes $(n_i^1)_{i\in\mathbb{N}}$ it is possible to extract a subsequence, that for the sake of notation[2] we will denote by $(n_i^2)_{i\in\mathbb{N}}$, such that

$$\lim_{i\to\infty} a^2_{n_i^2} = a^2,$$

for some $a^2 \in \mathbb{R}$. Note that, since $(n_i^2)_{i\in\mathbb{N}}$ is a subsequence of $(n_i^1)_{i\in\mathbb{N}}$, from (2.6) we get that

$$\lim_{i\to\infty} a^1_{n_i^2} = a^1.$$

Arguing N times in a similar way, we find a sequence of indexes $(n_i^N)_{i\in\mathbb{N}}$, and $a^1, \dots, a^N \in \mathbb{R}$ such that

$$\lim_{i\to\infty} a^k_{n_i^N} = a^k,$$

for all $k \in \{1, \dots, N\}$. Defining $a \coloneqq (a_1, \dots, a_N) \in \mathbb{R}^N$, and setting $n_i \coloneqq n_i^N$ for all $i \in \mathbb{N}$, we get that

$$\lim_{i\to\infty} a_{n_i} = a,$$

as desired. □

Remark 2.24. Note that the proof we used to prove the Bolzano-Weierstraß Theorem does not hold in infinite dimensional vector spaces with a basis that is more than countable. Such vector spaces are, however, the interesting and useful ones.

Remark 2.25. Let $(X, \|\cdot\|)$ be a normed space. Is the closed unit ball

$$\{x \in X \,:\, \|x\| \leq 1\}$$

[2] Otherwise we would have had to denote it by n_{i_j}.

sequentially compact? The answer is, in general, no. Indeed, consider the space $X = C^0([0,1])$ endowed with the C^0 norm, and the sequence $(f_n)_{n\in\mathbb{N}}$ given by

$$f_n(x) \coloneqq \begin{cases} 1 & \text{if } x \in \left[0, \frac{1}{2(n+1)}\right], \\ -2n(n+1)x + (n+1) & \text{if } x \in \left[\frac{1}{2(n+1)}, \frac{1}{2n}\right], \\ 0 & \text{if } x \in \left[\frac{1}{2n}, 1\right]. \end{cases}$$

Then $\|f_n\|_{C^0} = 1$ for all $n \in \mathbb{N}$, and

$$\|f_n - f_m\|_{C^0} = 1,$$

for all $n \neq m$. Therefore, there cannot be any converging subsequence.

This might seem quite surprising, since we expect the closed unit ball to be the stereotypical example of compact set. This is the case for finite dimensional vector spaces (see Theorem 2.23). Nevertheless, infinite dimensional vector spaces are more complicated objects, where our finite-dimensional fails, and we need to appeal to the power of mathematics. This is quite an issue when we have a sequence of objects $(a_n)_{n\in\mathbb{N}}$ with uniformly bounded norm, and we would like, up to a subsequence, to claim that it has a limit. To *fix* such an issue, mathematicians developed the notion of *weak convergence*, which is related to the continuity of scalar linear maps defined on the space.

Finally, we introduce the topological notion of compactness, and we show that, in metric space, it coincides with that of sequential compactness.

Definition 2.26. Let X be a set, $B \subset X$. A family of sets $(A_i)_{i\in I} \subset X$, where I is any set of indexes, such that

$$B \subset \bigcup_{i\in I} A_i$$

is said to *cover* B.

Definition 2.27. Let (X, d) be a metric space. For $x \in X$, and $r > 0$, the set

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

is called the *(open) ball of center x and radius r*.

Definition 2.28. Let (X, d) be a metric space. We say that a set $K \subset X$ is *compact* if, for every family $(A_i)_{i\in I}$ of open balls that covers K, it is possible to extract a finite subfamily $A_{i_1}, \dots, A_{i_k}$ that covers K.

Remark 2.29. Note that the set of indexes I can be more than countable.

Lemma 2.30. *A compact set in a metric space is complete.*

The proof is left as an exercise to the reader.

An important property of spaces is the possibility to approximate every element of the space by using a fixed *countable* family of objects. The reason why we want to do it with at most countably many objects is because that is the practical limit of human operations: you do one thing, then another, then another, and so on, and they can be at most countably many. Moreover, this is also what (standard) computers can handle: finite number of operations. Therefore, if we know that in countably many steps we can converge to any object of our space, in a large, but finite number of steps, we can approximate any object with any degree of accuracy we want.

Definition 2.31. Let (X, d) be a metric space. A set $A \subset X$ is said to be *dense* if every point $x \in X$ is the limit of a sequence $(a_i)_{i\in\mathbb{N}} \subset A$.

Remark 2.32. As for other notions like compactness and continuity, there is a topological and a sequential notion of closure of a set $A \subset X$. The *topological* notion is the following: the set $\bar{A} \subset X$ is the smallest closed set that contains A. The *sequential* notion is the following: the set $\bar{A} \subset X$ is the set of all limiting points of sequences in A. If X is a metric space, the two notions coincide. Try to prove it!

Definition 2.33. Let (X, d) be a metric space. A subset $A \subset X$ is called *separable* if there exists a countable set $\{a_i\}_{i\in\mathbb{N}} \subset A$ that is dense in A.

Lemma 2.34. *A compact set in a metric space is separable.*

Proof. Let $K \subset X$ be a compact set. We will construct a dense set $\{x_i\}_{i\in\mathbb{N}} \subset K$ as follows: for each $k \in \mathbb{N}\setminus\{0\}$, we will construct a *finite* sequence of points $y_1^k, \dots, y_{n_k}^k \in K$ such that, for each $x \in K$, there exists one of them whose distance from x is less than $1/k$. The set $\{x_i\}_{i\in\mathbb{N}} \subset K$ will then be the union of all of these points:

$$\{x_i\}_{i\in\mathbb{N}} \coloneqq \bigcup_{k=1}^{\infty} \left\{y_1^k, \dots, y_{n_k}^k\right\}.$$

By construction, this set is countable, and dense in K.

So, fix $k \in \mathbb{N} \setminus \{0\}$, and consider the coverings of K given by

$$\left(B\left(x, \frac{1}{k}\right)\right)_{x\in K}.$$

Since K is compact, there exists finitely many points $y_1^k, \dots, y_{n_k}^k \in K$ such that

$$K \subset \bigcup_{i=1}^{n_k} B\left(y_i^k, \frac{1}{k}\right).$$

These points satisfy the property claimed above. □

In a metric spaces, compactness is equivalent to sequential compactness. This will turn out to be useful for arguments that we will use later in the course.

Theorem 2.35. *Let (X, d) be a metric space. Then, a set $K \subset X$ is sequentially compact if and only if it is compact.*

Proof. Step 1. Assume $K \subset X$ is compact. Assume by contradiction that it is not sequentially compact. We will construct a covering that does not admit any finite subcovering as follows. Let $(a_i)_{i\in\mathbb{N}} \subset K$ be a sequence that does not admit any converging subsequence.

We first claim that, for each $x \in X$, it is possible to find $r(x) > 0$ such that the ball $B(x, r(x))$ contains only a finite number of elements of $(a_i)_{i\in\mathbb{N}}$. Indeed, if there was a point $x \in X$ for which $B(x, 1/n)$ contains infinitely many elements of the sequence $(a_i)_{i\in\mathbb{N}}$, for all $n \in \mathbb{N} \setminus \{0\}$, then, by selecting an element a_{i_n} from each of such intersections, we would get that the subsequence $(a_{i_n})_{n\in\mathbb{N}}$ converges to x.

Now, by compactness, from the covering $(B(x, r(x)))_{x\in X}$ we can extract a finite subcovering $B(x_1, r(x_1)), \dots, B(x_k, r(x_k))$ of K. Since, by construction, each ball $B(x_i, r(x_i))$ intersects only finitely many elements of the sequence $(a_i)_{i\in\mathbb{N}}$, also

$$\bigcup_{i=1}^{k} B(x_i, r(x_i))$$

intersects only a finite number of a_i's. Since $(a_i)_{i\in\mathbb{N}} \subset K$, this is a contradiction with the fact that $B(x_1, r(x_1)), \dots, B(x_k, r(x_k))$ covers K.

Step 2. Assume $K \subset X$ is sequentially compact. Let $(A_i)_{i\in I}$ be an open covering of K, where I is infinite. Assume that it does not admit any finite subcover. The idea is to understand when this is possible, and to get a contradiction. It turns out that not admitting a finite covering is

possible only in two cases: the open set case, and the unbounded set case. Think about the sets $(0,1)$, and $(0,+\infty)$, respectively. Both will be incompatible with the sequential compactness of the set. To identify which case we are in, we look at the *size* of elements in the covering around points of the set. If the size is shrinking, we are in the first case, while if it is constant, we are in the second. Note that the two cases are not mutually exclusive.

For each $x \in X$, let

$$r(x) \coloneqq \sup\{r > 0 : B(x,r) \subset A_i, \text{ for some } i \in I\}.$$

Note that $r(x) \in (0,+\infty]$, since there exists at least a set A_i that contains x. Define

$$\overline{r} \coloneqq \inf\{r(x) : x \in X\}.$$

We distinguish two cases: $\overline{r} = 0$ (the open set case), and the case $\overline{r} > 0$ (the unbounded set case). In the former, it is possible to find a sequence $(x_n)_{n\in\mathbb{N}\setminus\{0\}}$ such that $r(x_n) < 1/n$ for all $n \in \mathbb{N}\setminus\{0\}$. Since K is sequentially compact, up to a subsequence (that, for the sake of notation we do not relabel), we have that $x_n \to \overline{x}$, for some $\overline{x} \in K$. Then, there exists $n_0 \in \mathbb{N}$ such that

$$x_n \in B\left(\overline{x}, \frac{r(\overline{x})}{2}\right),$$

for all $n \geq n_0$. Thus, $r(x_n) \geq r(\overline{x})/2$ for all $n \geq n_0$. This contradicts the assumption $\overline{r} = 0$.

Next, assume $\overline{r} > 0$. Let $x_1 \in K$. Then, since by assumption $B(x_1, \overline{r}/2)$ does not cover the entire K, it is possible to find $x_2 \in K \setminus B(x_1, \overline{r}/2)$. We now find, for each $k \in \mathbb{N}$, a point

$$x_k \in K \setminus \bigcup_{i=1}^{k-1} B\left(x_i, \frac{\overline{r}}{2}\right).$$

This, the sequence $(x_k)_{k\in\mathbb{N}}$ is such that $\mathrm{d}(x_i, x_j) \geq \overline{r}$, for all $i \neq j \in \mathbb{N}$. Thus, it cannot admit any converging subsequence, contradicting the sequential compactness of K.

Therefore, we get a contradiction in both cases, and therefore the absurd assumption cannot hold. This concludes the proof. □

Corollary 2.36. *A compact set in a metric space is complete, bounded, and sequentially closed.*

Corollary 2.37. *A sequentially compact set in a metric space is separable.*

2.3. **Comparison of metrics and norms.** The notion of convergence gives a way to understand a space with respect to the metric d, and to compare two metrics on the same space. Heuristically, since metrics are a way to give a notion of convergence, two metrics d_1 and d_2 are equivalent if a sequence converge to a certain limit with respect to d_1 if and only if it converges to the same limit with respect to d_2.

Definition 2.38. We say that two metrics d_1 and d_2 on a set X are *equivalent* if, for every $x \in X$ there exist $\alpha, \beta > 0$ such that

$$\alpha\, \mathrm{d}_1(x,y) \leq \mathrm{d}_2(x,y) \leq \beta\, \mathrm{d}_1(x,y),$$

for all $y \in X$. If the constants α and β are independent of $x \in X$, we say that the metrics d_1 and d_2 are *strongly equivalent*.

Remark 2.39. For normed spaces, the condition for strong equivalence of the metric induced by the norm writes as

$$\alpha\|x\|_1 \leq \|x\|_2 \leq \beta\|x\|_1,$$

for all $x \in X$.

An important result is that all norms are equivalent on finite dimensional vector spaces.

Theorem 2.40. *All norms on $\mathbb{R}^N$ are strongly equivalent.*

Proof. Let $\|\cdot\|_1$, and $\|\cdot\|_2$ be two norms on $\mathbb{R}^N$. Let

$$B_1 \coloneqq \{x \in \mathbb{R}^N \,:\, \|x\|_1 \leq 1\}.$$

Then, by using the homogeneity of the norm, we get that B_1 is bounded in the $\|\cdot\|_2$ norm. Moreover, B_1 is closed in the Euclidean topology. Therefore, by Theorem 2.23 we get that B_1 is sequentially compact with respect to $\|\cdot\|_2$. Finally, using the fact that the continuous function $f(x) \coloneqq \|x\|_2$ (see Lemma 3.8 and Remark 3.9) admits a minimum and a maximum on B_1 (thanks to Theorem 3.15), we get that there exists $0 < \alpha \leq \beta < \infty$ such that

$$\alpha \leq \|x\|_2 \leq \beta,$$

for all $x \in B_1$. By using the homogeneity of the norm, we conclude. □

Remark 2.41. The above result does not hold in *infinite* dimensional vector spaces. An example (prove it!) is given by the C^0 and the C^1 norms on $C^1([0,1])$.

Remark 2.42. A similar results does not hold for metrics. Even in finite dimensional vector spaces, not all metrics are equivalent. For instance, in $\mathbb{R}$, consider the distances

$$\mathrm{d}_1(x,y) \coloneqq \begin{cases} 0 & \text{if } x = y, \\ 1 & \text{else,} \end{cases} \qquad\qquad \mathrm{d}_2(x,y) \coloneqq |x-y|.$$

Then, the two metrics are not equivalent (prove it!).

3. Continuous functions

Continuity, despite being a nowadays common notion, is not a trivial notion at all. Indeed, historically, it took a while for mathematicians to develop the correct notion of continuity. In particular, before the 20th century, the mathematical community had only a vague notion of continuity, mostly based on the Greek notion of *continuous variations*. In particular, mathematicians in the 18th century and beginning of the 19th century used infinitesimal analysis (nowadays called *non-standard analysis*, made rigorous by Robinson in the 60s). Mathematicians like Hermann Grassmann (in his *Ausdehnugslehre* - first edition 1844, second edition 1862), and Augustin-Louis Cauchy (in his *Course d'analyse* of 1821) proved theorems regarding the continuity of separately continuous and of linearly continuous functions. A function $f : \mathbb{R}^2 \to \mathbb{R}$ is said to be separately continuous at a point $x_0 \in \mathbb{R}^2$ if it is continuous along lines parallel to the orthogonal axes, while it is said to be *linearly continuous* if it continuous on every line passing by x_0. Already in 1870 Thomae presented a counterexample to the above claims in the case of separately continuous when standard analysis due to Heine, that considered the function $f : \mathbb{R}^2 \to \mathbb{R}$ given by

$$f(x,y) \coloneqq \begin{cases} \sin\left(4 \arctan \dfrac{x}{y}\right) & \text{if } y \neq 0, \\ \\ 0 & \text{else.} \end{cases}$$

Moreover, in the case of linearly continuous, the counterexample is the one given by Peano in 1884 (presented in his treatise on calculus of 1884): consider the function $f : \mathbb{R}^2 \to \mathbb{R}$ given by

$$f(x,y) \coloneqq \begin{cases} \dfrac{xy^2}{x^2+y^4} & \text{if } (x,y) \neq (0,0), \\ \\ 0 & \text{else.} \end{cases}$$

Then, f is continuous along every line passing through the origin, but it is not continuous.

All of the above examples, forced mathematicians (in particular analysts) to adopt more rigor in their investigations. It is indeed at the beginning of last century, that a huge work was done in order to lay down solid foundations for mathematics based on set theory. You will see the outcome of such enterprise in the course *Logic*. Using that work, Peano was able to provide the first axiomatization of natural numbers and to give the modern definition of vector spaces.

In this chapter, we will focus on continuous maps between metric spaces. Continuous functions form one of the easiest examples of a class of functions enjoying some regularity properties. In particular, they are extremely useful in the study of properties of metric spaces. We will also investigate two notions of convergence of sequences of continuous functions, pointwise and uniform convergence, as well as their relation, and the properties of limits of sequences with respect to each of the two notions of convergence.

3.1. Continuity in metric spaces.

As for closedness and compactness, there is a sequential and a topological notion of convergence. In metric spaces, we will show that they coincide.

Definition 3.1. Let $f : X \to Y$ be a function between two metric spaces (X, d_1) and (Y, d_2). Let $\bar{x} \in X$. We say that f is *continuous at* $\bar{x}$ with respect to the metrics d_1 and d_2, if for each $\varepsilon > 0$ there exists $\delta > 0$ such that

$$\mathrm{d}_2\left(f(x), f(\bar{x})\right) < \varepsilon$$

for all $x \in X$ with $\mathrm{d}_1(x, \bar{x}) < \delta$. If a function is continuous at each point, we say that it is *continuous*.

Definition 3.2. Let $f : X \to Y$ be a function between two metric spaces (X, d_1) and (Y, d_2). Let $\bar{x} \in X$. We say that f is *sequentially continuous at* $\bar{x}$ with respect to the metrics d_1 and d_2, if for any sequence $(x_n)_{n\in\mathbb{N}}$ in X with $\lim_{n\to\infty} x_n = \bar{x}$ it holds

$$\lim_{n\to\infty} f(x_n) = f(\bar{x}).$$

If a function is sequentially continuous at each point $x \in X$, we say that it is *sequentially continuous* in X.

Remark 3.3. Note that the condition for sequential continuity of a function $f : X \to Y$ at a point $\bar{x} \in X$ writes as

$$\lim_{n\to\infty} \mathrm{d}_1(x_n, \bar{x}) = 0 \qquad \Rightarrow \qquad \lim_{n\to\infty} \mathrm{d}_2\left(f(x_n), f(\bar{x})\right) = 0,$$

for all sequences $(x_n)_{n\in\mathbb{N}}$ in X. Equivalently,

$$\lim_{n\to\infty} f(x_n) = f\left(\lim_{n\to\infty} x_n\right),$$

for any converging sequence $(x_n)_{n\in\mathbb{N}}$.

Remark 3.4. Unless otherwise specified, Euclidean spaces are always endowed with the Euclidean metric.

Remark 3.5. Whether or not a function $f : X \to Y$ is continuous *depends on both* of the metrics d_1 and d_2. Different choices of metric might change the fact that a function is continuous or not. For instance, consider the function $f : \mathbb{R} \to \mathbb{R}$ given by

$$f(x) \coloneqq \begin{cases} 1 & \text{if } x = 0, \\ 0 & \text{else.} \end{cases}$$

Then, f is continuous if in the domain we consider the metric d defined in (1.17), but it is not continuous if we consider the Euclidean metric.

Remark 3.6. Also the set where the function is defined influences whether or not a function is continuous. For instance, the function $f : [0, 1] \to \mathbb{R}$ defined as

$$f(x) \coloneqq \mathbb{1}_{\mathbb{Q}\cap[0,1]}(x) \coloneqq \begin{cases} 1 & \text{if } x \in \mathbb{Q}, \\ 0 & \text{else,} \end{cases}$$

is continuous on the set $\mathbb{Q} \cap [0, 1]$, continuous on the set $[0, 1] \setminus \mathbb{Q}$, but not continuous on $[0, 1]$.

In general topological spaces the two notions are different; in particular, continuity is stronger than sequential continuity. Next result will show that in metric spaces the two notion of convergence coincide.

Theorem 3.7. *A function $f : X \to Y$ between two metric spaces (X, d_1), and (Y, d_2) is continuous at a point $\bar{x} \in X$ if and only if it is sequentially continuous at $\bar{x}$.*

Proof. Step 1. Assume that f is continuous at $\bar{x} \in X$. Fix $\varepsilon > 0$. By continuity of f, there exists $\delta > 0$ such that

$$\mathrm{d}_2\left(f(x), f(\bar{x})\right) < \varepsilon, \tag{3.1}$$

for all $x \in X$ with

$$\mathrm{d}_1(x, \bar{x}) < \delta. \tag{3.2}$$

Let $(x_n)_{n\in\mathbb{N}}$ be a sequence X with $\lim_{n\to\infty} x_n = \bar{x}$, namely such that

$$\lim_{n\to\infty} \mathrm{d}_1(x_n, \bar{x}) = 0. \tag{3.3}$$

From (3.3) we get that there exists $\bar{n} \in \mathbb{N}$ such that

$$\mathrm{d}_1(x_n, \bar{x}) < \delta,$$

for all $n \geq \bar{n}$. Therefore, for all $n \geq \bar{n}$ we get

$$\mathrm{d}_2\left(f(x_n), f(\bar{x})\right) < \varepsilon.$$

This proves that

$$\lim_{n\to\infty} f(x_n) = f(\bar{x}),$$

thus that f is sequentially continuous at $\bar{x}$.

Step 2. Assume that f is sequentially continuous. Assume by contradiction that it is not continuous. In particular, this implies that there exists $\varepsilon > 0$ with the following property: for each $n \in \mathbb{N}$, there exists $x_n \in X$ with

$$\mathrm{d}_1(x_n, \bar{x}) < \frac{1}{n}, \qquad \mathrm{d}_2\left(f(x_n), f(\bar{x})\right) > \varepsilon.$$

Therefore, the sequence $(x_n)_{n\in\mathbb{N}}$ converges to $\bar{x}$, but the sequence $(f(x_n))_{n\in\mathbb{N}}$ does not converge to $f(\bar{x})$. This contradicts the sequential continuity of f. □

An example of continuous map in a metric space is the distance from a fixed point.

Lemma 3.8. *Let (X, d) be a metric space, and $\bar{x} \in X$. Then, the function $f : X \to [0, \infty)$ given by*

$$f(x) \coloneqq \mathrm{d}(x, \bar{x})$$

is continuous.

The proof is left as an exercise for the reader.

Remark 3.9. For a normed space $(X, \|\cdot\|)$, the above results says the the map

$$x \mapsto \|x\|$$

is continuous in the metric induced by the norm. Similarly, for an inner product space $(X, \langle\cdot,\cdot\rangle)$, we get that, for any fixed $v \in V$, the map

$$w \mapsto \langle v, w\rangle$$

is continuous in the metric induced by the inner product.

Level sets, as well as sub and sup-level sets of continuous functions enjoy nice properties.

Lemma 3.10. *Let (X, d_1), (Y, d_2) be metric spaces, and let $f : X \to Y$ be continuous. Then, for each $y \in Y$, the level set*

$$\{x \in X \,:\, f(x) = y\}$$

is sequentially closed. Moreover, if $Y = \mathbb{R}$, and d_2 is the Euclidean metric, then the sub and sup-level sets

$$\{x \in X \,:\, f(x) < y\}, \qquad \{x \in X \,:\, f(x) > y\}$$

are open.

Remark 3.11. The above properties are not true if the function is not continuous. Can you find counterexamples?

If more metric spaces are involved, it is useful to obtain the continuity of the composition of functions by the continuity of each function.

Lemma 3.12. *Let (X, d_1), (Y, d_2), and (Z, d_3) be metric spaces. Let $f : X \to Y$, and $g : Y \to Z$ be continuous functions. Then, the composition $g \circ f : X \to Z$ is continuous.*

The proof is left as an exercise to the reader.

Remark 3.13. Note that the opposite is not true: if $g\circ f$ is continuous, then we cannot conclude anything about the continuity of neither f, nor g (Find an example!).

An important property of continuity is that it preserves compactness.

Proposition 3.14. *Let $f : X \to Y$ be a continuous functions between two metric spaces (X, d_1), and (Y, d_2). Let $K \subset X$ be a compact set. Then, $f(K) \subset Y$ is compact.*

Proof. Recall that, by Theorem 2.35, compactness is equivalent to sequential compactness. Let $(y_n)_{n\in\mathbb{N}}$ be a sequence in $f(K)$. For each $n \in \mathbb{N}$, let $x_n \in K$ such that $f(x_n) = y_n$. Note that we are not claiming uniqueness of such points x_n's, since we do not know whether or not the function f is injective. Then, by the compactness of K, there exist a subsequence $(x_{n_i})_{i\in\mathbb{N}}$ and $\bar{x} \in K$ such that

$$\lim_{i\to\infty} x_{n_i} = \bar{x}.$$

By the sequential continuity of f, we get that

$$\lim_{i\to\infty} f(x_{n_i}) = f(\bar{x}).$$

This proves that the subsequence $(y_{n_i})_{i\in\mathbb{N}}$ converges to the point $\bar{y} \in K$, where $\bar{y} \coloneqq f(\bar{x})$. □

As a corollary of the previous result, we get that a scalar function on a compact set achieves its maximum and its minimum. This result is very useful in Analysis.

Theorem 3.15 (Weierstraß Theorem)**.** *Let $f : X \to \mathbb{R}$ be a continuous function, and let $K \subset X$ be a compact set. Then, f achieves its minimum and its maximum on K.*

Proof. By using Proposition 3.14, we get that $f(K)$ is sequentially compact. Then, from Theorem 2.23, we know that $f(K)$ is bounded and sequentially closed. Therefore,

$$m \coloneqq \min\{x \in \mathbb{R} : x \in f(K)\}, \qquad M \coloneqq \max\{x \in \mathbb{R} : x \in f(K)\},$$

are well defined. Indeed, since $f(K)$ is bounded, we get

$$-\infty < \inf\{x \in \mathbb{R} : x \in f(K)\} \le \sup\{x \in \mathbb{R} : x \in f(K)\} < +\infty.$$

Moreover, infimum and supremum are attained because $f(K)$ is sequentially closed. Thus, there exist $x, y \in K$ such that $f(x) = m$, and $f(y) = M$. These are the point of minimum and maximum of f on K respectively. □

Remark 3.16. It is easy to see that the above result does not hold if both the continuity of f and the compactness of K are not in force.

We now introduce a stronger notion of continuity.

Definition 3.17. Let $f : X \to Y$ be a function between two metric space (X, d_1) and (Y, d_2). We say that f is *uniformly continuous* if for every $\varepsilon > 0$ there exists $\delta > 0$ such that

$$\mathrm{d}_2\left(f(x), f(y)\right) < \varepsilon,$$

for all $x, y \in X$ with $\mathrm{d}_1(x, y) < \delta$.

Remark 3.18. The difference with the notion of continuity is that, for each $\varepsilon > 0$, the number $\delta > 0$ is the same for all $x \in X$, while for continuity it might depend on the point $x \in X$.

In particular, a uniformly continuous function is continuous. The opposite is not true. Indeed, consider the function $f(x) \coloneqq x^2$ for $x \in \mathbb{R}$. Then, f is continuous, but not uniformly continuous.

Example 3.19. The function $f(x) \coloneqq \sin(x)$, for $x \in \mathbb{R}$, is uniformly continuous. Is the function $x \mapsto \sqrt{x}$ uniformly continuous in $[0, 1]$?

We now investigate the relation between the two notions of continuity. We have already seen examples of continuous functions that are not uniformly continuous. For instance, the function $f(x) \coloneqq x^2$ for $x \in \mathbb{R}$. The issue with such example, is that the so called *modulus of continuity* of f (namely the δ related to the ε in the definition of continuity) becomes larger and larger as $|x| \to \infty$. Next result will show that this is the only case where things can go wrong. Indeed, on compact sets, continuity is equivalent to uniform continuity.

Proposition 3.20. *Let (X, d_1) and (Y, d_2) be two metric spaces, and let $K \subset X$ be a compact set. Then, $f : K \to Y$ is continuous if and only if it is uniformly continuous.*

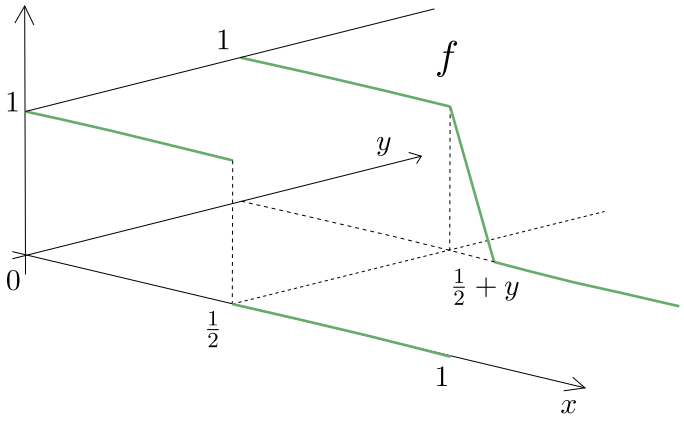

FIGURE 7. The function $f : (0,1)^2 \to \mathbb{R}$ of the example: it becomes less and less continuous as $y \to 0$.

Proof. Since uniformly continuous functions are continuous, we only need to prove the opposite implication. Let $f : K \to Y$ be a continuous function. Assume by contradiction that f is not uniformly continuous. Then, there exists $\varepsilon > 0$ and, for each $n \in \mathbb{N} \setminus \{0\}$, points $x_n, y_n \in K$ with

$$\mathrm{d}_1(x_n, y_n) < \frac{1}{n} \tag{3.4}$$

and

$$\mathrm{d}_2\left(f(x_n), f(y_n)\right) > \varepsilon. \tag{3.5}$$

Consider the sequences $(x_n)_{n\in\mathbb{N}}$ and $(y_n)_{n\in\mathbb{N}}$. Then, since K is compact, and hence sequentially compact (see Theorem 2.35), it is possible to extract a subsequence (why the same?) of indexes $(n_i)_{i\in\mathbb{N}}$, and find point $\bar{x}, \bar{y} \in K$ such that

$$\lim_{i\to\infty} x_{n_i} = \bar{x}, \qquad \lim_{i\to\infty} y_{n_i} = \bar{y}.$$

By using (3.4) we get that $\bar{x} = \bar{y}$. By continuity of f, this implies that

$$\lim_{i\to\infty} f(x_{n_i}) = f(\bar{x}) = f(\bar{y}) = \lim_{i\to\infty} f(y_{n_i}).$$

This is in contradiction with (3.5). □

We now want to investigate if, given a continuous function $f : A \to Y$, where A is a subset of the metric space X, it is possible to extend it to a continuous function defined in the whole space X. In general, this is not possible! Indeed, consider the function $f : (0,1)^2 \to \mathbb{R}$ defined as (see Figure 7)

$$f(x,y) \coloneqq \begin{cases} 1 & \text{if } x \in \left(0, \frac{1}{2}\right), \\ -\dfrac{x}{y} + 1 + \dfrac{1}{2y} & \text{if } x \in \left[\frac{1}{2}, \frac{1}{2} + y\right), \\ 0 & \text{if } x \in \left[\frac{1}{2} + y, 1\right). \end{cases}$$

What goes wrong is that the function becomes *less and less continuous* close to the boundary of A. Technically speaking, consider, for each $y \in (0,1)$, the restrictions $x \mapsto f(x,y)$. Such a function is continuous. Thus, for each $\varepsilon > 0$, there exists $\delta(y) > 0$ such that

$$|x_1 - x_2| < \delta(y) \qquad \Rightarrow \qquad |f(x_1) - f(x_2)| < \varepsilon.$$

Note that we stress the dependence of such continuity parameter $\delta(y)$ on y. The issue is that $\delta(y) \to 0$ as $y \to 0$.

In order to avoid such a pathology, we can require the function f to be uniformly continuous on A.

Proposition 3.21. *Let (X, d_1), and (Y, d_2) be metric spaces, and let $A \subset X$. Let $f : A \to Y$ be a uniformly continuous map. Then, there exists a uniformly continuous function $\widetilde{f} : \overline{A} \to Y$ such that $\widetilde{f} = f$ on A. The map $\widetilde{f}$ is unique.*

The proof is left as an exercise for the reader.

What about functions $f : A \to Y$ that are only continuous? Is there the chance to extend them to a continuous function defined in the entire space? Let us consider two examples. Let $f(x) : (0,1) \to \mathbb{R}$ given by $f(x) \coloneqq x^{-1}$. It is easy to see that there is no way to extend it in a continuous way to $[0,1]$. What makes it impossible to extend f, is that f blows up when it approaches the boundary of the set A.

Another example is the function $f : (0,1) \to \mathbb{R}$ given by $f(x) \coloneqq \sin(1/x)$. In this case, the function f oscillates too much close to the boundary of A.

The examples above show that if the set A is not closed, then there might be issue with having a continuous extension. A remarkable result, is that it is possible though to extend in a continuous way a continuous function defined on a *closed* set, at least when the target space is $\mathbb{R}$. Usually, this results is presented directly on topological spaces, and its proof requires fine arguments in topology (the so called Urysohn's lemma). You will see such a proof in the course *Topology* next semester.

In metric spaces, there are more direct constructions for such extension. In particular, in 1907 Lebesgue provided an extension of a continuous function defined on a closed subset of $\mathbb{R}^2$, while in 1905 Tietze extended such a result to the case of a general metric space. Nowadays there are several proofs of such a result, as well as several variants.

Theorem 3.22 (Tietze extension theorem)**.** *Let (X, d) be a metric space, and let $C \subset X$ be a sequentially closed set. Let $f : C \to \mathbb{R}$ be a continuous map. Then, there exists a continuous function $\widetilde{f} : X \to \mathbb{R}$ such that $\widetilde{f} = f$ on C.*

(Sketch of the proof). We present here some definitions of the extension, leaving the (sometimes not trivial) details to the reader.

The extension provided by Tietze is the following: for each $x \in X \setminus C$, define

$$\widetilde{f}(x) \coloneqq \sup_{y \in C} \frac{f(y)}{(1 + [\,\mathrm{d}(x,y)]^2)^{\frac{1}{\mathrm{dist}(x,C)}}},$$

where $\mathrm{dist}(x, C) > 0$ denotes the distance of the point x from the set C, defined as

$$\mathrm{dist}(x, C) \coloneqq \inf\{\, \mathrm{d}(x,y) : y \in C\}.$$

Hausdorff, in 1919 gave an easier version of the extension, by considering the function

$$\widetilde{f}(x) \coloneqq \inf_{y \in C} \left[f(y) + \frac{\mathrm{d}(y,x)}{\mathrm{dist}(y,C)} - 1 \right].$$

Riesz, for functions $f : C \to [1,2]$, used the function

$$\widetilde{f}(x) \coloneqq \sup_{y \in C} f(y) \frac{\mathrm{dist}(y,C)}{\mathrm{d}(y,x)},$$

while Dieudonnè in 1960 proposed the function

$$\widetilde{f}(x) \coloneqq \inf_{y \in C} f(y) \frac{\mathrm{d}(y,x)}{\mathrm{dist}(y,C)}.$$

All of the above definitions might made you think *"How did these people come up with such functions?"*. Try to draw a figure of what is going on in the case $X = \mathbb{R}^2$, and this might give you a better idea of what is going on. □

By combining the above two results, it is possible to provide a continuous extension of a uniformly continuous function defined on any subset A.

Corollary 3.23. *Let (X, d) be a metric space, and let $A \subset X$. Let $f : A \to \mathbb{R}$ be a uniformly continuous map. Then, there exists a continuous function $\widetilde{f} : X \to \mathbb{R}$ such that $\widetilde{f} = f$ on A.*

Remark 3.24. Note that the extension provided by the above corollary is only continuous. We cannot make sure that it is uniformly continuous. Can you find an example of a uniformly continuous function $f : A \to \mathbb{R}$, for some subset A of a metric space X, such that it cannot be extended to a uniformly continuous function to the whole space X?

Finally, we investigate properties of continuous functions with values in a normed spaces, since the linear structure allows us to add functions. Moreover, if the target space is $\mathbb{R}$, there are other natural operations among functions that maintain continuity.

Proposition 3.25. *Let (X, d) be a metric space, and let $(Y, \|\cdot\|)$ be a normed vector space. Let $f, g : X \to Y$ be continuous functions. Then, the function $f + g$ is continuous.*

Moreover, if $Y = \mathbb{R}$, and $\lambda \in \mathbb{R}$, also the functions

$$fg, \qquad \frac{f}{g}, \qquad \lambda f$$

are continuous, where they are defined.

The proof of the above result is left to the reader.

Remark 3.26. The above operations among functions are *finite* operations. What about the following case: consider a family of functions $(f_i)_{i \in \mathbb{N}}$ from a metric space (X, d), to $\mathbb{R}$, where I is any set of indexes (even more than countable). Are the functions $F, G : X \to Y$ defined as

$$F(x) \coloneqq \inf_{i \in I} f_i(x), \qquad G(x) \coloneqq \sup_{i \in I} f_i(x)$$

continuous? What about the case where $I = \mathbb{N}$, and we consider the function

$$H(x) \coloneqq \sum_{i \in \mathbb{N}} f_i(x),$$

Is that continuous?

3.2. **Pointwise and uniform convergence.** There are several notion of convergence for functions, each more suitable for the particular situation under investigation. Here we study two important basic notions of convergence for functions in a metric space: pointwise and uniform convergence. The former is easy to check, but not powerful enough to ensure continuity of the limit of a sequence of continuous functions. For this reason, we need the latter, stronger, type of convergence.

Definition 3.27. For each $n \in \mathbb{N}$, let $f_n : X \to Y$ be a function between two metric spaces, and let $f : X \to Y$. We say that the sequence $(f_n)_{n \in \mathbb{N}}$ *converges pointwise* to f if

$$\lim_{n \to \infty} f_n(x) = f(x),$$

for each $x \in X$.

Remark 3.28. The condition for pointwise converges writes as follows: for each $x \in X$, and for each $\varepsilon > 0$, there exists $\bar{n} \in \mathbb{N}$ such that

$$\mathrm{d}_2 \left(f_n(x), f(x) \right) < \varepsilon,$$

for all $n \geq \bar{n}$.

Example 3.29. Let $f_n : (0, +\infty) \to \mathbb{R}$ be defined as $f_n(x) \coloneqq e^{-nx}$. Then, f_n converges pointwise to the function $f \equiv 0$.

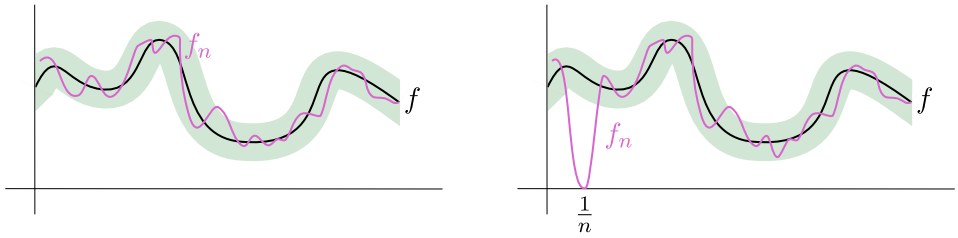

FIGURE 8. On the left: a sequence of functions $(f_n)_{n\in\mathbb{N}}$ converging uniformly to f will all have to lie inside the green region, for large enough n's. On the right: a sequence converging pointwise, but not uniformly.

Definition 3.30. For each $n \in \mathbb{N}$, let $f_n : X \to Y$ be a function between two metric spaces, and let $f : X \to Y$. We say that the sequence $(f_n)_{n\in\mathbb{N}}$ *converges uniformly* to f if for each $\varepsilon > 0$, there exists $\bar{n} \in \mathbb{N}$ such that

$$\mathrm{d}_2\left(f_n(x), f(x)\right) < \varepsilon,$$

for all $x \in X$ and $n \geq \bar{n}$.

Example 3.31. Let $f_n : \mathbb{R} \to \mathbb{R}$ be defined as $f_n(x) \coloneqq \frac{1}{n}\sin(x)$. Then, f_n converges uniformly to the function $f \equiv 0$.

Remark 3.32. Uniform convergece implies pointwise convergence, but the opposite is false. Indeed, consider the sequence $(f_n)_{n\in\mathbb{N}}$ defined in Example 3.29: it is converging pointwise, but not uniformly to the function $f \equiv 0$ (prove it!).

The difference between pointwise and uniform convergence is the following. Assume the sequence $(f_n)_{n\in\mathbb{N}}$ to converge pointwise to f. Fix $x \in X$. Then, for each $\varepsilon > 0$ there exists $\bar{n} \in \mathbb{N}$ such that

$$\mathrm{d}_2(f_n(x), f(x)) < \varepsilon \tag{3.6}$$

for all $n \geq \bar{n}$. This threshold $\bar{n}$ might change for point to point. In particular,

$$\sup_{x\in X} \{\, \bar{n} \in \mathbb{N} : (3.6) \text{ holds for all } n \geq \bar{n} \,\} \tag{3.7}$$

might be infinite. If, for each $\varepsilon > 0$, the quantity in (3.7) is not infinite, namely if it is possible to find a uniform threshold for all points $x \in X$ and all errors $\varepsilon > 0$, then the convergence is uniform (see Figure 8).

The notion of uniform convergence has been designed to preserve continuity.

Theorem 3.33. *Let $(f_n)_{n\in\mathbb{N}}$ be a sequence of functions converging uniformly to a function f. Let $\bar{x} \in X$. If each f_n is continuous at $\bar{x}$, then also f is continuous at $\bar{x}$. In particular, if each f_n is continuous, then also f is continuous.*

Proof. In order to prove that f is continuous, we argue as follows. Let $\bar{x} \in X$, and let $\varepsilon > 0$. By the uniform convergence of $(f_n)_{n\in\mathbb{N}}$ to f, there exists $\bar{n} \in \mathbb{N}$ such that

$$\mathrm{d}_2(f_n(x), f(x)) < \frac{\varepsilon}{3}, \tag{3.8}$$

for all $x \in X$, and $n \geq \bar{n}$. Fix $n \geq \bar{n}$. By continuity of f_n, there exists $\delta > 0$ such that

$$\mathrm{d}_2\left(f_n(x), f_n(\bar{x})\right) < \frac{\varepsilon}{3}, \tag{3.9}$$

for each $x \in X$ with $\mathrm{d}_1(x, \bar{x}) < \delta$. Therefore, from (3.8), (3.9), and by using the triangle inequality, we get

$$\begin{aligned}\mathrm{d}_2\left(f(x), f(\bar{x})\right) &\leq \mathrm{d}_2\left(f(x), f_n(x)\right) + \mathrm{d}_2\left(f_n(x), f_n(\bar{x})\right) + \mathrm{d}_2\left(f_n(\bar{x}), f(\bar{x})\right) \\ &\leq \frac{\varepsilon}{3} + \frac{\varepsilon}{3} + \frac{\varepsilon}{3} = \varepsilon,\end{aligned}$$

for all $x \in X$ with $\mathrm{d}_1(x, \bar{x}) < \delta$. This proves that f is continuous at $\bar{x}$. □

Remark 3.34. The pointwise limit of a sequence of continuous functions is not necessarily continuous. Indeed, consider the sequence

$$f_n(x) := \begin{cases} 1 & \text{if } x \in \left[0, \frac{1}{2}\right], \\ nx - \frac{n}{2} + 1 & \text{if } x \in \left[\frac{1}{2}, \frac{1}{2} + \frac{1}{n}\right], \\ 0 & \text{if } x \in \left[\frac{1}{2} + \frac{1}{n}, 1\right]. \end{cases}$$

Then, $(f_n)_{n\in\mathbb{N}}$ converges pointwise to the discontinuous function $f : [0,1] \to \mathbb{R}$ given by

$$f(x) := \begin{cases} 1 & \text{if } x \in \left[0, \frac{1}{2}\right], \\ 0 & \text{if } x \in \left(\frac{1}{2}, 1\right], \end{cases}$$

Note that a sequence of functions that are not continuous can converge to a continuous function. For instance, consider the sequence $f_n : \mathbb{R} \to \mathbb{R}$ given by

$$f_n(x) := \begin{cases} 0 & \text{if } x \leq 0, \\ \frac{1}{n} & \text{if } x > 0. \end{cases}$$

Then, despite each f_n not being continuous, the sequence $(f_n)_{n\in\mathbb{N}}$ converges uniformly to the continuous function $f \equiv 0$.

Moreover, uniform convergence to a continuous function allows to take the limit in a sequence of functions and a sequence of points at the same time.

Proposition 3.35. *Let $(f_n)_{n\in\mathbb{N}}$ be a sequence of functions from two metric spaces (X, d_1) and (Y, d_2) converging uniformly to a continuous $f : X \to Y$. Then,*

$$\lim_{n\to\infty} f_n(x_n) = f(\bar{x}),$$

for any sequence $(x_n)_{n\in\mathbb{N}} \subset X$ with $\lim_{n\to\infty} x_n = \bar{x}$.

The proof will be given as an exercise in the homework.

Remark 3.36. The above result is not true for sequences converging pointwise but not uniformly. Can you find a counterexample?

We now relate the notion of uniform convergence to convergence in the supremum norm.

Definition 3.37. Let (X, d_1) and (Y, d_2) be two metric spaces. On the space of functions $f : X \to Y$, we define the *uniform metric* d_∞ induced by d_2 as

$$\mathrm{d}_\infty(f, g) := \sup_{x\in X} \mathrm{d}_2(f(x), g(x)),$$

for $f, g : X \to Y$.

Remark 3.38. Note that the norm in the domain space does not influence the supremum norm. Moreover, this definition of supremum norm extends that given in Example 1.15.

Proposition 3.39. *Let $(f_n)_{n\in\mathbb{N}}$ be a sequence of functions from two metric spaces (X, d_1) and (Y, d_2), and $f : X \to Y$. Then $(f_n)_{n\in\mathbb{N}}$ converges to f uniformly if and only if it converges to f in the supremum norm.*

The proof is left as an exercise for the reader.

Remark 3.40. It can be shown that there exists a topology that induce the pointwise convergence. See Exercise 14.4.4 in the book *Analysis 2* by Terence Tao. Take a look at the file uploaded on Brightspace.

Finally, we consider the special case of *scalar* continuous functions, namely functions $f : X \to \mathbb{R}$. In this case, the fact that the target space $\mathbb{R}$ has *an order*, allows to talk about monotone sequences of functions.

Definition 3.41. Let $(f_n)_{n\in\mathbb{N}}$ be a sequence of functions $f_n : X \to \mathbb{R}$, where (X, d) is a metric space. We say that the sequence is *increasing* if, for all $x \in X$, it holds

$$f_n(x) \leq f_{n+1}(x),$$

for all $n \in \mathbb{N}$. We say that the function is *strictly increasing*, if the above inequality is strict.

Moreover, we say that the sequence is *decreasing* if, for all $x \in X$, it holds

$$f_n(x) \geq f_{n+1}(x),$$

for all $n \in \mathbb{N}$. We say that the function is *strictly decreasing*, if the above inequality is strict.

Finally, we say that the sequence is *monotone* if it is either monotone increasing or monotone decreasing.

A useful result, is that, for monotone sequences, pointwise convergence to a continuous function is actually a uniform convergence.

Theorem 3.42 (Dini's Theorem)**.** *Let $(f_n)_{n\in\mathbb{N}}$ be a monotone sequence of continuous functions $f_n : K \to \mathbb{R}$, where (X, d) is a metric space, and $K \subset X$ is compact. Assume that $(f_n)_{n\in\mathbb{N}}$ converges pointwise to a continuous function $f : K \to \mathbb{R}$. Then, the convergence is uniform.*

Proof. Without loss of generality, we can assume that the sequence is monotonically increasing. Assume by contradiction that the convergence is not uniform. Then, there exist $\varepsilon > 0$, an increasing sequence of indexes $(n_i)_{i\in\mathbb{N}}$, and a sequence of points $(x_i)_{i\in\mathbb{N}}$ such that

$$f(x_i) - f_{n_i}(x_{n_i}) > \varepsilon \tag{3.10}$$

for all $i \in \mathbb{N}$. Note that the above quantity is positive because the sequence is monotone. Since K is sequentially compact, up to extracting a subsequence (that we do not relabel), there exists $\bar{x} \in K$ such that $x_{n_i} \to \bar{x}$. Since f is continuous, there exists $\delta_1 > 0$ such that

$$|f(y) - f(\bar{x})| < \frac{\varepsilon}{2}, \tag{3.11}$$

for all $y \in K$ with $\mathrm{d}(y, \bar{x}) < \delta_1$. Moreover, by using the pointwise convergence of $(f_n)_n \in \mathbb{N}$ to f together with the monotonicity of the sequence, we get that there exists $\bar{n} \in \mathbb{N}$ such that

$$f(\bar{x}) - f_n(\bar{x}) < \frac{\varepsilon}{2}, \tag{3.12}$$

for all $n \geq \bar{n}$. Since $f_{\bar{n}}$ is continuous, there exists $\delta_2 > 0$ such that

$$|f_{\bar{n}}(y) - f_{\bar{n}}(\bar{x})| < \frac{\varepsilon}{2}, \tag{3.13}$$

for all $y \in K$ with $\mathrm{d}(y, \bar{x}) < \delta_2$. In particular, by using (3.11), (3.12), (3.13), and the fact that the sequence is monotone, we get that

$$f(y) - f_n(y) < \varepsilon, \tag{3.14}$$

for all $n \geq \bar{n}$, and all $y \in K$ with $\mathrm{d}(y, \bar{x}) < \delta$, where $\delta := \min\{\delta_1, \delta_2\}$. This is in contradiction with (3.10), for i large enough. □

Remark 3.43. The above result is not true if we drop the assumption of continuity, either for the functions in the sequence, or for the limiting function. Indeed, for each $n \in \mathbb{N} \setminus \{0\}$ consider the function $f_n : [-1, 1] \to \mathbb{R}$ defined as $f_n := \mathbb{1}_{[1/n,1]}$. Then, the monotone sequence $(f_n)_{n\in\mathbb{N}}$ converges pointwise, but not uniformly, to the function $f := \mathbb{1}_{[0,1]}$.

Moreover, it is possible to construct a sequence of continuous functions converging pointwise, but not uniformly, to the function f.

Finally, also the compactness of the domain of the function is crucial for the validity of the result. To see that, consider the functions $f_n : (0,1) \to \mathbb{R}$ defined as

$$f_n(x) \coloneqq \begin{cases} nx & \text{if } x \in \left(0, \frac{1}{n}\right), \\ 1 & \text{else.} \end{cases}$$

Then, the monotone sequence $(f_n)_{n\in\mathbb{N}}$ converges pointwise, but not uniformly, to the function $f \coloneqq \mathbb{1}_{(0,1)}$.

Finally, we recall that uniform convergence allows to pass to the limit in integrals and derivatives.

Proposition 3.44. *Let $(f_n)_{n\in\mathbb{N}}$ be a sequence of scalar continuous functions on $[0,1]$ converging uniformly to f. Then,*

$$\lim_{n\to\infty} \int_0^1 f_n(x)\,dx = \int_0^1 f(x)\,dx.$$

The proof of the above result was presented in Analysis 1.

Proposition 3.45. *Let $(f_n)_{n\in\mathbb{N}}$ be a sequence of scalar differentiable functions on $[0,1]$ converging pointwise to f, and such that the sequence $(f_n')_n$ converges uniformly to g. Then, f is differentiable, and $f' = g$.*

Try to prove the above proposition.

Remark 3.46. The above results do not hold if the sequence converges only pointwise, but not uniformly. Find counterexamples!

4. The space of continuous functions

In this section we study an important example of a functional space: the space of bounded continuous functions. This is an important example of *functional space*, namely a space of functions satisfying certain properties. Functional spaces are ubiquitous in almost all areas of modern mathematics: Lipschitz space to Hölder space, Sobolev space, functions with bounded variations, to mention some. These are where old and new problems in mathematics and physics found the proper framework to be stated and solved, from PDEs, to machine learning, from minimization problems, to quantum mechanics. We will investigate three main properties of the space of bounded continuous functions: completeness, characterization of compact sets, and density of a special class of functions (polynomial functions). For the latter two properties, we will restrict our attention to the case of scalar functions, to focus on the main ideas without getting lost in technical details.

4.1. Completeness.

We start by investigating the space of bounded functions.

Definition 4.1. Let (X, d_1) and (Y, d_2) be metric spaces. We define the set of *bounded functions* as

$$B(X,Y) \coloneqq \{f : X \to Y \,:\, \operatorname{diam}(f(X)) < \infty\},$$

When the target space is $\mathbb{R}$ with the Euclidean metric, we will simply denote the space by $B(X)$.

Remark 4.2. Note that $f \in B(X,Y)$ if and only if, for fixed $y \in Y$, there exists $R > 0$ such that

$$\mathrm{d}_2(f(x), y) < R,$$

for all $x \in X$.

Remark 4.3. Note that functions in $B(X,Y)$ are not necessarily continuous. Moreover, it is easy to see that $B(X,Y)$ is closed with respect to the uniform convergence. Namely, if $(f_n)_{n\in\mathbb{N}} \subset B(X,Y)$ converges uniformly to a function $f : X \to Y$, then $f \in B(X,Y)$. Is $B(X,Y)$ also closed with respect to the pointwise convergence?

As a first result, we show that the space of bounded functions is complete with respect to the supremum norm.

Proposition 4.4. *Let (X, d_1) be a metric space, and let (Y, d_2) be a complete metric space. Then, the space $B(X,Y)$ is a complete metric space with respect to the supremum norm.*

Proof. Let $(f_n)_{n\in\mathbb{N}}$ be a Cauchy sequence in $B(X,Y)$ with respect to the supremum norm (see Definition 3.37). Fix $\varepsilon > 0$. Then, there exists $\bar{n} \in \mathbb{N}$ such that

$$\mathrm{d}_2\left(f_n(x), f_m(x)\right) < \varepsilon. \tag{4.1}$$

for all $x \in X$, and all $n, m \geq \bar{n}$. From (4.1) we get that, for each $x \in X$, the sequence $(f_n(x))_{n\in\mathbb{N}}$ is a Cauchy sequence. Since Y is complete, the sequence admits a limit, denoted by $f(x)$. This defines a function $f : X \to Y$.

We now claim that $(f_n)_{n\in\mathbb{N}}$ converges uniformly to f. Fix $\varepsilon > 0$. By using (4.1) we get that there exists $\bar{n} \in \mathbb{N}$ such that

$$\mathrm{d}_2\left(f_n(x), f_m(x)\right) < \varepsilon,$$

for all $x \in X$, and all $n, m \geq \bar{n}$. Thus, by sending $n, m \to \infty$ in the above equation, we get

$$\mathrm{d}_2\left(f(x), f(x)\right) < \varepsilon,$$

for all $x \in X$.

Finally, to prove that f is bounded, we argue as follows. Since $f_{\bar{n}}$ is bounded, it follows that, fixed $y \in Y$, there exist $M > 0$ such that

$$\mathrm{d}_2(f_n(x), y) \leq M, \tag{4.2}$$

for all $x \in X$. Thanks to (4.2), we have that $f \in B(X,Y)$. □

Remark 4.5. If Y is a vector space, then also $B(X,Y)$ is.

We now investigate two families of continuous functions.

Definition 4.6. Let (X, d_1) and (Y, d_2) be metric spaces. We define

$$C^0(X,Y) \coloneqq \{f : X \to Y : f \text{ is continuous }\},$$

and

$$C^0_b(X,Y) \coloneqq \{f \in B(X,Y) : f \text{ is continuous }\}.$$

In case the target space is $\mathbb{R}$ with the Euclidean metric, we denote the above spaces by $C^0(X)$ and $C^0_b(X)$ respectively.

Remark 4.7. In general, $C^0_b(X,Y)$ is a proper subset of $C^0(X,Y)$. If X is compact, then $C^0_b(X,Y) = C^0(X,Y)$.

By using Theorem 3.33 and a similar argument as those used in the proof of Proposition 4.4, we get that both spaces are closed with respect to the uniform convergence. Moreover, they are also complete.

Proposition 4.8. *Let (X, d_1) be a metric space, and let (Y, d_2) be a complete metric space. Then, the space $C^0(X,Y)$, and the space $C^0_b(X,Y)$ are complete with respect to the supremum norm.*

Proof. Let $(f_n)_{n\in\mathbb{N}}$ be a Cauchy sequence in $C^0_b(X,Y)$. Thanks to Proposition 4.4 we know that it converges in the supremum norm to a function $f \in C_b(X;Y)$. Thanks to Theorem 3.7, we know that f is continuous. This concludes the proof. A similar argument proves the result for $C^0(X,Y)$. □

4.2. **Characterization of compact sets: the Ascoli-Arzelà Theorem.** As we have seen, compact sets play an important role in metric spaces and for continuous functions between them. We now investigate a characterization of sequentially compact sets in $C^0_b(X)$ known as the Ascoli-Arzelà Theorem[3]. The importance of the Ascoli-Arzelà Theorem is that it ensure that, up to a subsequence, a sequence of equibounded and equicontinuous functions converge to a limiting function. As an application of such result, we will sketch the proof of the Peano Theorem on the existence of solutions to an Ordinary Differential Equation.

For, we need to introduce three notions.

Definition 4.9. Let (X, d_1), and (Y, d_2) be two metric spaces, and let $\mathcal{F} \subset C^0(X,Y)$. We say that the family $\mathcal{F}$ is *equicontinuous* if, for every $\varepsilon > 0$, there exists $\delta > 0$ such that

$$\mathrm{d}_2\left(f(x), f(y)\right) < \varepsilon,$$

for all $x, y \in X$ with $\mathrm{d}_1(x,y) < \delta$, and all $f \in \mathcal{F}$.

Remark 4.10. The above definition means that all functions in an equicontinuous family have the same modulus of continuity. A *modulus of continuity* for a function $f : X \to Y$ between two metric spaces (X, d_1) and (Y, d_2) is a monotone function $\omega : [0,\infty) \to (0,\infty)$, such that

$$\mathrm{d}_2\left(f(x), f(y)\right) \leq \omega\left(\mathrm{d}_1(x,y)\right),$$

for all $x, y \in X$.

Definition 4.11. Let (X, d_1) and (Y, d_2) be two metric spaces, and let $\mathcal{F} \subset C^0(X,Y)$. Let $\mathcal{F} \subset C^0(X,Y)$. We say that the family $\mathcal{F}$ is *equibounded* if there exists $D < \infty$ such that

$$\operatorname{diam}\left(f(X)\right) \leq D,$$

for all $f \in \mathcal{F}$.

[3]It might seem strange that the names are not in alphabetical order. The reason is that, in Italian (since both mathematicians are Italians), Arzelà-Ascoli sounds very strange, since accents tend to go at the end of the composite word.

We are now in position to prove one of the main results of this section. The sufficient condition for compactness was established by Ascoli in 1884, while Arzelà in 1895 proved also the necessity. Both of these results were for compact intervals of $\mathbb{R}$. The generalization to family of functions on compact metric spaces was obtained by Fréchet in 1906. As anticipated above, we will restrict our attention to the case of scalar functions, in order to grasp better the ideas behind the result.

Theorem 4.12 (Ascoli-Arzelà Theorem)**.** *Let* (X, d) *be a compact metric space. Let* $\mathcal{F} \subset C^0(X)$ *be a closed set with respect to the uniform convergence. Then,* $\mathcal{F}$ *is compact in the uniform norm if and only if* $\mathcal{F}$ *is equibounded and equicontinuous.*

Proof. Necessity. Assume the family $\mathcal{F}$ to be compact with respect to the uniform convergence. Then, it is left as an exercise for the reader to check that it is equibounded and equicontinuous.

Sufficiency. Assume that $\mathcal{F}$ is equibounded and equicontinuous. Since $(C^0(X), \|\cdot\|_{C^0})$ is a metric space, we show that the family $\mathcal{F}$ is sequentially compact. Let $(f_n)_{n\in\mathbb{N}}$ be a sequence of functions in $\mathcal{F}$. We want to prove that it is possible to extract a subsequence $(f_{n_k})_{k\in\mathbb{N}}$ such that $f_{n_k} \to f$ uniformly, Note that, by using the fact that $\mathcal{F}$ is closed, we get that $f \in \mathcal{F}$. The idea of the proof is the following: given a point $x \in X$, we can extract a subsequence $(f_{n_k})_{k\in\mathbb{N}}$ such that $(f_{n_k}(x))_{k\in\mathbb{N}}$ converges to some value $f(x) \in \mathbb{R}$. The problem is that the subsequence $(n_k)_{k\in\mathbb{N}}$ might depend on the point $x \in X$. Thus, if the metric space X has more than countably many points, we cannot just conclude by using a diagonal argument. This is where the assumption of equicontinuity of the family $\mathcal{F}$, together with the separability of X, comes to the rescue. Given a dense set $(x_i)_{i\in\mathbb{N}}$ of X, by a diagonal argument we can extract a subsequence $(n_k)_{k\in\mathbb{N}}$ such that

$$\lim_{k\to\infty} f_{n_k}(x_i) = f(x_i)$$

for all $i \in \mathbb{N}$, where $f(x_i) \in \mathbb{R}$. The function f is defined only on the set $\{x_i\}_{i\in\mathbb{N}}$. It is possible to see that it is uniformly continuous. Thus, we can uniquely extend it to a continuous function defined on the entire space X. By using the compactness of the space X, such convergence will turn out to be uniform. Let us now use this idea to get the rigorous proof.

Since the metric space X is compact, it is separable (see Lemma 2.34. Let $(x_i)_{i\in\mathbb{N}}$ be a dense set in X. Since $\mathcal{F}$ is equibounded, there exists $D < \infty$ such that

$$|f_n(x_i)| \leq D, \tag{4.3}$$

for all $i \in \mathbb{N}$ and $n \in \mathbb{N}$.

Step 1. We will construct a subsequence of indexes $(n_k)_{k\in\mathbb{N}}$ such that

$$\exists \lim_{k\to\infty} f_{n_k}(x_i)$$

for all $i \in \mathbb{N}$ by using a diagonal argument as follows. By (4.3), we get that

$$\{f_n(x_1)\}_{n\in\mathbb{N}} \subset [-D, D].$$

Thus, from the Bolzano Weierstraß Theorem (see Theorem 2.23) applied to the set $[-D, D] \subset \mathbb{R}$, we can find a subsequence of indexes $(n_k^1)_{k\in\mathbb{N}}$, and a point $y_1 \in [-D, D]$, such that

$$\lim_{k\to\infty} f_{n_k^1}(x_1) = y_1. \tag{4.4}$$

Now, consider the sequence $\{f_{n_k^1}(x_2)\}_{k\in\mathbb{N}}$. As before, from $(n_k^1)_{k\in\mathbb{N}}$ it is possible to extract a further subsequence $(n_k^2)_{k\in\mathbb{N}}$ such that

$$\lim_{k\to\infty} f_{n_k^2}(x_2) = y_2, \tag{4.5}$$

for some $y_2 \in [-D, D]$. Note that, since $(n_k^2)_{k\in\mathbb{N}}$ is a subsequence of $(n_k^1)_{k\in\mathbb{N}}$, we get that

$$\lim_{k\to\infty} f_{n_k^2}(x_1) = y_1.$$

Namely, extracting further subsequence does not invalidate the convergence of the previous steps. We now repeat the same argument for $i \in \{3, 4, \dots\}$. Since we do this *countably* many times, by using a diagonal argument, we obtain a subsequence of indexes $(n_k)_{k\in\mathbb{N}}$, and points $\{y_i\}_{i\in\mathbb{N}} \subset [-D, D]$ such that

$$\lim_{k\to\infty} f_{n_k}(x_i) = y_i, \tag{4.6}$$

for all $i \in \mathbb{N}$.

Step 2. We now define the limiting function $f : X \to \mathbb{R}$ as follows. In step 1 we constructed the values of f on the dense set $(x_i)_{i\in\mathbb{N}}$. We then define the function $f : X \to \mathbb{R}$ as follows: for $x \in X$, we set

$$f(x) := \lim_{j\to\infty} y_{i_j}, \tag{4.7}$$

where $(x_{i_j})_{j\in\mathbb{N}}$ is a sequence of points in the dense set $\{x_i\}_{i\in\mathbb{N}}$ converging to x, and the y_{i_j}'s are defined in (**??**). We will prove the followings:

(1) f is well defined;
(2) $f(x_i) = y_i$ for all $i \in \mathbb{N}$.

Warning: it would be tempting to use the identity

$$f(x) = \lim_{j\to\infty} f_{n_{k_j}}(z_j),$$

for any sequence $(z_j)_{j\in\mathbb{N}}$ converging to x. This is true, but it will follow from the uniform convergence of $(f_{n_k})_{k\in\mathbb{N}}$ to f and from Proposition 3.35. Therefore, in the following, we cannot use it (or we would have to prove it first).

Let's start with (1). To prove that f is well defined, we need to show that the limit exists, and that it does not depend on the sequence converging to x.

Fix $x \in X$. Let $(x_{i_j})_{j\in\mathbb{N}}$ be such that $x_{i_j} \to x$. Then, since by (4.3), the sequence $(f(x_{i_j}))_{j\in\mathbb{N}}$ is bounded, we can apply the Bolzano-Weirstraß Theorem (see 2.23) and get that there exists a subsequence $(x_{i_{j_r}})_{r\in\mathbb{N}}$ such that $(y_{i_{j_r}})_{r\in\mathbb{N}}$ converges to some point.

We now prove that, if $(x_{i_p})_{p\in\mathbb{N}}$ and $(x_{i_q})_{q\in\mathbb{N}}$ are two sequences converging to x, and such that

$$y_{i_p} \to y^1, \qquad y_{i_q} \to y^2, \tag{4.8}$$

then, $y^1 = y^2$. Fix $\varepsilon > 0$. We will show that $|y^1 - y^2| < \varepsilon$. By using the equicontinuity of the family $\mathcal{F}$, there exists $\delta > 0$ such that

$$|f_n(s) - f_n(t)| < \varepsilon, \tag{4.9}$$

for all $s, t \in X$ with $\mathrm{d}(s, t) < \delta$, and all $n \in \mathbb{N}$. Since the sequences $(x_{i_p})_{p\in\mathbb{N}}$ and $(x_{i_q})_{q\in\mathbb{N}}$ converge to x, there exist $\bar{p}, \bar{q} \in \mathbb{N}$ such that

$$\mathrm{d}(x_{i_p}, x) < \frac{\delta}{2}, \qquad \mathrm{d}(x_{i_q}, x) < \frac{\delta}{2},$$

for all $p \geq \bar{p}$, and all $q \geq \bar{q}$. In particular, $\mathrm{d}(x_{n_p}, x_{n_q}) < \delta$, for all $p \geq \bar{p}$, and all $q \geq \bar{q}$, and thus, from (4.9), we get

$$|f_{n_k}(x_{i_p}) - f_{n_k}(x_{i_q})| < \varepsilon, \tag{4.10}$$

for all $p \geq \bar{p}$, all $q \geq \bar{q}$, and all $k \in \mathbb{N}$. Thus, using the definition of y_{i_p} and y_{i_q}, from (4.10) we get

$$|y_{i_p} - y_{i_q}| = \lim_{k\to\infty} |f_{n_k}(x_{i_p}) - f_{n_k}(x_{i_q})| < \varepsilon, \tag{4.11}$$

for all $p \geq \bar{p}$, all $q \geq \bar{q}$. Therefore, (4.8) and (4.11) yield

$$|y^1 - y^2| = \lim_{p,q\to\infty} |y_{i_p} - y_{i_q}| < \varepsilon.$$

Since $\varepsilon > 0$ is arbitrary, we conclude that $y^1 = y^2$.

Finally, combining the two claims above, and using the Urysohn property (see Proposition 2.6) we get that f is well defined. Indeed, we proved that there exists a value $y \in \mathbb{R}$ with the following property: from any sequence $(x_{i_j})_{j\in\mathbb{N}}$ converging to x, we can extract a subsequence $(x_{i_{j_p}})_{p\in\mathbb{N}}$ such that $y_{i_{j_p}} \to y$ as $p \to \infty$.

Let us now prove (2). Fix x_i. Since we just proved that the definition of $f(x)$ does not depend on the sequence converging to x, we can just take the constant $x_{i_k} := x_i$ for each $k \in \mathbb{N}$. This gives that $f(x_i) = y_i$.

Step 3. We now prove that f is continuous. Thanks to Theorem 3.7, we can equivalently show that f is sequentially continuous. Let $x \in X$, and let $(z_j)_{j\in\mathbb{N}} \subset X$ be a sequence converging to x. We want to prove that

$$\lim_{j\to\infty} f(z_j) = f(x). \tag{4.12}$$

By using the definition of f, for each $j \in \mathbb{N}$ it is possible to find a point x_{i_j} belonging to the dense set $\{x_i\}_{i\in\mathbb{N}}$ such that

$$\mathrm{d}(x_{i_j}, z_j) \le \frac{1}{j} \tag{4.13}$$

and

$$\left|y_{i_j} - f(z_j)\right| \le \frac{1}{j}. \tag{4.14}$$

Then, by using the fact that z_j converges to x, together with (4.13), we get that $x_{i_j} \to x$ as $j \to \infty$. Thus, by the definition of $f(x)$, we get that

$$\lim_{j\to\infty} y_{i_j} = f(x). \tag{4.15}$$

Therefore, from (4.14) we get

$$|f(z_j) - f(x)| \le \left|f(z_j) - y_{i_j}\right| + \left|y_{i_j} - f(x)\right| \le \frac{1}{j} + \left|y_{i_j} - f(x)\right|.$$

From (4.15), we get that the right-hand side converges to zero as $j \to \infty$. This yields (4.12).

Step 4. We now prove that $(f_{n_k})_{k\in\mathbb{N}}$ converges to f uniformly. Assume not. Then, there would exist $\varepsilon > 0$ and points $(z_k)_{k\in\mathbb{N}}$ such that

$$|f_{n_k}(z_k) - f(z_k)| > \varepsilon, \tag{4.16}$$

for all $k \in \mathbb{N}$. Since X is compact, and hence, by Theorem 2.35, sequentially compact, there exists a subsequence $(z_{k_j})_{j\in\mathbb{N}}$ and a point $x \in X$ such that z_{k_j} converges to x.

By using the equicontinuity of the family $\mathcal{F}$, it is possible to find $\delta > 0$ such that

$$|f_n(s) - f_n(t)| < \frac{\varepsilon}{3}, \tag{4.17}$$

for all $s, t \in X$ with $\mathrm{d}(s,t) < \delta$, and all $n \in \mathbb{N}$. For each $j \in \mathbb{N}$, by using the density of $\{x_i\}_{i\in\mathbb{N}}$, we can find $x_{k_j} \in \{x_i\}_{i\in\mathbb{N}}$ such that

$$\mathrm{d}(x_{k_j}, z_{k_j}) < \delta. \tag{4.18}$$

Thus, (4.18) together with (4.17) yields that

$$|f_{n_{k_j}}(x_{k_j}) - f_{n_{k_j}}(z_{k_j})| < \frac{\varepsilon}{3}, \tag{4.19}$$

for all $j \in \mathbb{N}$.

Now, by using the continuity of f at x, up to reducing $\delta > 0$, we can also assume that

$$|f(s) - f(x)| < \varepsilon, \tag{4.20}$$

for all $s \in X$ with $\mathrm{d}(s,x) < \delta$. Since $(z_{k_j})_{j\in\mathbb{N}}$ converges to x, we can find $j_0 \in \mathbb{N}$ such that

$$\mathrm{d}(z_{k_j}, x) < \delta, \tag{4.21}$$

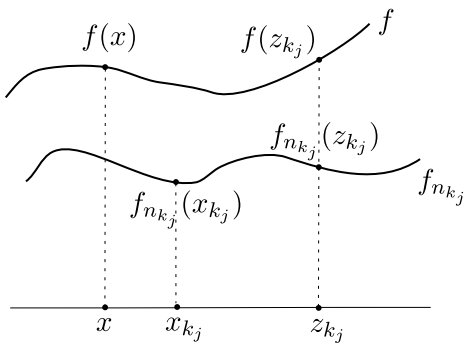

FIGURE 9. The idea behind the strategy of Step 4 of the proof of Theorem 4.12 to show the uniform convergence.

for all $j \geq j_0$. Thus, from (4.21), and (4.20), we get that

$$|f(x) - f(z_{k_j})| < \frac{\varepsilon}{3}, \tag{4.22}$$

for all $j \geq j_0$.

Finally, since $z_{k_j} \to x$, by using the definition of $f(x)$, we get that

$$\lim_{j\to\infty} f_{n_{k_j}}(x_{n_{k_j}}) = f(x).$$

Therefore, up to increasing j_0, we can assume that

$$|f(x) - f_{n_{k_j}}(x_{k_j})| < \frac{\varepsilon}{3}, \tag{4.23}$$

for all $j \geq j_0$.

We are now ready to conclude. Using (4.19), (4.22), and (4.23), we get (see Figure 9)

$$|f(z_{k_j}) - f_{n_{k_j}}(z_{k_j})| \leq |f(z_{k_j}) - f(x)| + |f(x) - f_{n_{k_j}}(x_{k_j})| + |f_{n_{k_j}}(x_{k_j}) - f_{n_{k_j}}(z_{k_j})| < \varepsilon,$$

for all $j \geq j_0$. This is in contradiction with (4.16). Thus, the proof of the theorem is concluded. □

Remark 4.13. The Ascoli-Arzelà Theorem can be generalized to the case where the target is a metric space.

All of the assumptions of the Ascoli-Arzelà Theorem are sharp. Namely, if they are not in force, the result does not hold. Try to find a counterexample in the case all assumptions are in force except one of the followings:

(i) The equiboundeness of the family $\mathcal{F}$;
(ii) The equicontinuity of the family $\mathcal{F}$;
(iii) The compactness of the space X.

Remark 4.14. The Ascoli-Arzelà Theorem gives, in particular, sufficient conditions for a sequence of functions to converge uniformly, up to a subsequence. In practice, the difficult part is to check the equicontinuity of the family. Something that comes at handy in this regard is a particular subspace of $C^0(X, Y)$, known as the space of *Lipschitz functions*. This is defined as the family of functions $f : X \to Y$ such that

$$[f]_{\mathrm{Lip}} := \sup_{x\neq y} \frac{\mathrm{d}_2(f(x), f(y))}{\mathrm{d}_1(x, y)} < \infty.$$

In particular, if a family has a uniform bound on the Lipschitz norm

$$\|f\|_{\mathrm{Lip}} := \|f\|_{C^0} + [f]_{\mathrm{Lip}},$$

then, the family satisfies the sufficient conditions of the Ascoli-Arzelà Theorem. More properties of the space of Lipschitz functions will be investigated in the exercises.

As anticipated above, as an application of the Ascoli-Arzelà Theorem, we will sketch the proof of Peano's Theorem (1890) on the existence of a solution to the initial value problem

$$\begin{cases} y'(t) = f(t, y(t)), \\ y(x_0) = y_0, \end{cases} \tag{4.24}$$

The full proof will be given as an exercise (and you will see it again in the course *Ordinary Differential Equations*). The importance of such a result lies in the weak assumption on the term f: indeed, in the proof by Picard and Lindelöf, the function f is required to be Lipschitz in the second variable. This additional assumption gives existence and *uniqueness* of the solution. On the other hand, existence of a solution follows from continuity alone of f. In such a case, though, uniqueness might fail dramatically. The strategy of the proof is based on an approximation scheme known as *forward Euler* that is common in Numeric (you will see it in the course *Numerical Methods* for ODEs, and in *Numerical Methods for PDEs* for PDEs).

Theorem 4.15 (Peano's Theorem)**.** *Consider the initial value problem* (4.24)*. Assume that there exists $L > 0$ such that f is continuous in $(x_0 - L, x_0 + L)$. Then, there exists $\varepsilon \in (0, L)$ and a C^1 function $y : (x_0 - \varepsilon, x_0 + \varepsilon) \to \mathbb{R}$ solving* (4.24)*.*

(Sketch of the proof). The idea is to construct a sequence $(u_n)_{n\in\mathbb{N}}$ of *approximating* solutions, show that they satisfy the assumptions of the Ascoli-Arzelà Theorem, and prove that any limit solves the initial value problem (4.24). The strategy is based on the discretization of the time derivative. Namely, for $n \in \mathbb{N} \setminus \{0\}$, we consider $(t_i)_{i\in\mathbb{Z}}$ defined as

$$t_i := x_0 + \frac{i}{n}.$$

Let $i_0 \in \mathbb{N}$ be such that $t_{i_0} \in (x_0 - L, x_0 + L)$, while $t_{i_0+1} \notin (x_0 - L, x_0 + L)$. Let $u_0^n := y_0$, and, for $|i| < i_0$, define recursively

$$u_{i+1}^n := u_i^n + \frac{1}{n} f(t_i, u_i^n),$$

and the function $u_n : (x_0 - L/2, x_0 + L/2) \to \mathbb{R}$ as the linear interpolation of the u_i^n's. Then, the sequence $(u_n)_{n\in\mathbb{N}}$ turns out to be equi-Lipschitz. It is possible to show that any limit is a solution to (4.24). □

Remark 4.16. There is another strategy that requires the validity of the Picard and Lindelöf Theorem, and it is based on the approximation of the term f by polynomial functions. Such approximation is possible thanks to Theorem 4.17.

4.3. **Separability: the Theorems of Weierstraß.** Separability of a space is an extremely important property. Think about the real numbers $\mathbb{R}$. It is separable, and a countable dense set is given by the rationals $\mathbb{Q}$. These are objects that are simpler than a general real number, and that can also be implemented on a computer. In the same spirit, having a good countable dense set on a general space allows to access, with some (hopefully controlled) error, all of the objects of that space.

In particular, the 'analogous' of rational numbers for real valued continuous functions on an interval is the family of polynomials with rational coefficients.

We will first start with presenting the proof for continuous functions $f : [0, 1] \to \mathbb{R}$. In 1885 Weierstraß, at the tender age of 70, proved that algebraic polynomials are dense in $C([0, 1])$, and that trigonometric polynomials are dense in the subclass of functions in $C([0, 1])$ with periodic boundary conditions. With these results, he started the line of research called *approximation theory*.

Theorem 4.17 (Weierstraß Theorem)**.** *Let $f : [0, 1] \to \mathbb{R}$ be a continuous function, and $\varepsilon > 0$. Then, there exists a polynomial $P : [0, 1] \to \mathbb{R}$ such that $\|f - P\|_{C^0} < \varepsilon$.*

Many proofs and generalization of such results are now available. The strategy that we will follow is that of Lebesgue (in his first published paper in 1898, when he was still a 23 year old

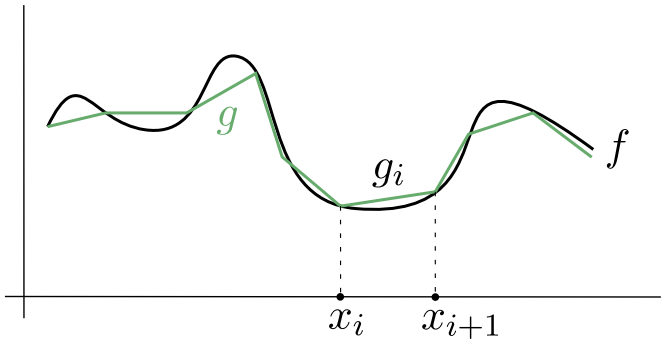

FIGURE 10. The approximation of a continuous function by a spline.

student), with a bit of Bourbaki[4] (in 1949) for the approximation of the absolute value (see Lemma 4.18).

The idea is the following:

- *Step 1:* Any continuous function can be approximated uniformly by a *spline*, namely a piecewise affine function (see Figure 10);
- *Step 2:* Any spline can be written as a linear combination of a linear function (polynomial of degree one), and absolute values;
- *Step 3:* The absolute value can be approximated uniformly by a sequence of polynomials.

Note how ingenious such a proof is: it reduces the problem of approximating *any* continuous function, to the problem of approximating a *specific* function, namely the absolute value.

Let us carry on the above strategy. Let $f : [0,1] \to \mathbb{R}$, and $\varepsilon > 0$. Step 1 requires us to find a piecewise affine function $g : [0,1] \to \mathbb{R}$ such that

$$\|f - g\|_{C^0} < \varepsilon. \tag{4.25}$$

This is left as an exercise for the reader.

We now write a spline g in a suitable form. The first way that comes to your mind to write a piecewise affine function is the following. Let $0 = x_0 < x_1 < \cdots < x_k = 1$ be the nodes of the spline. Then,

$$g(x) = g_1(x) + \sum_{i=1}^{k-1} [g_{i+1}(x) - g_i(x)]h(x - x_i), \tag{4.26}$$

where

$$g_i(x) = g(x_{i-1}) + \frac{x - x_{i-1}}{x_i - x_{i+1}}(g(x_i) - g(x_{i-1})),$$

and

$$h(x) := \begin{cases} 1 & \text{if } x \geq 0, \\ 0 & \text{if } x < 0. \end{cases}$$

It is possible to rewrite (4.26) as

$$g(x) = ax + b + \sum_{i=1}^{k-1} c_i (x - x_i)_+, \tag{4.27}$$

for some $a, b, c_i \in \mathbb{R}$, where

$$t_+ := \begin{cases} t & \text{if } t \geq 0, \\ 0 & \text{if } t < 0. \end{cases}$$

[4]Bourbaki is one example where asking *Who is Bourbaki?* is not correct, while asking *What is Bourbaki?* is. Look it up!

Now, since $2t_+ = |t| + t$, it is possible to write (4.27) as

$$g(x) = Ax + B + \sum_{i=1}^{k-1} C_i |x - x_i|, \tag{4.28}$$

for some $A, B, C_i \in \mathbb{R}$.

We are now left with the problem of approximating the absolute value with a polynomial. Here we do not follow what Lebesgue did, but the easier solution found by Bourbaki.

Lemma 4.18. *There exists a sequence $(P_n)_{n\in\mathbb{N}}$, where each $P_n : [-1,1] \to \mathbb{R}$ is a polynomial, such that P_n converges uniformly on $[-1,1]$ to g, where $g(t) \coloneqq |t|$.*

Proof. The idea of the proof is the following: since $|t| = \sqrt{t^2}$, it suffices to approximate $f(t) \coloneqq \sqrt{t}$. To do that, we construct the polynomial iteratively, starting with $P_0 \equiv 0$, and adding half of the distance that separates it from f. In such a way, we get an increasing sequence of functions that converge pointwise, and hence uniformly by Dini's Theorem, to the desired function. The problem with this, is that the approximating functions are not polynomials. To fix that, we consider the distance from P_n^2 to f^2.

Step 1. Set $P_0 \equiv 0$. Define, for $n \in \mathbb{N} \setminus \{0\}$, P_n inductively as follows:

$$P_{n+1}(t) \coloneqq P_n(t) + \frac{1}{2}\left[\, t - P_n^2(t) \,\right].$$

Then, it is easy to see that each P_n is a polynomial.

Step 2. We claim that $P_n(t) \leq \sqrt{t}$ for all $t \in [0,1]$, and all $n \in \mathbb{N}$. We will prove the claim by induction on n. For $n = 0$ it is trivial. Then, assume that the claim is true for n, and write

$$\begin{aligned} \sqrt{t} - P_{n+1}(t) &= \sqrt{t} - P_n(t) - \frac{1}{2}\left[\, t - P_n^2(t) \,\right] \\ &= \left[\sqrt{t} - P_n(t)\right]\left[1 - \frac{1}{2}\left[\sqrt{t} + P_n(t)\right]\right]. \end{aligned} \tag{4.29}$$

Since we are assuming $P_n(t) \leq \sqrt{t}$, and in particular that $P_n(t) \leq 1$, we get that the right-hand side is non-negative.

Step 3. We claim that $P_n \leq P_{n+1}$ for all $n \in \mathbb{N}$. This follows directly from (4.29).

Step 4. We claim that P_n converges to f pointwise. Indeed, from step 3, we get that

$$\sqrt{t} - P_{n+1}(t) = \theta_n(t)\left[\sqrt{t} - P_n(t)\right],$$

for some $\theta_n(t) \in [c, 1)$, for some $c > 0$.

Step 5. Finally, we conclude the proof of the lemma, since, by Dini's Theorem (see Theorem 3.42), the pointwise convergence of the monotone sequence $(P_n)_{n\in\mathbb{N}}$ is actually uniform. □

Remark 4.19. Why, in the above proof, we approximated $\sqrt{t}$ instead of $|t|$ directly?

By using (4.25), (4.28), and Lemma 4.18, we conclude the proof of Theorem 4.17.

We now consider the problem of approximating periodic functions. The proof we present is due to de la Vallèe Poussin (1919), and it is based on the existence of algebraic polynomials approximating a continuous function.

Theorem 4.20 (Weierstraß Theorem)**.** *Let $f : [0, 2\pi] \to \mathbb{R}$ be a continuous function with $f(0) = f(2\pi)$, and $\varepsilon > 0$. Then, there exists a trigonometric polynomial $T : [0, 2\pi] \to \mathbb{R}$ such that $\|f - T\|_{C^0} < \varepsilon$.*

Proof. Extend f as a function to the entire $\mathbb{R}$ in a 2π periodic way, Namely, with an abuse of notation, set $f(x+2k\pi) \coloneqq f(x)$, for all $x \in [0,2\pi]$, and all $k \in \mathbb{Z}$. Fix $\varepsilon > 0$. Consider the 2π-periodic continuous functions

$$g(x) \coloneqq \frac{f(x)+f(-x)}{2}, \qquad h(x) \coloneqq \frac{f(x)-f(-x)}{2}\sin x.$$

Set, for $x \in [-1,1]$,

$$\varphi(x) \coloneqq g(\arccos x), \qquad \psi(x) \coloneqq h(\arccos x).$$

By Weierstraß Theorem 4.17 there exist algebraic polynomial $P, Q : [-1,1] \to \mathbb{R}$ such that

$$\|\varphi - P\|_{C^0} < \frac{\varepsilon}{4}, \qquad \|\psi - Q\|_{C^0} < \frac{\varepsilon}{4}. \tag{4.30}$$

Since g and h are even, from (4.30) we get

$$\|g - P\circ\cos\|_{C^0} < \frac{\varepsilon}{4}, \qquad \|h - Q\circ\cos\|_{C^0} < \frac{\varepsilon}{4}. \tag{4.31}$$

Using the definitions of g and h, together with (4.31), we get

$$\begin{aligned}
&\sup_{x\in[0,2\pi]} \|f(x)\sin^2 x - \left[P(\cos x)\sin^2 x + Q(\cos x)\sin x\right]\| \\
&= \sup_{x\in[0,2\pi]} \left\| f(x)\sin^2 x + \frac{f(-x)\sin^2 x}{2} - \frac{f(-x)\sin^2 x}{2} - \left[P(\cos x)\sin^2 x + Q(\cos x)\sin x\right]\right\| \\
&\leq \sup_{x\in[0,2\pi]} \|g(x)\sin^2 x - P(\cos x)\sin^2 x\| + \sup_{x\in[0,2\pi]} \|h(x)\sin x - Q(\cos x)\sin x\| \\
&\leq \sup_{x\in[0,2\pi]} \|g(x) - P(\cos x)\| + \sup_{x\in[0,2\pi]} \|h(x) - Q(\cos x)\| \\
&< \frac{\varepsilon}{2}.
\end{aligned} \tag{4.32}$$

By using the same argument for the function $x \mapsto f(x+\pi/2)$, we find algebraic polynomials $R, S : [0,2\pi] \to \mathbb{R}$ such that

$$\sup_{x\in[0,2\pi]} \|f(x+\pi/2)\sin^2 x - \left[R(\cos x)\sin^2 x + S(\cos x)\sin x\right]\| < \frac{\varepsilon}{2}. \tag{4.33}$$

Thus, from (4.32) and (4.33) we get that

$$\|f - T\|_{C^0} < \varepsilon,$$

where $T : [0,2\pi] \to \mathbb{R}$ is the trigonometric polynomial defined as

$$T(x) \coloneqq P(\cos x)\sin^2 x + Q(\cos x)\sin x + R(\sin x)\cos^2 x - S(\sin x)\cos x,$$

for $x \in [0,2\pi]$. □

Remark 4.21. It is possible to prove that the validity of Theorem 4.20 implies the validity of Theorem 4.17. Namely, the density of algebraic polynomials in the space of continuous functions is *equivalent* to the density of trigonometric polynomials in the space of periodic continuous functions.

Several other proofs were given over the years of Weierstraß Approximation Theorem. In particular, we point out two strategies of proofs. The first one is based on convolution: namely approximation of the identity by polynomials by using singular integrals. The result presented below, namely the choice of the particular singular kernel, is due to Landau (1908).

Theorem 4.22. *Let $f \in C^0([0,1])$. Define $\widetilde{f} : \mathbb{R} \to \mathbb{R}$ as*

$$\widetilde{f}(x) \coloneqq \begin{cases} f(x) \coloneqq f(x) - f(0) - (f(1)-f(0))\,x & \text{if } x \in [0,1], \\ 0 & \text{else.} \end{cases}$$

Namely, $\widetilde{f}$ vanishes at $x = 0$ and at $x = 1$. For each $n \in \mathbb{N}$, define

$$P_n(t) := \int_{-1}^{1} \widetilde{f}(x+t) c_n (1-x^2)^n \, \mathrm{d}x,$$

where

$$c_n^{-1} := \int_{-1}^{1} (1-x^2)^n \, \mathrm{d}x.$$

Then, the sequence of polynomials $(P_n)_{n\in\mathbb{N}}$ converges to $\widetilde{f}$ uniformly on $[0,1]$.

The other strategy is based on the use of the values of f on rational points: it gives an almost explicit way to construct the approximating polynomials, as well as error estimated. The result presented below is due to Bernstein (1913), who used probabilistic arguments to prove it.

Theorem 4.23. *Let $f \in C^0([0,1])$. For $n \in \mathbb{N}$, define the Bernstein polynomial $B_n : [0,1] \to \mathbb{R}$ as*

$$B_n(x) := \sum_{j=0}^{n} f\left(\frac{j}{n}\right) C_j^n x^j (1-x)^{n-j},$$

where

$$C_j^n := \frac{n!}{j!(n-j)!}.$$

Then, the sequence of polynomials $(B_n)_{n\in\mathbb{N}}$ converges to f uniformly on $[0,1]$. In particular,

$$\|f - B_n\|_{C^0} \leq \frac{5}{4} \omega_f \left(\frac{1}{\sqrt{n}}\right),$$

for all $n \in \mathbb{N}$, where ω_f is the modulus of continuity of f, defined as follows: for $t > 0$, $\omega_f(t)$ is the greatest $\delta > 0$ such that

$$|f(x) - f(y)| < t,$$

for all $x, y \in [0,1]$ with $|x - y| < \delta$.

Finally, approximation results by algebraic polynomials hold also in higher dimension, but the strategy of the proof is more involved.

4.4. **Separability: the Theorem of Stone.** We now wonder if the approximation result of Theorem 4.17 uses in essential way the structure of $\mathbb{R}$ or if it can be generalized to metric spaces. Unfortunately, in a general metric space, we do not have polynomials at our disposal. Nevertheless, it is possible to obtain an approximation result by using a type of family of functions that resembles the properties of the family of algebraic polynomials.

Definition 4.24. A subset $\mathcal{A} \subset C^0(X)$ is called a *subalgebra* if:

(i) It is a linear space: namely, if $f, g \in \mathcal{A}$, and $\lambda, \mu \in \mathbb{R}$ then $\lambda f + \mu g \in \mathcal{A}$;
(ii) It is closed under multiplication: namely, if $f, g \in \mathcal{A}$, then $fg \in \mathcal{A}$.

We now look for necessary conditions for an algebra to be dense in $C^0(X)$. First of all, we note that, if we take two points $x \neq y \in X$, we need to have an element $f \in \mathcal{A}$ such that

$$f(x) \neq f(y).$$

Moreover, for each $x \in X$ there must be at least one element $f \in \mathcal{A}$ with $f(x) \neq 0$. We give names to these two properties.

Definition 4.25. Let $\mathcal{A}$ be a subalgebra of $C^0(X)$. We say that $\mathcal{A}$ *separates points* if, for all $x \neq y \in X$, there exists $f \in \mathcal{A}$ such that

$$f(x) \neq f(y).$$

Definition 4.26. Let $\mathcal{A}$ be a subalgebra of $C^0(X)$. We say that $\mathcal{A}$ satisfies the *non-vanishing property* if, for each $x \in X$, there exists $f \in \mathcal{A}$ with $f(x) \neq 0$.

It turns out that these two properties ensure density of the subalgebra.

Theorem 4.27 (Stone Theorem)**.** *Let $\mathcal{A}$ be a subalgebra of $C^0(X)$. Then, $\mathcal{A}$ is dense in $C^0(X)$ with respect to the uniform norm if and only if it separates points and satisfies the non-vanishing property.*

In order to focus on the main ideas of the proof of Stone Theorem, we first isolate two technical results.

Lemma 4.28. *Let $\mathcal{A}$ be a subalgebra of $C^0(X)$ that separates points. Then, for every $x \neq y \in X$, and every $a, b \in \mathbb{R}$, there exists $f \in \mathcal{A}$ such that $f(x) = a$, and $f(y) = b$.*

Proof. Since $\mathcal{A}$ separates points, there exist $g \in \mathcal{A}$ with $g(x) \neq g(y)$. Then, define $f : X \to \mathbb{R}$ by

$$f(z) \coloneqq a + (b-a)\frac{g(z) - g(x)}{g(y) - g(x)},$$

for all $z \in X$. Note that, since $g(y) \neq g(x)$, we have that the denominator is different from zero. Then, $f \in C^0(X)$ satisfies the desired property. □

Lemma 4.29. *Let $\mathcal{A}$ be a subalgebra of $C^0(X)$. Then, for each $f, g \in \mathcal{A}$, we have that the functions*

$$\max\{f, g\}, \qquad \min\{f, g\}.$$

are uniform limits of sequences in $\mathcal{A}$.

Proof. Write

$$\max\{f, g\} = \frac{1}{2}(f + g + |f - g|), \qquad \min\{f, g\} = \frac{1}{2}(f + g - |f - g|).$$

Therefore, if we prove that, for every $h \in \mathcal{A}$, the function $|h|$ is a limit of elements in $\mathcal{A}$, we are done. For, consider the sequence

$$u_n \coloneqq P_n\left(\frac{h^2}{\|h\|_\infty^2}\right),$$

where P_n is the polynomial defined in Lemma 4.18. Then, by using Lemma 4.18, we get that the sequence $(u_n)_{n\in\mathbb{N}}$ converges uniformly to $|h|$. This concludes the proof □

We can now present the proof of Theorem 4.27. Stone first proved it in 1937, and it was one of the first examples of the use of algebraic ideas in analysis. The proof we present here is that that Stone wrote in 1948 and 1962 and benefits from suggestions from Kakutani and Chevalley. It is based on a double compactness argument, that allows to construct a function that satisfies the upper and the lower bound, respectively.

Proof of Theorem 4.27. Step 1. Assume that the subalgebra $\mathcal{A}$ is dense in C^0. Then, it is easy to see that it separates points and satisfies the non-vanishing property.

Step 2. Let $f \in C^0(X)$, and $\varepsilon > 0$. We claim that there exists a function $g \in \overline{\mathcal{A}}$ such that

$$\|f - g\|_{C^0} < \varepsilon.$$

The idea is to obtain the lower and the upper bound separately, by using for each a compactness argument.

Step 1. Let $x \in X$. We claim that there exists $g_x \in \overline{\mathcal{A}}$ such that

$$g_x(x) = f(x), \qquad g_x(y) \leq f(y) + \varepsilon,$$

for all $y \in X$. Indeed, since $\mathcal{A}$ separates points, for each $z \in X$ there exists $h_z \in \mathcal{A}$ such that

$$h_z(x) = f(x), \qquad h_z(z) = f(z) + \frac{\varepsilon}{2}.$$

Since h_z and f are continuous, there exists $r(z) > 0$ such that

$$h_z(y) \leq f(y) + \varepsilon,$$

for all $y \in B(z, r(z))$. Then, consider the open cover of X given by $(B(z, r(z)))_{z \in X}$. Since X is compact, there exists a finite subfamily $B(z_1, r(z_1)), \dots, B(z_k, r(z_k))$ that still covers X. Define

$$g_x \coloneqq \min\{g_{z_i} : i = 1, \dots, k\}.$$

By using Lemma 4.29 we get that $g_x \in \overline{\mathcal{A}}$. Moreover, by construction, it satisfies the desired property.

Step 2. For each $x \in X$, let g_x be the function given by the previous step. By continuity of g_x and f, there exists $s(x) > 0$ such that

$$g_x(y) \geq f(y) - \varepsilon,$$

for all $y \in B(x, s(x))$. As before, consider the open cover of X given by $(B(x, s(x)))_{x \in X}$. Since X is compact, there exists a finite subfamily $B(x_1, s(x_1)), \dots, B(x_j, s(x_j))$ that still covers X. Define

$$g \coloneqq \max\{g_{x_i} : i = 1, \dots, j\}.$$

By using Lemma 4.29 we get that $g \in \overline{\mathcal{A}}$. Moreover, by construction, it satisfies the desired property. □

Corollary 4.30. *Let* (X, d) *be a compact metric space. Then,* $C^0(X)$ *is separable.*

Proof. We want to construct a subalgebra $\mathcal{A}$ of $C^0(X)$ that separates points, and satisfies the non-vanishing property.

Since X is compact, by Lemma 2.34 we know that there exists a dense set $(x_n)_{n \in \mathbb{N}}$. For each $n \in \mathbb{N}$, consider the countable family of balls $(B(x_n, 1/k))_{k \in \mathbb{N} \setminus \{0\}}$. Then, the set

$$\big((B(x_n, 1/k))_{k \in \mathbb{N} \setminus \{0\}}\big)_{n \in \mathbb{N}}$$

is countable. We denote if by $(A_i)_{i \in \mathbb{N}}$. For each $i \in \mathbb{N}$, define $g_i : X \to \mathbb{R}$ as

$$g_i(x) \coloneqq \operatorname{dist}(x, X \setminus A_i).$$

Then, each g_i is continuous. Consider the countable set

$$\mathcal{A} \coloneqq \{g_1^{d_1} \cdots g_k^{d_k} \ : \ k \in \mathbb{N},\, d_i \in \mathbb{N} \text{ for all } i = 1, \dots, k\}.$$

Then, it is easy to see that $\mathcal{A}$ is a subalgebra of $C^0(X)$.

We claim that $\mathcal{A}$ separates points. Let $x \neq y \in X$. By definition of $(A_i)_{i \in \mathbb{N}}$, there exists A_j such that $x \in A_j$, $y \notin A_j$. Therefore, $g_j(x) \neq 0$, and $g(y) = 0$. This also proves that $\mathcal{A}$ satisfies the non-vanishing property. Thus, Theorem 4.27 ensures that the countable set $\mathcal{A}$ is dense in $C^0(X)$. □

Remark 4.31. Note that $B(X)$, the space of bounded functions, is not separable. Prove it!

Remark 4.32. For continuous functions taking values in $\mathbb{C}$, more assumptions on the algebra are needed in order for the analogue of the Stone-Weierstraß Theorem to hold.

4.5. **Nowhere differentiable functions.** It came as a huge shock for the mathematical community, when in in the nineteenth century the first examples of functions that are nowhere differentiable were constructed. If you think about it, it is easy to accept that there are functions that are not differentiable at finitely many points, like piecewise affine functions. But a function for which the limit of the different quotient did not exist was something that, today as at that time, has a bit of surprise in it. In particular, it calls for deep rigorous mathematics to understand the behavior of such functions, that are nowadays commonly used in Brownian motion, chaos theory, fractals, among others areas. It seems that Bolzano was the first one to gave an example of a function nowhere differentiable in the 1830's.

Two are the examples that are usually presented today: one by Weierstraß (1861), and one by Takagi (1903). The former is given by

$$f(x) \coloneqq \sum_{n=0}^{\infty} b^n \cos(a^n \pi x),$$

for $b \in (0,1)$, a odd, and $ab > 1 + 3\pi/2$. The latter by

$$f(x) \coloneqq \sum_{i=0}^{\infty} \frac{g(2^i x)}{2^i},$$

where $g(x) \coloneqq \operatorname{dist}(x, \mathbb{Z})$.

Once the existence of a nowhere differentiable function is settled, next question is to investigate how many of such functions there are. It turns out that there are a lot of them (in a topologically precise sense). To be precise, we say that a set is of *first category*, if it can be written as a countable union of closed sets each of which does not contain any ball in the uniform metric. It is a result in topology (called Baire's Category Theorem) that first category sets have empty interior. In particular, they cannot be dense. What can be proved is the following.

Theorem 4.33. *The complement in the space of continuous function of the family of nowhere differentiable functions is of first category.*

Think about the meaning of the above result in light of the approximation results by Weierstraß and Stone.

5. Differentiation of functions of several variables

The goal of this section is to investigate the local behaviour of functions $f : \mathbb{R}^N \to \mathbb{R}^M$. In particular, we focus on functions that locally behave like affine maps. The reason for such a choice lies on the fact that linear functions can be investigated by the tools of Linear Algebra, and therefore form a class of *easy* objects that are well known.

We will introduce the notion of differentiability for functions $f : \mathbb{R}^N \to \mathbb{R}^M$ by mimicking what you know from the one dimensional case $N = 1$. In the one dimensional case, differentiability is characterized by the existence of a non-vertical tangent line to the graph of the function. Is it also the case in higher dimension? Also, for a scalar function of several variables $f : \mathbb{R}^N \to \mathbb{R}$, it is possible to consider the restriction over a line, namely $t \mapsto f(x + tv)$, or $x, v \in \mathbb{R}^N$. What is the relation between the derivatives of such one dimensional restrictions (called directional derivatives), and the differential of the function?

We will challenge a naïve intuition about differentials, tangent hyperplanes, and directional derivatives by showing that rigor in definitions and proofs is needed in order to not incur in contradictions. This is also what happened historically.

We will detail all the proofs for scalar functions, namely functions $f : \mathbb{R}^N \to \mathbb{R}$, since the general case of $f : \mathbb{R}^N \to \mathbb{R}^M$ follows easily by arguing component by component.

5.1. **Differentiability in the one dimensional case $N = 1$.** Here we recall what is done in the one dimensional case as a motivation for the definition of differentiability in the case $N > 1$, and for the characterizations of such a notion we will investigate. Let $\Omega \subset \mathbb{R}$ be an open interval, and let $\bar{x} \in \Omega$. We recall that f is said to be *differentiable* at $\bar{x}$ if the limit

$$\lim_{x \to \bar{x}} \frac{f(x) - f(\bar{x})}{|x - \bar{x}|} \tag{5.1}$$

exists, and is a real number (namely it is not $\pm\infty$). In this case, the limit is denoted by $f'(\bar{x})$. The existence of the limit in (5.1) has a geometrical interpretation: the existence of a non-vertical tangent line to the graph of f at the point $(\bar{x}, f(\bar{x}))$ (see Figure 11). Such a tangent line is given by

$$\mathrm{Tan}(\mathrm{graph}(f), (\bar{x}, f(\bar{x}))) := \{\, (v, y) \in \mathbb{R} \times \mathbb{R} \,:\, y = L[v - \bar{x}] + f(\bar{x}) \,\}, \tag{5.2}$$

where the linear map $L : \mathbb{R} \to \mathbb{R}$ is defined as

$$L[w] := f'(\bar{x})w,$$

for $w \in \mathbb{R}$. In a more geometrical fashion, this can be stated by saying that all tangent vectors to the graph of f at $(\bar{x}, f(\bar{x}))$ are multiples of the vector $(1, f'(\bar{x})) \in \mathbb{R} \times \mathbb{R}$. Namely, that

$$\lim_{k \to \infty} \frac{f(x_k) - f(\bar{x})}{\lambda_k} = f'(\bar{x})v \tag{5.3}$$

for all sequences $(x_k)_{k \in \mathbb{N}} \subset \mathbb{R}$ with $x_k \to \bar{x}$, and all infinitesimal sequences $(\lambda_k)_{k \in \mathbb{N}}$ such that

$$\lim_{k \to \infty} \frac{x_k - \bar{x}}{\lambda_k} = v.$$

In particular, we say that L is the linear map that best approximates f at $\bar{x}$ at first order, meaning that

$$\lim_{x \to \bar{x}} \frac{|f(x) - f(\bar{x}) - L[x - \bar{x}]|}{|x - \bar{x}|} = 0. \tag{5.4}$$

The above limit can be also written in the following form

$$f(x) = f(\bar{x}) + L[x - \bar{x}] + o(|x - \bar{x}|),$$

little o notation[5].

[5]The *little o* notation was introduced by Landau.

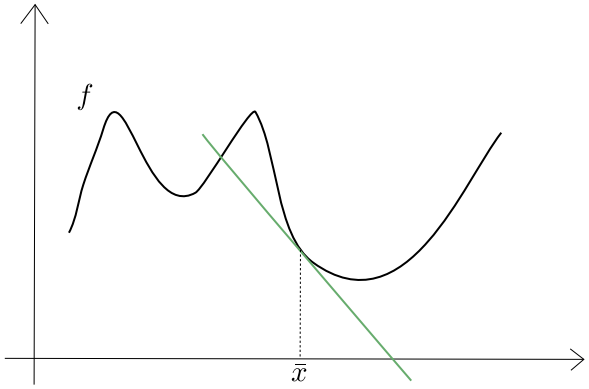

FIGURE 11. The differential of a scalar function defined on $\mathbb{R}$: it has the geometric meaning of non-vertical tangent line.

5.2. **Differentiability in the general case $N \geq 1$.** In the case of several variables, we introduce the notion of differentiability by using the analogous of expression (5.4).

Definition 5.1. Let $\Omega \subset \mathbb{R}^N$ be an open set, and let $\bar{x} \in \Omega$. We say that $f : \Omega \to \mathbb{R}$ is *differentiable* at $\bar{x} \in \Omega$ if there exists a linear map $L : \mathbb{R}^N \to \mathbb{R}$ such that

$$\lim_{x \to \bar{x}} \frac{|f(x) - f(\bar{x}) - L[x - \bar{x}]|}{\|x - \bar{x}\|} = 0.$$

In such a case, we call the map L the *differential of f at $\bar{x}$*, and we denote it by $\mathrm{d}f(\bar{x})$.

Remark 5.2. It is easy to see that if a function is differentiable at a point, then the differential is unique.

Remark 5.3. Note that we are able to define the differential of a function only at an interior point of the domain, since we need to be able to *go in all directions* around the point.

Since linear maps $L : \mathbb{R}^N \to \mathbb{R}$ are identified, by duality, with a vector $w \in \mathbb{R}^N$ via the condition

$$L[v] = \langle w, v \rangle$$

for all $v \in \mathbb{R}^N$, we give a special name to the vector identifying the differential of a map.

Definition 5.4. Let $\Omega \subset \mathbb{R}^N$ be an open set, and let $\bar{x} \in \Omega$. Let $f : \Omega \to \mathbb{R}$ be differentiable at $\bar{x}$. We call *gradient of f at $\bar{x}$*, and we denote it by $\nabla f(\bar{x})$, the vector $w \in \mathbb{R}^N$ that identifies the differential $\mathrm{d}f(\bar{x})$. Namely[6],

$$\mathrm{d}f(\bar{x})[v] = \langle \nabla f(\bar{x}), v \rangle,$$

for all $v \in \mathbb{R}^N$.

Differentiable functions are continuous functions with additional regularity properties, as the following result shows. The proof is left as an exercise to the reader.

Lemma 5.5. *Let $\Omega \subset \mathbb{R}^N$ be an open set, and let $\bar{x} \in \Omega$. Let $f : \Omega \to \mathbb{R}$ be differentiable at $\bar{x}$. Then f is continuous at $\bar{x}$.*

We now collect basic algebraic properties of the differential. For the sake of notation, we state them by using the gradient.

Proposition 5.6. *Let $\Omega \subset \mathbb{R}^N$ be an open set, and let $\bar{x} \in \Omega$. Let $f, g : \Omega \to \mathbb{R}$ be differentiable at $\bar{x}$. Then, the followings hold:*

[6]Note that here we are using the notation with square brackets only for aesthetical reasons. It is the same as writing $\mathrm{d}f(x_0)(v)$.

- Linearity. *For each $\lambda, \mu \in \mathbb{R}$, the function $\lambda f + \mu g$ is differentiable at $\bar{x}$, and*

$$\nabla(\lambda f + \mu g)(\bar{x}) = \lambda \nabla f(\bar{x}) + \mu \nabla g(\bar{x});$$

- Leibniz's rule. *The function fg is differentiable at $\bar{x}$, and*

$$\nabla(fg)(\bar{x}) = g(\bar{x})\nabla f(\bar{x}) + f(\bar{x})\nabla g(\bar{x}).$$

Moreover, let $\varphi : \mathbb{R} \to \mathbb{R}$ be differentiable at $f(\bar{x})$. Then $\varphi \circ f : \mathbb{R}^N \to \mathbb{R}$ is differentiable at $\bar{x}$, and the following chain rule *holds*

$$\nabla(\varphi \circ f)(\bar{x}) = \varphi'(f(\bar{x}))\nabla f(\bar{x}).$$

Finally, let $\gamma : \mathbb{R} \to \mathbb{R}^N$ written as

$$\gamma(t) = (\gamma_1(t), \dots, \gamma_N(t))$$

for $t \in \mathbb{R}$. Assume that there exists $\bar{t} \in \mathbb{R}$ such that $\gamma(\bar{t}) = \bar{x}$ and that, for each $i = 1, \dots, N$, the function $\gamma_i : \mathbb{R} \to \mathbb{R}$ is differentiable at $\bar{t} \in \mathbb{R}$. Then, $f \circ \gamma : \mathbb{R} \to \mathbb{R}$ is differentiable at $\bar{t}$, and the following chain rule *holds*

$$(f \circ \gamma)' = \langle \nabla f(\gamma(\bar{t})), \gamma'(\bar{t}) \rangle,$$

where $\gamma'(t) = (\gamma_1'(t), \dots, \gamma_N'(t))$.

Proof. Step 1: Linearity. From the inequality

$$\begin{aligned} &\frac{|(\lambda f + \mu g)(x) - (\lambda f + \mu g)(\bar{x}) - (\lambda \nabla f(\bar{x}) + \mu \nabla g(\bar{x}))\,[x - \bar{x}]|}{\|x - \bar{x}\|} \\ &\qquad \le |\lambda| \frac{|f(x) - f(\bar{x}) - \nabla f(\bar{x})[x - \bar{x}]|}{\|x - \bar{x}\|} + |\mu| \frac{|g(x) - g(\bar{x}) - \nabla g(\bar{x})[x - \bar{x}]|}{\|x - \bar{x}\|}, \end{aligned}$$

and by using the differentiability of f and g at $\bar{x}$, we get the desired result.

Step 2: Leibniz's rule. It follows in the same way as for the one dimensional case.

Step 3: Chain rule - First case. Write, for $x \in \mathbb{R}^N$ such that $f(x) \neq f(\bar{x})$,

$$\frac{\varphi \circ f(x) - \varphi \circ f(\bar{x})}{\|x - \bar{x}\|} = \frac{\varphi(f(x)) - \varphi(f(\bar{x}))}{f(x) - f(\bar{x})} \frac{f(x) - f(\bar{x})}{\|x - \bar{x}\|}.$$

By using the differentiability of φ and of f, we get that, in the limit as $x \to \bar{x}$, the right-hand side converges to $\varphi'(f(\bar{x}))\nabla f(\bar{x})$.

Step 4: Chain rule - Second case. We have that

$$\frac{f(\gamma(t)) - f(\gamma(\bar{t}))}{t - \bar{t}} = \frac{f(\gamma(t)) - f(\gamma(\bar{t})) - \langle \nabla f(\bar{x}), \gamma(t) - \gamma(\bar{t}) \rangle}{|\gamma(t) - \gamma(\bar{t})|} \frac{|\gamma(t) - \gamma(\bar{t})|}{t - \bar{t}} + \langle \nabla f(\bar{x}), \frac{\gamma(t) - \gamma(\bar{t})}{t - \bar{t}} \rangle.$$

Now, the first fraction converges to zero as $t \to \bar{t}$, thanks to the differentiability of f at $\bar{x}$. Moreover, since

$$\lim_{t \to \bar{t}} \frac{\gamma(t) - \gamma(\bar{t})}{t - \bar{t}} = \gamma'(\bar{t}),$$

we get that the second fraction is bounded uniformly in t close to $\bar{t}$, and that the last term converges to

$$\langle \nabla f(\bar{x}), \gamma'(\bar{t}) \rangle.$$

This concludes the proof. □

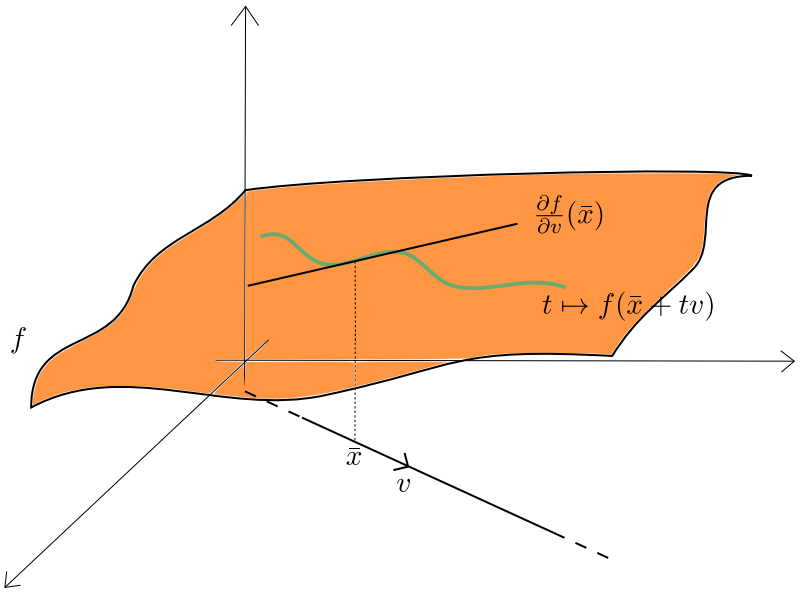

FIGURE 12. The geometric meaning of the directional derivative: it is the derivative of f at $\bar{x}$ along the direction v of the one dimensional function $t \mapsto f(\bar{x} + tv)$.

5.3. **Partial derivatives.** We now turn our attention to the gradient. It encodes all of the information of the differential, and it would then be nice to have a practical way to compute it. We notice that if $f : \Omega \to \mathbb{R}$ is differentiable at $\bar{x}$, and we take $i \in \{1, \dots, N\}$, then

$$\lim_{k \to \infty} \frac{f(\bar{x} + t_k e_i) - f(\bar{x})}{t_k} = \langle \nabla f(\bar{x}), e_i \rangle, \tag{5.5}$$

for all infinitesimal sequences $(t_k)_{k \in \mathbb{N}}$. Here $e_1, \dots, e_N$ denotes the canonical basis of $\mathbb{R}^N$. By using (5.5), we can then compute the components of $\nabla f(\bar{x})$. Geometrically, this corresponds to looking at f only along the directions e_i's, namely to reduce to considering a one dimensional function. This procedure can be carried out for any vector.

Definition 5.7. Let $\Omega \subset \mathbb{R}^N$ be an open set, and let $\bar{x} \in \Omega$. Let $f : \Omega \to \mathbb{R}$, and $v \in \mathbb{R}^N \setminus \{0\}$. We say that f has *directional derivatives* at $\bar{x}$ in the direction v if the limit

$$\lim_{t \to 0} \frac{f(\bar{x} + tv) - f(\bar{x})}{t}$$

exists and is finite. In such a case, it is denoted by $\frac{\partial f}{\partial v}(\bar{x})$, or by $\partial_v f(\bar{x})$.

In case v is one element of the canonical basis of $\mathbb{R}^N$, we call the directional derivative a *partial derivative*. The partial derivative with respect to e_i is usually denoted by $\partial_i f(\bar{x})$.

Remark 5.8. Directional derivatives look at the behavior of the function along *lines*. In particular, this narrows down the kind of singular behaviors of the function. This is to say, that the existence of directional derivatives is a very limited information on the behavior of the function in a neighborhood of a point, since in $\mathbb{R}^N$ with $N > 1$, there is a lot of more space to behave badly, than in just dimension one. This is not surprising, since we already saw in Chapter 3 that even continuity along straight lines is not sufficient to ensure continuity. We will soon see better what consequences that has.

Since the operation to compute partial derivatives is to reduce to one dimensional functions, the following mean value theorem holds. The proof follows directly from the same result in dimension one.

Theorem 5.9 (Lagrange's Mean Value Theorem). *Let $\Omega \subset \mathbb{R}^N$ be an open set, $f : \Omega \to \mathbb{R}$, and $x, y \in \Omega$. Assume that the segment $S := \{x + tv : t \in [0, 1]\}$ is contained in Ω, and that f has directional derivatives in the direction $v := y - x$ at all the points of S. Then,*

$$f(y) - f(x) = \frac{\partial f}{\partial v}(z) = \langle \nabla f(z), (y - x)\rangle,$$

for some $z \in S$.

We now investigate the relation between the existence of partial derivatives and differentiability. By using the same argument as in (5.5), we obtain that differentiability is stronger than admitting partial derivatives.

Lemma 5.10. *Let $\Omega \subset \mathbb{R}^N$ be an open set, and let $\bar{x} \in \Omega$. Let $f : \Omega \to \mathbb{R}$ be differentiable at x_0. Then, f has directional derivatives at $\bar{x}$ in every direction $v \in \mathbb{R}^N$. Moreover, it holds that*

$$\frac{\partial f}{\partial v}(\bar{x}) = \langle \nabla f(\bar{x}), v\rangle,$$

for all $v \in \mathbb{R}^N$. Finally, the map

$$v \mapsto \frac{\partial f}{\partial v}(\bar{x})$$

is linear.

Proof. Let $v \in \mathbb{R}^N \setminus \{0\}$. We have that

$$\begin{aligned}\lim_{t\to 0} \frac{f(\bar{x} + tv) - f(\bar{x})}{t} &= \lim_{t\to 0} \frac{f(\bar{x} + tv) - f(\bar{x}) - \langle \nabla f(\bar{x}), (tv)\rangle}{t} + \langle \nabla f(\bar{x}), v\rangle \\ &= \langle \nabla f(\bar{x}), v\rangle,\end{aligned}$$

where, in the last step, we used the differentiability of f at $\bar{x}$. Thus, we get that

$$\frac{\partial f}{\partial v}(\bar{x}) = \langle \nabla f(\bar{x}), v\rangle,$$

for all $v \in \mathbb{R}^N$. Then, the linearity of the map

$$v \mapsto \frac{\partial f}{\partial v}(\bar{x}) = \langle \nabla f(\bar{x}), v\rangle$$

follows from the linearity of the Euclidean scalar product. □

The opposite is not true in general, as the following remarks show.

Remark 5.11. Directional derivatives at a point can exist, but the function might not be differentiable at that point. Indeed, consider the function $f : \mathbb{R}^2 \to \mathbb{R}$ defined as

$$f(x, y) := yx^{\frac{1}{3}},$$

admits partial derivatives at the origin, but the map is not differentiable at that point.

Remark 5.12. Directional derivative at a point in all directions can exist, but the map $v \mapsto \partial_v f$ might not be linear. Indeed, consider the function $f : \mathbb{R}^2 \to \mathbb{R}$ defined as

$$f(x, y) := \begin{cases} x & \text{if } y = x, \\ \\ 0 & \text{else.} \end{cases}$$

Then, f has directional derivative in all directions at the origin, but the map $v \mapsto \frac{\partial f}{\partial v}((0, 0))$ is not linear.

Remark 5.13. Directional derivative at a point in all directions can exist, the map $v \mapsto \partial_v f$ can be linear, but the function is not continuous at a point. Indeed, consider the function $f : \mathbb{R}^2 \to \mathbb{R}$ defined as

$$f(x, y) := \begin{cases} 1 & \text{if } y = x^2 \neq 0, \\ 0 & \text{else.} \end{cases}$$

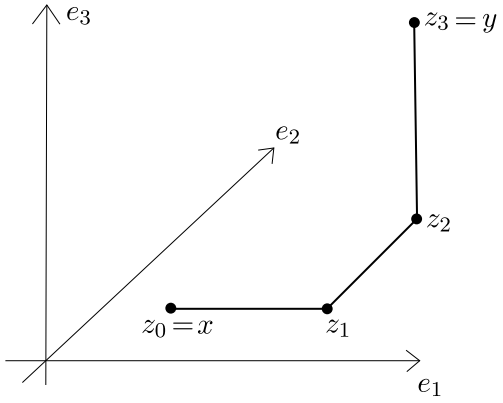

FIGURE 13. The geometric idea of the proof of Theorem 5.15: we go from the point x to the point y by following a path of segments parallel to the orthogonal axes.

Then, f has directional derivatives in all directions at the origin, the map $v \mapsto \frac{\partial f}{\partial v}((0,0))$ is linear, but the function is not continuous at the origin.

The reason why directional derivatives give very weaker information on the local behavior of the function is because we are only looking at the function *restricted to lines*, and not to the behavior of the function close to each line, and not even to the relation between the behavior of the function restricted to lines. This is the difference between directional derivatives and tangent vectors (see Definition 5.19).

As the previous remarks showed, additional assumptions are required for a function admitting partial derivatives at a point in order to be differentiable at the same.

Definition 5.14. Let $\Omega \subset \mathbb{R}^N$ be an open set, and let $f : \Omega \to \mathbb{R}$. We say that f is *of class* C^1 *in* Ω if it has partial derivatives $\partial_1 f(x), \dots, \partial_N f(x)$ for all $x \in \Omega$, and the function

$$x \mapsto \partial_i f(x)$$

is continuous, for all $i = 1, \dots, N$.

The above properties ensures differentiability, as the following result shows.

Theorem 5.15. *Let $\Omega \subset \mathbb{R}^N$ be an open set, and let $f : \Omega \to \mathbb{R}$ be of class C^1 on Ω. Then, f is differentiable at each point of Ω. Moreover, it holds that*

$$\lim_{x \neq y \to \bar{x}} \frac{f(y) - f(x) - \langle \nabla f(\bar{x}), (y - x) \rangle}{\|x - y\|} = 0, \tag{5.6}$$

for all $\bar{x} \in \Omega$.

Proof. Fix $\bar{x} \in \Omega$. We want to prove the validity of (5.6) which, in turn, implies the differentiability of f at $\bar{x}$. Let

$$A := \bar{x} + (-\mu, \mu)^N,$$

for some $\mu > 0$ such that $A \subset \Omega$. For $x, y \in A$, consider the points $z_0, \dots, z_N \in \mathbb{R}^N$ defined as (see Figure 13)

$$z_0 := x, \qquad z_i := x + \sum_{k=1}^{i} \left(\langle (y - x), e_k \rangle \right) e_k.$$

Note that $z_i \in A$ for all $i = 0, \dots, N$, that $z_N = y$, and that $z_i \to \bar{x}$ as $x, y \to \bar{x}$. Moreover, we can write

$$\langle \nabla f(\bar{x}), (y - x) \rangle = \sum_{i=1}^{N} \left(\langle (y - x), e_i \rangle \right) \partial_i f(\bar{x}).$$

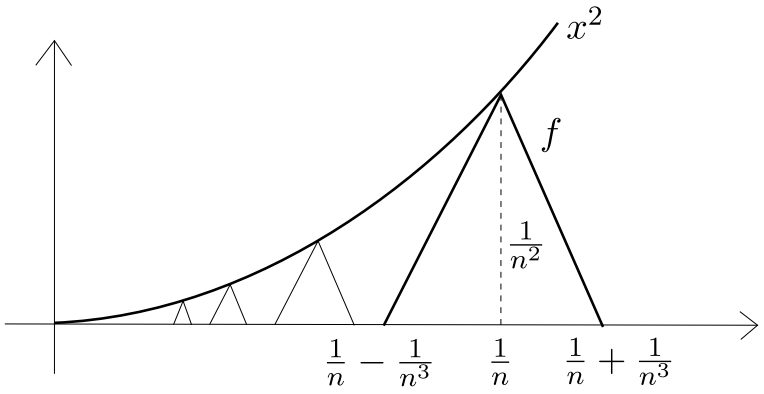

FIGURE 14. A differentiable function for which (5.6) does not hold.

Now, we notice that, since for each $i = 0, \dots, N$ it holds

$$z_i - z_{i-1} = (\langle (y-x), e_i \rangle)\, e_i,$$

the existence of partial derivatives in the directions $e_1, \dots, e_N$ at every point of Ω, allows to apply Theorem 5.9 to get points $\xi_1, \dots, \xi_N \in A$ such that

$$f(z_i) - f(z_{i-1}) = (\langle (y-x), e_i \rangle)\, \partial_i f(\xi_i).$$

Note that, for each $i = 1, \dots, N$, it holds that $\xi_i \to \bar{x}$ as $x, y \to \bar{x}$. We now write

$$\begin{aligned}
\frac{f(y) - f(x) - \langle \nabla f(\bar{x}), (y-x) \rangle)}{\|x-y\|} &= \frac{1}{\|x-y\|} \left[\sum_{i=1}^{N} \left(f(z_i) - f(z_{i-1}) \right) - \sum_{i=1}^{N} (\langle (y-x), e_i \rangle)\, \partial_i f(\bar{x}) \right] \\
&= \frac{1}{\|x-y\|} \left[\sum_{i=1}^{N} (\langle (y-x), e_i \rangle)\, \partial_i f(\xi_i) - \sum_{i=1}^{N} (\langle (y-x), e_i \rangle)\, \partial_i f(\bar{x}) \right] \\
&= \sum_{i=1}^{N} \frac{\langle (y-x), e_i \rangle}{\|x-y\|} \left(\partial_i f(\xi_i) - \partial_i f(\bar{x}) \right). \qquad (5.7)
\end{aligned}$$

Note that

$$\left| \frac{\langle (y-x), e_i \rangle}{\|x-y\|} \right| \leq 1.$$

By letting $x, y \to \bar{x}$ in (5.7), we get that the right-hand side goes to zero, since $\xi_i \to \bar{x}$, and by assumption the partial derivatives are continuous. This concludes the proof of the theorem. □

Remark 5.16. Note that there are functions that are differentiable in an open set, but not C^1(Find one!).

Moreover, note that we cannot remove the assumption that the partial derivatives exists also at the point $\bar{x}$. Indeed, the function

$$f(x,y) := \begin{cases} \dfrac{x^3 y}{\sqrt{x^4 + y^2}} & \text{if } (x,y) \neq (0,0), \\ 0 & \text{if } (x,y) = (0,0). \end{cases}$$

is of class C^1 in $\mathbb{R}^2 \setminus \{(0,0)\}$, but it is not differentiable at the origin (Prove it!).

Remark 5.17. Note that condition (5.6) is stronger than differentiability, since both x and y are free to move. In particular, there are functions that are differentiable at $\bar{x}$ but for which (5.6) does not hold.

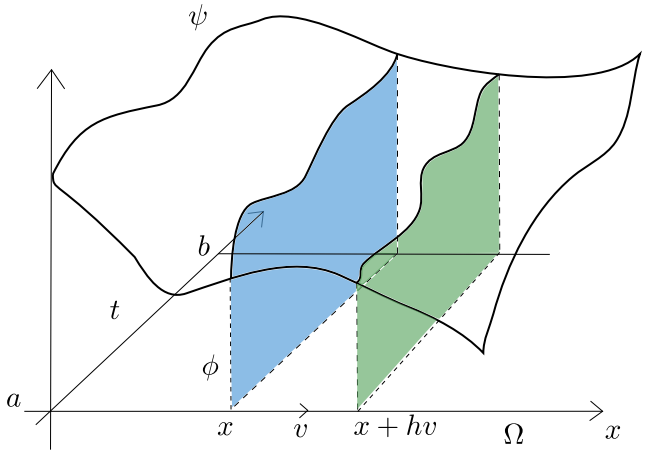

FIGURE 15. The idea behind the proof of the Leibniz formula.

For instance, consider the function $f : \mathbb{R} \to \mathbb{R}$ depicted in Figure 14. Then, f is differentiable at the origin, but if we consider the sequences of points

$$x_n := \frac{1}{n} - \frac{1}{n^3}, \qquad y_n := \frac{1}{n},$$

we have that

$$\lim_{n\to\infty} \frac{f(y_n) - f(x_n)}{\|x_n - y_n\|} = \frac{-\frac{1}{n^2}}{\frac{1}{n^3}} = -n \neq 0.$$

Thus, (5.6) doesn't hold for f.

Finally, we present a formula that will be used later (see Figure 15).

Proposition 5.18 (Leibniz formula)**.** *Let $\psi : [a,b] \times \Omega \to \mathbb{R}$ be a continuous function, where $\Omega \subset \mathbb{R}^N$ is an open set. Then, the function $\phi : \Omega \to \mathbb{R}$ defined as*

$$\phi(x) := \int_a^b \psi(t,x)\, dt$$

is continuous. Moreover, let $v \in \mathbb{R}^N$, and assume that $\partial_v \phi$ is continuous on $[a,b] \times \Omega$. Then, Leibniz formula *holds:*

$$\partial_v \phi(x) = \int_a^b \partial_v \psi(t,x)\, dt.$$

The proof is left to the reader as an exercise.

5.4. **Tangent hyperplane.** Finally, we want to investigate the relation between differentiability and the existence of a tangent hyperplane to the graph. First, we need to clarify the notion of tangent hyperplane.

Definition 5.19. Let $M \geq 1$, $E \subset \mathbb{R}^M$ be a set, and let $\bar{x} \in \mathbb{R}^M$. We say that a vector $v \in \mathbb{R}^M$ is *tangent* to E at $\bar{x}$ if there exist $(x_k)_{k\in\mathbb{N}} \subset E$ with $x_k \to \bar{x}$, and an infinitesimal sequence $(\lambda_k)_{k\in\mathbb{N}}$ such that

$$\lim_{k\to\infty} \frac{x_k - \bar{x}}{\lambda_k} = v.$$

Moreover, we denote by $\mathrm{Tan}(E, \bar{x})$ the set of all such tangent vectors to E at $\bar{x}$ (see Figure 16).

Remark 5.20. The tangent cone at the set E at the point $\bar{x}$ is what directions we *see* the set E when we zoom in at the point $\bar{x}$.

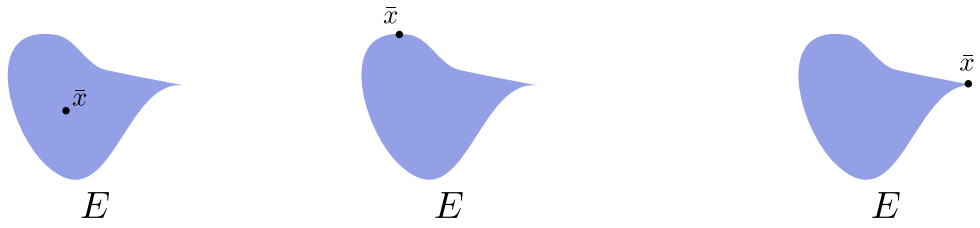

FIGURE 16. A set $E\subset\mathbb{R}^2$ and the tangent cone to it at different points $\bar{x}$. in the left, when $\bar{x}$ is in the interior of E, the tangent cone is the entire space, since E is locally at $\bar{x}$ in all directions. In the middle, when the point $\bar{x}$ belongs to a regular part of the boundary of E, we just see locally E on one side of the tangent plane to the boundary of E. On the right, when $\bar{x}$ is a cusp of the boundary of E, we see locally E as an half-line.

A simple property of the tangent set is the following.

Lemma 5.21. *Let $M\geq 1$, $E\subset\mathbb{R}^M$ be a set, and let $\bar{x}\in\mathbb{R}^M$. Then $\mathrm{Tan}(E,\bar{x})$ is a cone. Namely, if $v\in\mathrm{Tan}(E,\bar{x})$, then $\lambda v\in\mathrm{Tan}(E,\bar{x})$ for all $\lambda\geq 0$.*

Proof. Let $v\in\mathrm{Tan}(E,\bar{x})$ and $\lambda\geq 0$. Then, by definition, there exist $(x_k)_{k\in\mathbb{N}}\subset E$ with $x_k\to\bar{x}$, and an infinitesimal sequence $(\lambda_k)_{k\in\mathbb{N}}$ such that

$$\lim_{k\to\infty}\frac{x_k-\bar{x}}{\lambda_k}=v.$$

Then, by setting

$$\mu_k:=\frac{\lambda_k}{\lambda},$$

we get that

$$\lim_{k\to\infty}\frac{x_k-\bar{x}}{\mu_k}=\lambda v,$$

and thus that $\lambda v\in\mathrm{Tan}(E,\bar{x})$ as desired. □

Remark 5.22. Roughly speaking, the *cone* of tangent vectors is what you see when you zoom in the set E at the point $\bar{z}$. Note that, for a general set E, there is no reason why such vector should exist. And even if the tangent cone is not empty, it could be very wild.

We now focus our attention to the case of the tangent cone to a graph. We first prove a characterization of differentiability in terms of images of tangent vectors via the differential, in the same spirit as (5.3). In particular, we justify the sentence: *the differential sends tangent vectors to tangent vectors*, that will be used repetitively.

Theorem 5.23. *Let $\Omega\subset\mathbb{R}^N$ be an open set, and let $\bar{x}\in\Omega$. Let $f:\Omega\to\mathbb{R}$, and $w\in\mathbb{R}^N$. Then, the followings are equivalent*

(i) f is differentiable at $\bar{x}$ and $\nabla f(\bar{x})=w$;

(ii) For all sequences $(x_k)_{k\in\mathbb{N}}\subset\mathbb{R}$ with $x_k\to\bar{x}$, and all infinitesimal sequences $(\lambda_k)_{k\in\mathbb{N}}$ such that

$$\lim_{k\to\infty}\frac{x_k-\bar{x}}{\lambda_k}=v,$$

for some $v\in\mathbb{R}^N$, it holds that

$$\lim_{k\to\infty}\frac{f(x_k)-f(\bar{x})}{\lambda_k}=\langle w,v\rangle.$$

Proof. Step 1: (i)⇒(ii). If $x_k \neq \bar{x}$, write

$$\frac{f(x_k) - f(\bar{x})}{\lambda_k} = \frac{f(x_k) - f(\bar{x}) - \langle \nabla f(\bar{x}), (x_k - \bar{x})\rangle}{\|x_k - \bar{x}\|} \frac{\|x_k - \bar{x}\|}{\lambda_k} + \langle \nabla f(\bar{x}), \frac{x_k - \bar{x}}{\lambda_k}\rangle.$$

If $x_k = \bar{x}$, note that the left-hand side is zero. Let

$$Z := \{k \in \mathbb{N} : x_k = \bar{x}\}.$$

By the differentiability of f at $\bar{x}$, and the fact that, by (ii),

$$\sup_{k \notin Z} \frac{\|x_k - \bar{x}\|}{\lambda_k} < \infty,$$

we get that the first term on the right-hand side vanishes in the limit $k \to \infty$. Moreover, by (ii), we get that the second term on the right-hand side converges to $\langle \nabla f(\bar{x}), v\rangle$.

Step 2: (ii)⇒(i). Let $(x_k)_{k\in\mathbb{N}} \subset \mathbb{R}$ with $x_k \to \bar{x}$. We would like to prove that

$$\lim_{k\to\infty} \frac{|f(x_k) - f(\bar{x}) - \langle w, (x_k - \bar{x})\rangle|}{\|x_k - \bar{x}\|} = 0. \tag{5.8}$$

For, we apply Urysohn's property (see Proposition 2.6) to the sequence

$$\left(\frac{|f(x_k) - f(\bar{x}) - \langle w, (x_k - \bar{x})\rangle|}{\|x_k - \bar{x}\|}\right)_{k\in\mathbb{N}}.$$

Let $(k_i)_{i\in\mathbb{N}}$ be a subsequence. Since the sequence of vectors

$$\left(\frac{x_{k_i} - \bar{x}}{\|x_{k_i} - \bar{x}\|}\right)_{i\in\mathbb{N}},$$

is uniformly bounded in norm, by using Bolzano-Weierstraß Theorem (see Theorem 2.23) it is possible to find a further subsequence $(x_{k_{i_j}})_{j\in\mathbb{N}}$ such that

$$\lim_{j\to\infty} \frac{x_{k_{i_j}} - \bar{x}}{\|x_{k_{i_j}} - \bar{x}\|} = v, \tag{5.9}$$

for some $v \in \mathbb{R}^N$. Then,

$$\begin{aligned} &\lim_{j\to\infty} \frac{|f(x_{k_{i_j}}) - f(\bar{x}) - \langle w, (x_{k_{i_j}} - \bar{x})\rangle|}{\|x_{k_{i_j}} - \bar{x}\|} \\ &= \lim_{j\to\infty} \left| \frac{f(x_{k_{i_j}}) - f(\bar{x})}{\|x_{k_{i_j}} - \bar{x}\|} - \langle w, \frac{x_{k_{i_j}} - \bar{x}}{\|x_{k_{i_j}} - \bar{x}\|}\rangle \right| \\ &= 0. \end{aligned} \tag{5.10}$$

where in the last step we used (5.9) together with the assumption. Since the limit in (5.10) is independent of the subsequence $(k_i)_{i\in\mathbb{N}}$, Urysohn's property yields (5.8) as desired. □

We now show that differentiability implies that the tangent cone to a graph is a linear space of dimension N. The proof follows directly from the definition of the tangent cone together with the characterization of differentiability provided in Theorem 5.23.

Lemma 5.24. *Let $\Omega \subset \mathbb{R}^N$ be an open set, and let $\bar{x} \in \Omega$. Let $f : \Omega \to \mathbb{R}$ be differentiable at $\bar{x}$. Then,* $\mathrm{Tan}(\mathrm{graph}(f), (\bar{x}, f(\bar{x})))$ *is a linear space of dimension N. In particular,*

$$\mathrm{Tan}\left(\mathrm{graph}(f), (\bar{x}, f(\bar{x}))\right) = \left\{ \left(v, \frac{\partial f}{\partial v}(\bar{x})\right) : v \in \mathbb{R}^N \right\}.$$

In particular, it is the linear space generated by the vectors $(e_i, \partial_{e_i} f(\bar{x}))$, for $i = 1, \dots, N$.

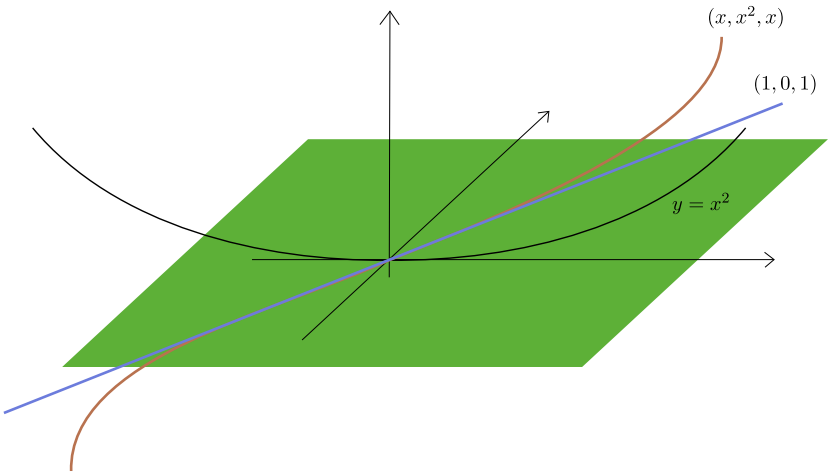

FIGURE 17. The function f of Remark 5.25: its tangent cone at the origin is the horizontal plane (depicted in green), and the line along the segment $(1, 0, 1)$ (depicted in blue).

Remark 5.25. Note that, for a function that is not differentiable at a point, the tangent space at the corresponding point on the graph may fail to be a linear space. Indeed, consider the function $f : (-1, 1)^2 \to \mathbb{R}$ defined as

$$f(x, y) \coloneqq \begin{cases} x & \text{if } y = x^2, \\ 0 & \text{else.} \end{cases}$$

Then, the tangent cone at the graph of f at the origin is the set (see Figure 17)

$$\{(x, y, 0) : (x, y) \in \mathbb{R}^2\} \cup \{(x, 0, x) : x > 0\},$$

which is very far from being a linear space.

Remark 5.26. It is possible to use the idea of the previous remark, and make the tangent cone fail to be a linear space even in a more spectacular way. Indeed, consider the function $f : \mathbb{R}^2 \to \mathbb{R}$ defined as

$$f(x, y) \coloneqq \begin{cases} \operatorname{sign}(y)\theta x & \text{if } |y| = \theta x^2 \text{ for some } \theta \in (0, 1], \\ \operatorname{sign}(y)(2 - \theta)x & \text{if } |y| = \theta x^2 \text{ for some } \theta \in (1, 2], \\ 0 & \text{else,} \end{cases}$$

where $\operatorname{sign}(y)$ is the sign of y. Then, f is continuous at the origin, that it admits directional derivatives in all directions, and that the map $v \mapsto \frac{\partial f}{\partial v}((0, 0))$ is linear.

Nevertheless, the tangent space to the graph of f at the origin is given by

$$\{(x, y, 0) : (x, y) \in \mathbb{R}^2\} \cup \{(x, 0, z) : |z| \leq |x|\}.$$

Remark 5.27. In the one dimensional case $N = 1$, differentiability at a point is equivalent to the existence of a non-vertical tangent line. In the case of several variables, the situation is more complicated, since the existence of a tangent hyperplane to the graph of a function is not sufficient to ensure differentiability. Indeed, consider the function $f : \mathbb{R}^2 \to \mathbb{R}$ defined as

$$f(x, y) \coloneqq \begin{cases} 1 & \text{if } y = x^2 \neq 0, \\ 0 & \text{else.} \end{cases}$$

Set $\bar{x} \coloneqq (0,0)$. Then it holds (prove it!) that $\mathrm{Tan}(\mathrm{graph}(f),(\bar{x},f(\bar{x}))$ is a linear space of dimension N, but f cannot be differentiable at $\bar{x}$, since it is not continuous at that point.

We now present a weaker version of the previous result, that will be useful in the following.

Lemma 5.28. *Let $f:\Omega\to\mathbb{R}$, where $\Omega\subset\mathbb{R}^N$ is an open set. Fix $\bar{x}\in\Omega$, and $v\in\mathbb{R}^N$, and assume that there exists $\partial_v f(\bar{x})$. Then,*

$$(v,\partial_v f(\bar{x}))\in \mathrm{Tan}(\mathrm{graph}(f),(\bar{x},f(\bar{x})).$$

We now provide a characterization of differentiability based on three necessary condition that we derived above (see Lemma 5.5, Lemma 5.10, and Lemma 5.24).

Theorem 5.29. *Let $f:\Omega\to\mathbb{R}$, where $\Omega\subset\mathbb{R}^N$ is an open set. Then, f is differentiable at $\bar{x}\in\Omega$ if and only if*

(i) f is continuous at $\bar{x}$;

(ii) For each $v\in\mathbb{R}^N$ the directional derivative $\frac{\partial f}{\partial v}(\bar{x})$ exists. Moreover, the map

$$v\mapsto \frac{\partial f}{\partial v}(\bar{x})$$

is linear;

(iii) The tangent cone $\mathrm{Tan}\,(\mathrm{graph}(f),(\bar{x},f(\bar{x}))$ is a linear space of dimension N.

Proof. Step 1: Necessity. Assume f is differentiable at $\bar{x}$. Then, (i), (ii), and (iii) follows from Lemma 5.5, Lemma 5.10, and Lemma 5.24, respectively.

Step 2: Sufficiency. Assume that f satisfies (i), (ii), and (iii). We divide the argument in several steps.

Step 2.1. We claim that

$$\mathrm{Tan}\,(\mathrm{graph}(f),(\bar{x},f(\bar{x}))=\left\{\left(v,\frac{\partial f}{\partial v}(\bar{x})\right)\,:\,v\in\mathbb{R}^N\right\}. \tag{5.11}$$

Indeed, from (ii) together with Lemma 5.28, we get that the right-hand side is included in the left-hand side. Since by (ii) we have that the map

$$v\mapsto \frac{\partial f}{\partial v}(\bar{x})$$

is linear, we obtain that the right-hand side is a linear space of dimension N and, thanks to (i), also the left-hand side is, we get also the opposite inequality.

Step 2.2. We claim that there exists a linear map $L:\mathbb{R}^N\to\mathbb{R}$ such that

$$\left\{\left(v,\frac{\partial f}{\partial v}(\bar{x})\right)\,:\,v\in\mathbb{R}^N\right\}=\mathrm{graph}(L). \tag{5.12}$$

Namely, the tangent cone to the graph of f at $(\bar{x},f(\bar{x})$ is the graph of the linear function L.

Indeed, from (5.11), we have that the tangent cone to the graph of f at $(\bar{x},f(\bar{x}))$ is a non-vertical hyperplane in $\mathbb{R}^N\times\mathbb{R}$. Thus, it is the graph of a linear map $L:\mathbb{R}^N\to\mathbb{R}$. In particular, we have that

$$L[v]=\sum_{i=1}^N \partial_i f(\bar{x})v_i,$$

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

Step 2.3. We now prove that, given a sequence $(x_k)_{k\in\mathbb{N}}\subset\Omega$ converging to $\bar{x}$, it holds

$$\lim_{k\to\infty}\frac{|f(x_k)-f(\bar{x})-L[x_k-\bar{x}]|}{\|x_k-\bar{x}\|}=0.$$

We will do that by showing that, from each of such sequence, we can extract a subsequence such that the above limit holds. Thus, we conclude that the limit holds for the entire sequence by using the Urysohn property (see Proposition 2.6).

Since the vectors

$$\frac{x_k - \bar{x}}{\|x_k - \bar{x}\|}$$

have uniformly bounded norm, thanks to the Bolzano-Weierstraß Theorem (see Theorem 2.23), we can extract a subsequence $(x_{k_i})_{i\in\mathbb{N}}$ such that

$$\lim_{i\to\infty} \frac{x_{k_i} - \bar{x}}{\|x_{k_i} - \bar{x}\|} = v \in \mathbb{R}^N.$$

We now have two possibilities. Either there exists a further subsequence (not relabeled) such that

$$\sup_{i\in\mathbb{N}} \frac{|f(x_{k_i}) - f(\bar{x})|}{\|x_{k_i} - \bar{x}\|} < \infty, \tag{5.13}$$

or for all possible subsequences, the above quantity is infinite. In the former case, thanks again to the Bolzano-Weierstraß Theorem (see Theorem 2.23), we can assume that

$$\lim_{i\to\infty} \frac{f(x_{k_i}) - f(\bar{x})}{\|x_{k_i} - \bar{x}\|} = w \in \mathbb{R}.$$

Thus, $(v, w) \in \mathbb{R}^N \times \mathbb{R}$ is a tangent vector to the graph of f at the point $(\bar{x}, f(\bar{x}))$, and by (5.12) it holds $w = L[v]$. Therefore,

$$\lim_{i\to\infty} \frac{|f(x_{k_i}) - f(\bar{x}) - L(x_{k_i} - \bar{x})|}{\|x_k - \bar{x}\|} = 0.$$

In case there is no subsequence for which (5.13) is satisfied, we argue as follows. By continuity of f at $\bar{x}$, the sequence $|f(x_{k_i}) - f(\bar{x})|$ is infinitesimal. Again by the Bolzano-Weierstraß Theorem (see Theorem 2.23), we can assume that

$$\lim_{i\to\infty} \frac{f(x_{k_i}) - f(\bar{x})}{|f(x_{k_i}) - f(\bar{x})|} = w \in \{\pm 1\}. \tag{5.14}$$

Now, by assumption

$$\lim_{i\to\infty} \frac{\|x_{k_i} - \bar{x}\|}{|f(x_{k_i}) - f(\bar{x})|} = 0.$$

Thus, by considering the infinitesimal sequence $\lambda_i := |f(x_{k_i}) - f(\bar{x})|$, we get that

$$(0, w) \in \mathrm{Tan}\left(\mathrm{graph}(f), (\bar{x}, f(\bar{x}))\right).$$

This gives the desired contradiction. Indeed, by (5.11), we would get

$$w = L[0].$$

By the linearity of the map L, we have that $L[0] = 0$, but, by (5.14), $w \neq 0$. This concludes the proof. □

As it turns out, continuity and the existence of a non vertical tangent hyperplane are sufficient to ensure differentiability.

Proposition 5.30. *Let $\Omega \subset \mathbb{R}^N$ be an open set, and let $\bar{x} \in \Omega$. Let $f : \Omega \to \mathbb{R}$ be a function that is continuous at $\bar{x}$ and such that $\mathrm{Tan}(\mathrm{graph}(f), (\bar{x}, f(\bar{x}))$ is the graph of a linear map $L : \mathbb{R}^N \to \mathbb{R}$. Then, f is differentiable at $\bar{x}$.*

The proof of this result follows easily from Theorem 5.29 and is left to the reader.

5.5. **Differentiability of functions from $\mathbb{R}^N$ to $\mathbb{R}^M$.** The notion of differentiability for vector valued functions mimic that for scalar functions.

Definition 5.31. Let $\Omega \subset \mathbb{R}^N$ be an open set, and let $\bar{x} \in \Omega$. We say that $f : \Omega \to \mathbb{R}^M$ is *differentiable* at $\bar{x} \in \Omega$ if there exists a linear map $L : \mathbb{R}^N \to \mathbb{R}^M$ such that

$$\lim_{x \to \bar{x}} \frac{\|f(x) - f(\bar{x}) - L(x - \bar{x})\|}{\|x - \bar{x}\|} = 0.$$

In such a case, we call the map L the *differential of f at $\bar{x}$*, and we denote it by $\mathrm{d}f(\bar{x})$.

As for the case of scalar functions, the differential can be identified by duality by a matrix.

Definition 5.32. Let $f : \Omega \to \mathbb{R}^M$ be differentiable at $x \in \Omega$. We denote by $Jf(x)$, and we call it the *Jacobian matrix* of f at x, the $M \times N$ matrix defined by the relation

$$\mathrm{d}f(\bar{x})[v] = Jf(x)v,$$

for all $v \in \mathbb{R}^N$.

The differentiability of a vector valued function is equivalent to the differentiability of its components.

Lemma 5.33. *Let $f : \Omega \to \mathbb{R}^M$, where $\Omega \subset \mathbb{R}^N$ is an open set, and write $f = (f_1, \dots, f_M)$. Then, f is differentiable at a point $x \in \Omega$ if and only if each of its components $f_1, \dots f_M$ are differentiable at x.*

The proof is left as an exercise for the reader.

Remark 5.34. In particular, the above result allows us to write

$$Jf(x) = \begin{pmatrix} \nabla f_1(x) \\ \nabla f_2(x) \\ \dots \\ \nabla f_M(x) \end{pmatrix}$$

Namely, the ith row of $Jf(x)$ is $\nabla f_i(x)$.

Remark 5.35. A difference with scalar functions, is that, in general, the Lagranges' Mean Value Theorem (see Theorem 5.9) is not valid. Indeed, consider the function $f : \mathbb{R} \to \mathbb{R}^2$ given by $f(t) \coloneqq (\cos t, \sin t)$. Then, $f(0) = f(2\pi)$, but f' never vanishes.

The fact is that, for vectorial functions, we cannot make sure that *all* of the components of the function obey that Lagranges' Mean Value Theorem at the *same* point.

As for scalar functions, it is possible to introduce the concept of partial (and directional) derivative.

Definition 5.36. Let $\Omega \subset \mathbb{R}^N$ be an open set, and let $\bar{x} \in \Omega$. Let $f : \Omega \to \mathbb{R}^M$, and $v \in \mathbb{R}^N$. We say that f has *directional derivatives* at $\bar{x}$ in the direction v if the limit

$$\lim_{t \to 0} \frac{f(\bar{x} + tv) - f(\bar{x})}{t}$$

exists and is finite. In such a case, it is denoted by $\frac{\partial f}{\partial v}(\bar{x})$, or by $\partial_v f(\bar{x})$.

In case v is one element of the canonical basis of $\mathbb{R}^N$, we call the directional derivative a *partial derivative*.

Remark 5.37. Partial derivatives allow to give a different point of view on the Jacobian matrix:

$$Jf(x) = (\partial_1 f(x) \dots \partial_N f(x))$$

Namely, the ith column of $Jf(x)$ is $\partial_i f(x)$.

Finally, despite tangent planes to the graphs of a vector valued function are not easy to imagine, the analogous of Theorem 5.23 holds. Thus, also for vector value functions, *the differential sends tangent vectors to tangent vectors.*

Theorem 5.38. *Let $\Omega \subset \mathbb{R}^N$ be an open set, and let $\bar{x} \in \Omega$. Let $f : \Omega \to \mathbb{R}^M$, and A be an $M \times N$ matrix. Then, the followings are equivalent*

(i) f is differentiable at $\bar{x}$ and $Jf(\bar{x}) = A$;

(ii) For all sequences $(x_k)_{k\in\mathbb{N}} \subset \mathbb{R}^N$ with $x_k \to \bar{x}$, and all infinitesimal sequences $(\lambda_k)_{k\in\mathbb{N}}$ such that

$$\lim_{k\to\infty} \frac{x_k - \bar{x}}{\lambda_k} = v,$$

for some $v \in \mathbb{R}^N$, it holds that

$$\lim_{k\to\infty} \frac{f(x_k) - f(\bar{x})}{\lambda_k} = Av.$$

The chain rule (see Proposition 5.6) also holds for vector valued differentiable functions.

Proposition 5.39. *Let $f : \mathbb{R}^N \to \mathbb{R}^M$, and $g : \mathbb{R}^M \to \mathbb{R}^k$ be functions such that f is differentiable at $\bar{x} \in \mathbb{R}^N$, and g is differentiable at $f(\bar{x})$. Then, $g \circ f$ is differentiable at $\bar{x}$, and*

$$\mathrm{d}(g \circ f)(\bar{x})[v] = \mathrm{d}g(f(\bar{x}))\left[\mathrm{d}f(\bar{x})[v]\right],$$

for all $v \in \mathbb{R}^N$. In particular, it holds that

$$J(g \circ f)(\bar{x}) = Jg(f(\bar{x}))\, Jf(\bar{x}).$$

To state the Leibniz rule for vector valued functions, we first need to introduce a convenient notation to express the Jacobian matrix of the product between a scalar and a vectorial function.

Definition 5.40. Let $a \in \mathbb{R}^M$ and $b \in \mathbb{R}^N$. The *tensor product* between a and b is the $M \times N$ matrix $a \otimes b$ defined as

$$(a \otimes b)_{ij} \coloneqq a_i b_j,$$

for $i = 1, \dots, M$ and $j = 1 \dots, N$.

Proposition 5.41. *Let $\varphi : \mathbb{R}^N \to \mathbb{R}$ and $f : \mathbb{R}^N \to \mathbb{R}^M$ be differentiable functions at a point $\bar{x} \in \mathbb{R}^N$. Then, the function $\varphi f : \mathbb{R}^N \to \mathbb{R}^M$ is differentiable at $\bar{x}$, and follows*

$$J(\varphi f) = f \otimes \nabla\varphi + \varphi\, Jf.$$

Remark 5.42. Note that the tensor product is *not* commutative! In particular, in the above formula it is important to take $f \otimes \nabla\varphi$ and not $\nabla\varphi \otimes f$!

6. Vector fields and gradients

Are all vector fields gradients of scalar functions? Namely, given a vector field $V : \mathbb{R}^N \to \mathbb{R}^N$, is it possible to find a function $f : \mathbb{R}^N \to \mathbb{R}$ such that $V = \nabla f$?

Note that, for continuous vector fields $V : \mathbb{R} \to \mathbb{R}$, the question is trivial (why?).

The answer to the above question is: no, not all vector fields are gradient vector fields. In order to understand what can go wrong, we will consider two examples in $\mathbb{R}^2$. The first is the vector field $V : \mathbb{R}^2 \to \mathbb{R}^2$ defined as

$$V(x, y) := (y^2, x^2).$$

We claim that V is not a gradient vector field. Indeed, if by absurd there existed a function $f : \mathbb{R}^2 \to \mathbb{R}$ such that $V = \nabla f$, then we would have

$$\partial_1 f(x, y) = y^2 \quad \Rightarrow \quad f(x, y) = xy^2 + g_1(y),$$

and

$$\partial_2 f(x, y) = x^2 \quad \Rightarrow \quad f(x, y) = yx^2 + g_2(x),$$

for some functions $g_1, g_2 : \mathbb{R} \to \mathbb{R}$. Clearly, the above two conditions are not compatible with each other. This is because V does not satisfy an *algebraic* condition that holds gradient vector fields.

The second example we consider is the vector field $V : \mathbb{R}^2 \setminus \{0\} \to \mathbb{R}^2$ defined as

$$V(x, y) := \left(\frac{-y}{x^2 + y^2}, \frac{x}{x^2 + y^2} \right).$$

We claim that V is not a gradient vector field. Indeed, if we assume by contradiction that there existed a function $f : \mathbb{R}^2 \to \mathbb{R}$ such that $V = \nabla f$. Consider the curve $\gamma : [0, 2\pi] \to \mathbb{R}^2$ given by

$$\gamma(t) := (\cos t, \sin t),$$

and the composite function $f \circ \gamma$. Then, since $\gamma(0) = \gamma(2\pi)$, we would have

$$\begin{aligned} 0 = f \circ \gamma(2\pi) - f \circ \gamma(0) &= \int_0^{2\pi} (f \circ \gamma)' \, dt \\ &= \int_0^{2\pi} \langle \nabla f(\gamma(t)), \gamma'(t) \rangle \, dt = \int_0^{2\pi} \langle V(\gamma(t)), \gamma'(t) \rangle \, dt = \int_0^{2\pi} 1 \, dt = 2\pi, \end{aligned}$$

which gives the desired contradiction.

The problem here is that, if you follow a gradient vector field along a closed path, you have to go up as much as you go down. The above computation shows that this is not the case for the vector field V (see Figure 18).

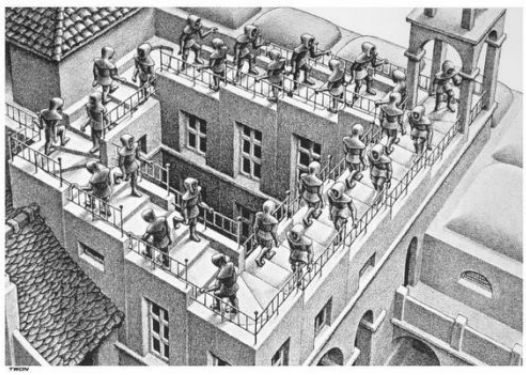

Figure 18. If you follow a gradient vector field over a closed path, you cannot just go up, like the people in the *Ascending and Descending* illustration by Escher of 1960. Same ideas were developed by Oscar Reutersvärd.

Nevertheless, contrary to the previous example, if we consider it in the half space $\{(x, y) \in \mathbb{R}^2 : y > 0\}$, we get that $V = \nabla f$, where

$$f(x, y) \coloneqq -\arctan\left(\frac{x}{y}\right).$$

Thus, there is an interaction between the vector field and the *topology* of the domain where we consider it.

As we will see, these two examples are the prototypes of things that can go wrong when we try to find a *primitive* to a vector field. Nevertheless, we will see that it is always possible to decompose a vector field in a gradient part, and another part. For the case $N = 3$, we will relate this latter to the rotation of the vector field around points. This is the so called Helmholtz Decomposition Theorem. Such a result has both an interest in mathematics and in physics. For the former, it is a way to understand the relation between the family of vector fields, and the family of gradients. For the latter, it is related to conservative forces, fluid dynamics, and electromagnetism. We will talk about these applications at the end of the chapter.

In order to investigate the relation between vector fields and gradients, we will introduce the notion of *forms*: these are objects of extreme importance in several areas of science, like analysis, geometry, and physics, to mention some.

6.1. **Schwarz's Theorem.** In order to answer the above question, we first derive a necessary condition that a gradient field has to satisfy.

Theorem 6.1 (Schwarz's Theorem)**.** *Let $f : \mathbb{R}^N \to \mathbb{R}$ be of class C^2. Then,*

$$\frac{\partial^2 f}{\partial e_i \partial e_j}(x) = \frac{\partial^2 f}{\partial e_j \partial e_i}(x),$$

for all $x \in \mathbb{R}^N$, and all $i, j = 1, \dots, N$.

Proof. Without loss of generality, we can assume $N = 2$, and let $x = (a, b)$. The idea of the proof is based on the following trivial identity:

$$\begin{aligned} &f(a+h, b+k) - f(a+h, b) + f(a+h, b) - f(a, b) \\ &\qquad = f(a+h, b+k) - f(a, b+k) + f(a, b+k) - f(a, b), \end{aligned} \tag{6.1}$$

for $h, k \in \mathbb{R}$. Note that the identity holds because, geometrically (see Figure 19), if you make one step east, one north, one west, and one south, you can back to the same place. This is because $\mathbb{R}^N$ is *flat*!

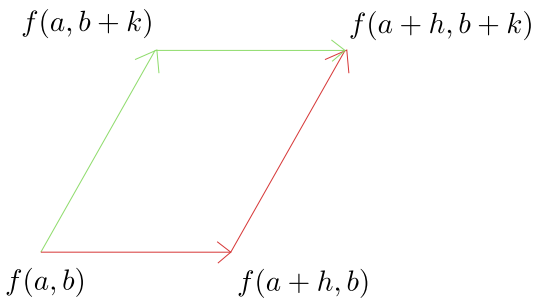

FIGURE 19. The geometric idea behind the proof of Schwarz's Theorem. Following the green path takes you to the same point as following the red path.

Rearranging the terms in (6.1), we obtain

$$f(a+h, b+k) - f(a+h, b) - f(a, b+k) + f(a, b)$$

$$= f(a+h,b+k) - f(a,b+k) - f(a+h,b) + f(a,b). \tag{6.2}$$

Since f is a C^1 function, the idea is to use Lagrange's Mean Value Theorem (see Theorem 5.9) to rewrite the several differences

$$f(a+h,b+k) - f(a+h,b), \qquad f(a,b+k) - f(a,b),$$

on the left-hand side, and

$$f(a+h,b+k) - f(a,b+k), \qquad f(a+h,b) - f(a,b),$$

on the right-hand side Note that, if we consider them separately, we would obtain different points where the derivatives are computed. This would make the computations a bit more involved. A better way to do that, is to see the left hand side as

$$u(h,k) - u(0,k),$$

where

$$u(h,k) \coloneqq f(a+h,b+k) - f(a+h,b),$$

and the right-hand side as

$$v(h,k) - v(h,0),$$

where

$$v(h,k) \coloneqq f(a+h,b+k) - f(a,b+k).$$

In such a way, (6.2) becomes

$$h\left[\partial_1 f(a+\widetilde{h},b+k) - \partial_1 f(a+\widetilde{h},b)\right] = k\left[\partial_2 f(a+h,b+\widetilde{k}) - \partial_2 f(a,b+\widetilde{k})\right],$$

for some $|\widetilde{h}| < |h|$, and $|\widetilde{k}| < |k|$. Using again Lagrange's Mean Value Theorem (see Theorem 5.9), since f is a C^2 function, we get

$$hk\,\partial_2\partial_1 f(a+\widetilde{h},b+\bar{k}) = kh\,\partial_1\partial_2 f(a+\bar{h},b+\widetilde{k}),$$

for some $|\bar{h}| < |h|$, and $|\bar{k}| < |k|$. Thus, dividing both sides by hk, we get

$$\partial_2\partial_1 f(a+\widetilde{h},b+\bar{k}) = \partial_1\partial_2 f(a+\bar{h},b+\widetilde{k}).$$

Thus, by sending $h,k \to 0$, and using the continuity of the second partial derivatives, we get

$$\partial_2\partial_1 f(a,b) = \partial_1\partial_2 f(a,b).$$

This concludes the proof of the theorem. □

Remark 6.2. In the above proof, one could ask why introducing the functions u and v to estimate the difference of the f in equation (6.2). Well, try to estimate it in a different way and see what goes wrong.

Remark 6.3. The continuity of the second order partial derivatives is needed in order for the above result to be true. Indeed, Peano found the following counterexample. Consider the function $f : \mathbb{R}^2 \to \mathbb{R}$ defined as (see Figure 20)

$$f(x,y) \coloneqq \begin{cases} \dfrac{xy(x^2-y^2)}{x^2+y^2} & \text{if } (x,y) \neq (0,0), \\ \\ 0 & \text{else,} \end{cases}$$

Then, f is twice differentiable, but the second order derivatives are not continuous at the origin. It is easier to understand the above function in polar coordinates: it writes as

$$f(r,\theta) \coloneqq \frac{r\sin(4\theta)}{4}.$$

Namely, for each fixed $r > 0$, the function $\theta \mapsto f(r,\theta)$ oscillates in a way that is not possible to approximate with a parabola (namely by using second derivatives).

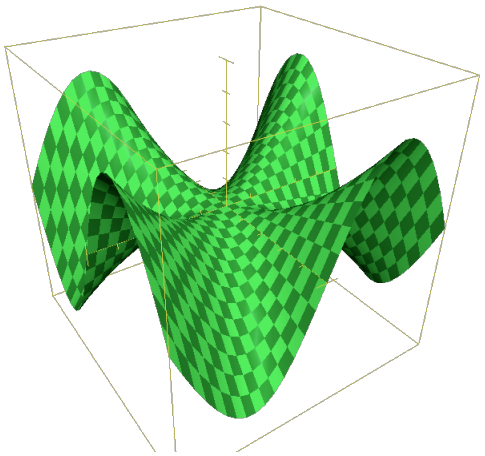

FIGURE 20. The example by Peano: the function f oscillates around the origin in such a way that it is not possible to approximate with a parabola (namely by using second derivatives).

Remark 6.4. Schwarz's Theorem implies that the *Hessian matrix*, namely the $N \times N$ matrix of the second order partial derivatives, is symmetric.

6.2. **Differential forms.** We now want to take a different look at the differential of a scalar map. Let $f : \mathbb{R}^N \to \mathbb{R}$ be a C^1 map. At each point $x \in \mathbb{R}^N$, the differential $\mathrm{d}f(x)$ is a linear map from $\mathbb{R}^N \to \mathbb{R}$, namely an element of $\mathcal{L}(\mathbb{R}^N; \mathbb{R})$ (see Example 1.16). Indeed,

$$\mathrm{d}f(x)[v] = \langle \nabla f(x), v \rangle,$$

for all $v \in \mathbb{R}^N$. Therefore, the function

$$x \mapsto \mathrm{d}f(x)$$

is a map from $\mathbb{R}^N$ to the space $\mathcal{L}(\mathbb{R}^N; \mathbb{R})$. Since the function f is C^1, such a map is continuous, if we equip the target space with the operator norm (see Example 1.16 - Check this fact!). Such an object is a particular case of what is called a *(multilinear) form*, and they are fundamental in Analysis, Geometric Integration Theory, as well as in Geometry, and Physics. This was pioneered in the work by Grassmann. For the moment, we will only consider the class of forms to which the differential belongs to, namely 1-forms.

Definition 6.5. A map $\omega : \mathbb{R}^N \to \mathcal{L}(\mathbb{R}^N; \mathbb{R})$ is called a 1-form. The space of 1-forms is denoted by $\Lambda^1(\mathbb{R}^N)$.

Remark 6.6. What we are saying is this. Take a linear map $L : \mathbb{R}^N \to \mathbb{R}$. Then, by duality, there exists a vector $V \in \mathbb{R}^N$ such that

$$L[v] = \langle V, v \rangle,$$

for all $v \in \mathbb{R}^N$. Now, consider a map $L : \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}$ such that, for each $x \in \mathbb{R}^N$, the map

$$v \mapsto L(x, v) = L(x)[v],$$

is linear. The right-hand side is just a different way to write the same object, where the two variables, x and v are separated, because they play a different role. Then, by duality as before, for each $x \in \mathbb{R}^N$ we can find a vector $V(x) \in \mathbb{R}^N$ such that

$$L(x)[v] = \langle V(x), v \rangle,$$

for all $v \in \mathbb{R}^N$. This is what a 1-form is!

Motivated by the previous example, we want to define a standard way to write a 1-form. In the same way as there is a standard way to write vectors in $\mathbb{R}^N$, namely by using the basis $e_1, \dots, e_N$, there are special 1-forms that allow to write all of the others. First, we remark that the space of 1-forms is a linear vector space.

Definition 6.7. On $\Lambda^1(\mathbb{R}^N)$ we define a natural notion of addition of 1-forms, and of multiplication of a 1-form with a scalar as follows. Given $\omega_1, \omega_2 \in \Lambda^1(\mathbb{R}^N)$ and $\lambda \in \mathbb{R}$, we define

$$(\omega_1 + \omega_2)(x)[v] \coloneqq \omega_1(x)[v] + \omega_2(x)[v],$$

and

$$(\lambda\omega_1)(x)[v] \coloneqq \lambda\omega_1(x)[v],$$

for every $v \in \mathbb{R}^N$, respectively.

Lemma 6.8. *The space $\Lambda^1(\mathbb{R}^N)$ is a linear vector space, with respect to the notion of addition and scalar multiplication defined above.*

We are now in position to define the standard basis of the linear vector space $\Lambda^1(\mathbb{R}^n)$.

Definition 6.9. We define $\mathrm{d}x_1, \dots, \mathrm{d}x_N \in \Lambda^1(\mathbb{R}^N)$ as

$$\mathrm{d}x_i(y)[v] \coloneqq v_i,$$

for $y \in \mathbb{R}^N$, and $i = 1, \dots, N$, where $v = (v_1, \dots, v_N) \in \mathbb{R}^N$.

Remark 6.10. Let $\omega \in \Lambda^1(\mathbb{R}^N)$. Then, it is possible to write

$$\omega(x) = \sum_{i=1}^N \omega_i(x)\,\mathrm{d}x_i,$$

for $\omega_i : \mathbb{R}^N \to \mathbb{R}$, for $i = 1, \dots, N$. In particular, we have that

$$\omega(x)[v] = \sum_{i=1}^N \omega_i(x) v_i,$$

for all vectors $v \in \mathbb{R}^N$, and all $x \in \mathbb{R}^N$.

By identifying the 1-form $\omega \in \Lambda^1(\mathbb{R}^N)$ with the vector of its coordinates $(\omega_1, \dots, \omega_N)$, that we will denote with an abuse of notation with $\omega : \mathbb{R}^N \to \mathbb{R}^N$, we can write

$$\omega(x)[v] = \langle \omega(x), v \rangle,$$

for all $v \in \mathbb{R}^N$. On the left-hand side, with ω we intend the 1-form, while on the right-hand side, with ω we intend the vector of its coordinates with respect to the standard basis.

Remark 6.11. In particular, we get that

$$\mathrm{d}f(x) = \sum_{i=1}^N \partial_i f(x)\,\mathrm{d}x_i,$$

for a C^1 function f. Note that we omitted the dependence of the $\mathrm{d}x_i$'s on the space variable $x \in \mathbb{R}^N$, since they are constant linear maps. Moreover,

$$\mathrm{d}f(x)[v] = \sum_{i=1}^N \partial_i f(x) v_i = \langle \nabla f(x), v \rangle,$$

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

Remark 6.10 allows us to define the notion of regularity of a 1-form by using the regularity of the coordinates functions.

Definition 6.12. We say that a 1-form $\omega \in \Lambda^1(\mathbb{R}^N)$ is continuous (or differentiable, or C^1), if its coordinates $\omega_1, \dots, \omega_N : \mathbb{R}^N \to \mathbb{R}$ are continuous (or differentiable, or C^1).

Finally, we define the subclass of 1-forms that are gradients.

Definition 6.13. A 1-form $\omega \in \Lambda^1(\mathbb{R}^N)$ is called *exact* if there exists $f \in C^1(\mathbb{R}^N)$ such that $\omega = \mathrm{d}f$.

6.3. **Poincarè Lemma.** By using the language of 1-forms, we are in position to prove two characterizations of gradients vector fields.

The idea of how to construct the primitive (or potential) in both cases is based on the Fundamental Theorem of Calculus. The difference between the two characterizations is in the conditions ensuring that the potential is well defined.

Assume that $f : \mathbb{R}^N \to \mathbb{R}$ is a C^1 function. Fix $x, y \in \mathbb{R}^N$, and consider the function $g : [0,1] \to \mathbb{R}$ defined as

$$g(t) \coloneqq f(y + t(x-y)).$$

Then,

$$\begin{aligned} f(x) - f(y) = g(1) - g(0) &= \int_0^1 g'(t)\,dt \\ &= \int_0^1 \langle \nabla f(y + t(x-y)), (x-y)\rangle\,dt \\ &= \int_0^1 \mathrm{d}f(y + t(x-y))[x-y]\,dt. \end{aligned}$$

Now, take a generic 1-form $\omega \in \Lambda^1(\mathbb{R}^N)$. By using the above formula with a fixed $y \in \mathbb{R}^N$, we can define

$$F(x) \coloneqq \int_0^1 \omega(x + t(y-x))[y-x]\,dt.$$

This will give a function $F : \mathbb{R}^N \to \mathbb{R}$ that, hopefully, satisfies $\mathrm{d}F = \omega$ (we will check this later). Before worrying about this, we first need to check that F is a well defined function, namely that its value *does not depend on the choice of the path from y to x*. In order to state such property, we need to introduce the notion of integration of a 1-form along a curve.

Definition 6.14. Let $\omega \in \Lambda^1(\mathbb{R}^N)$ be continuous and let $\gamma : [0,1] \to \mathbb{R}^N$ be a piecewise-C^1 curve. We define

$$\int_\gamma \omega \coloneqq \int_0^1 \omega(\gamma(t))[\gamma'(t)]\,dt.$$

Similarly, we define the integration over a piecewise C^1 curve.

Remark 6.15. In the case $\omega = \mathrm{d}f$, for a C^1 function $f : \mathbb{R}^n \to \mathbb{R}$, by using the Chain Rule (see Proposition 5.6) we get (see Figure 21)

$$\int_\gamma \mathrm{d}f = \int_0^1 \langle \nabla f(\gamma(t)), \gamma'(t)\rangle\,dt = \int_0^1 (f\circ\gamma)'(t)\,dt = f(\gamma(1)) - f(\gamma(0)),$$

for any piecewise C^1 curve $\gamma : [0,1] \to \mathbb{R}^N$.

We are now in position to prove the first characterization of gradient vector fields.

Theorem 6.16. *Let $\omega \in \Lambda^1(\mathbb{R}^N)$ be continuous, and let $\gamma, \mu : [0,1] \to \mathbb{R}^N$ be piecewise C^1 curves. The following are equivalent:*

(i) If $\gamma(0) = \gamma(1)$, then

$$\int_\gamma \omega = 0;$$

(ii) If $\gamma(0) = \mu(0)$, and $\gamma(1) = \mu(1)$, then

$$\int_\gamma \omega = \int_\mu \omega;$$

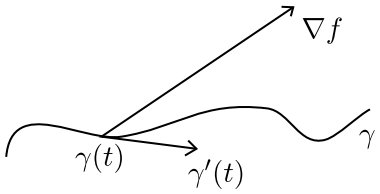

FIGURE 21. The geometric idea behind the definition of the integral of a gradient along a curve.

(iii) ω is exact; namely there exists $f \in C^1(\mathbb{R}^N)$ such that $\omega = \mathrm{d}f$.

Proof. Step 1: (i)⇒(ii). The idea is to follow γ forward, and then μ backwards. This will give a closed path that allows us to use (i). Define the curve $\widetilde{\mu} : [0,1] \to \mathbb{R}^N$ defined as

$$\widetilde{\mu}(t) \coloneqq \mu(1-t).$$

Namely, $\widetilde{\mu}$ is μ traveled backwards. Then, it holds

$$\int_{\widetilde{\mu}} \gamma = -\int_{\mu} \gamma. \tag{6.3}$$

Since $\widetilde{\mu}(0) = \gamma(1)$, and $\widetilde{\mu}(1) = \gamma(0)$, the curve $\lambda : [0,1] \to \mathbb{R}^N$ defined as

$$\lambda(t) \coloneqq \begin{cases} \gamma(2t) & \text{if } t \in [0,1/2], \\ \\ \widetilde{\mu}(2t-1) & \text{if } t \in [1/2,1]. \end{cases}$$

Then, λ satisfies $\lambda(0) = \lambda(1)$. Thus, from (i) and (6.3), we get the desired result.

Step 2: (ii)⇒(iii). Define the function $f : \mathbb{R}^N \to \mathbb{R}$ by

$$f(x) \coloneqq \int_0^1 \omega(tx)[x]\, dt.$$

Thanks to (ii), the function f is well defined. Indeed, its definition is independent on the path taken to from the origin to the point x. We now prove that $\mathrm{d}f = \omega$. For, we want to show that

$$\frac{\partial f}{\partial v}(x) = \omega(x)[v],$$

for all $x \in \mathbb{R}^N$ and all $v \in \mathbb{R}^N$. Let $h \neq 0$. By using (ii), we can compute the value $f(x+hv)$ by connecting the point $x+hv$ with the origin to the point x with a segment, and then connecting the point x to the origin with another segment. Thus, we get that

$$f(x+hv) = \int_0^1 \omega(tx)[x]\, dt + \int_0^1 \omega(x+shv)[hv]\, ds.$$

Thus,

$$\frac{f(x+hv)-f(x)}{h} = \int_0^1 \omega(x+shv)[v]\, ds = \frac{1}{h}\int_0^h \omega(x+rv)[v]\, ds, \tag{6.4}$$

where in the last step we used the change of variables $sh = r$. Now, fix $\varepsilon > 0$. By using the continuity of ω (namely the continuity of its components), there exists $\delta > 0$ such that if $r \in \mathbb{R}$ is such that $|rv| < \delta$, then

$$|\omega(x)[v] - \omega(x+rv)[v]| < \varepsilon. \tag{6.5}$$

Note that we are considering a *fixed* vector $v \in \mathbb{R}^N$. Thus, if

$$|h| < \frac{\delta}{|v|},$$

we get that

$$\begin{aligned}\left|\frac{1}{h}\int_0^h \omega(x+rv)[v]\,ds-\omega(x)[v]\right| &= \left|\frac{1}{h}\int_0^h \omega(x+rv)[v]\,ds-\frac{1}{h}\int_0^h \omega(x)[v]\,ds\right| \\ &\le \frac{1}{h}\int_0^h |\omega(x+rv)[v]-\omega(x)[v]|\,ds \\ &\le \varepsilon,\end{aligned}$$

where in the last step we used (6.5) to estimate the first integral. Thus, since $\varepsilon>0$ is arbitrary, we conclude that

$$\lim_{h\to 0}\frac{1}{h}\int_0^h \omega(x+rv)[v]\,ds=\omega(x)[v], \tag{6.6}$$

Thus, from (6.4) and (6.6) we conclude.

Step 3: (iii)⇒(i). This follows from the computations in Remark 6.15. □

Remark 6.17. The same result holds also if instead of having ω defined in the whole $\mathbb{R}^N$, it is defined only on an open set $\Omega\subset\mathbb{R}^N$.

We now have a characterization of 1-forms that are gradients. The problem is that the above conditions are not that easy to check. We would like to have a more manageable condition that characterizes exactness of a 1-form. For, we would like to use the algebraic condition given by Schwarz's Theorem (see Theorem 6.1).

Definition 6.18. We say that a C^1 1-form $\omega\in\Lambda^1(\mathbb{R}^N)$ is *closed*, if

$$\partial_i\omega_j=\partial_j\omega_i,$$

where $\omega=(\omega_1,\dots,\omega_N)$ with respect to the basis $(\mathrm{d}x_1,\dots\mathrm{d}x_N)$.

We now want to understand what is the issue in proving that a closed form $\omega\in\Lambda^1(\Omega)$ is exact in an *entire* open set Ω. To start with, let us notice that if two functions $f,g:\mathbb{R}^N\to\mathbb{R}$ are such that

$$\omega=\nabla f,\qquad \omega=\nabla g,$$

then, $f=g+c$, for some $c\in\mathbb{R}$. Assume that Ω is connected (namely, it is just *one piece*). Thus, since two potentials differs by a constant, the problem of passing from the local exactness of ω to its global exactness is a matter of *matching constants* of the local potentials. This is an issue that involves the *topology* of the set Ω. We will present the proof for a subclass of open sets for which the full characterization of exactness given by the Poincaré Lemma holds. This is for simplicity of exposition.

In order to present the result for forms defined on a subset of $\mathbb{R}^N$, to carry out all of the above constructions, we need to make sure that it is possible to connect each point in our set with a base point. This will restrict the type of domains that we will consider.

Definition 6.19. A set $\Omega\subset\mathbb{R}^N$ is called *star-shaped*, if there exists a point $x\in\Omega$ such that (see Figure 22)

$$tx+(1-t)y\in\Omega,$$

for all $y\in\Omega$ and $t\in[0,1]$.

We can now prove the characterization we wanted.

Theorem 6.20 (Poincaré Lemma)**.** *Let $\Omega\subset\mathbb{R}^N$ be a star-shaped domain, and let $\omega\in\Lambda^1(\Omega)$ be of class C^1. Then, ω is exact if and only if it is closed.*

Proof. Step 1. Assume ω is exact. Then, by Schwarz Theorem (see Theorem 6.1), it is closed.

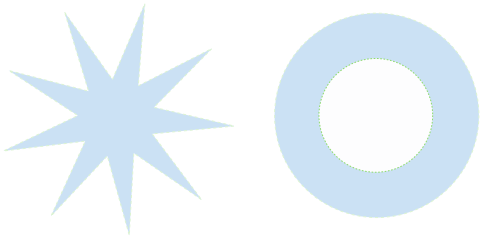

FIGURE 22. A star-shaped set on the left, and a non star-shaped set on the right

Step 2. Assume ω is closed. Define

$$f(x) := \int_0^1 \omega(tx)[x]\, dt.$$

We need to check that $\omega = \mathrm{d}f$. As we did in the proof of Theorem 6.16, this will be achieved by showing that

$$\partial_i f(x) = \omega_i(x),$$

for all $x \in \mathbb{R}^N$ and all $i = 1, \dots, N$. By using Leibniz formula (see Proposition 5.18), we have that

$$\begin{aligned}
\partial_i f(x) &= \partial_i \int_0^1 \sum_{j=1}^N \omega_j(tx) x_j \, dt \\
&= \int_0^1 \sum_{j=1}^N \partial_i [\omega_j(tx) x_j] \, dt \\
&= \int_0^1 \left[\omega_i(tx) + \sum_{j=1}^N t \partial_i \omega_j(tx) x_j \right] dt \\
&= \int_0^1 \left[\omega_i(tx) + \sum_{j=1}^N t \partial_j \omega_i(tx) x_j \right] dt \\
&= \int_0^1 \partial_t (t \omega_i(tx)) \, dt \\
&= \omega_i(x),
\end{aligned}$$

where in the fourth step we used the fact that $\partial_i \omega_j(tx) = \partial_j \omega_i(tx)$ for all $j = 1, \dots, N$, since ω is closed. This concludes the proof of the theorem. □

Remark 6.21. Poincaré Lemma holds for more general domains. Indeed, it is also valid for domains *without holes*. The precise notion that you will see in *Topology*. This gives yet another link between Analysis and Topology: the validity of Poincaré Lemma depends on topological properties of domain we consider.

As a corollary, we get that a closed 1-form is *locally* exact.

Corollary 6.22. *Let $\omega \in \Lambda^1(\Omega)$ be a closed C^1 1-form, where $\Omega \subset \mathbb{R}^N$ is an open set. Then, ω is locally exact. Namely, for each $x \in \Omega$, there exists a radius $r > 0$ and a function $f : \mathbb{R}^N \to \mathbb{R}$ such that $\omega = \nabla f$ in $B(x, r)$.*

6.4. **Helmholtz Decomposition Theorem.** Poincaré Lemma ensures that a closed 1-form in a star-shaped domain is the gradient of a function. Since the viceversa is also true, this is a characterization of C^1 vector fields that are gradients of a C^2 function. What about regular vector fields that are not closed? Can we say something about their structure?

The answer is in the computations we did before.

Proposition 6.23. *Let $\omega \in \Lambda^1(\mathbb{R}^N)$ be of class C^1. Then, there exists a function $f \in C^2(\mathbb{R}^N)$, and a vector field $V \in C^1(\mathbb{R}^N; \mathbb{R}^N)$ such that*

$$\omega = \nabla f + V.$$

Note that we are identifying a 1-form with its coordinates. In particular, if ω is closed, then $V = 0$.

Proof. Define

$$f(x) \coloneqq \int_0^1 \omega(tx)[x]\, dt.$$

Then, for $i = 1, \dots, N$, we get the definition of V_i by looking at the difference between $\partial_i f$ and ω_i.

$$\begin{aligned}
\partial_i f(x) &= \partial_i \int_0^1 \sum_{j=1}^N \omega_j(tx) x_j \, dt = \int_0^1 \sum_{j=1}^N \partial_i[\omega_j(tx)x_j]\, dt \\
&= \int_0^1 \left[\omega_i(tx) + \sum_{j=1}^N t\partial_i\omega_j(tx)x_j \right] dt \\
&= \int_0^1 \left[\omega_i(tx) + \sum_{j=1}^N t\partial_j\omega_i(tx)x_j \right] dt + \int_0^1 t \sum_{j=1}^N \left[\partial_i\omega_j(tx)x_j - \partial_j\omega_i(tx)x_j \right] dt \\
&= \int_0^1 \partial_t(t\omega_i(tx))\, dt + \int_0^1 t \sum_{j=1}^N \left[\partial_i\omega_j(tx)x_j - \partial_j\omega_i(tx)x_j \right] dt \\
&= \omega_i(x) + \int_0^1 t \sum_{j=1}^N \left[\partial_i\omega_j(tx)x_j - \partial_j\omega_i(tx)x_j \right] dt,
\end{aligned}$$

Thus, by defining

$$V_i \coloneqq \int_0^1 t \sum_{j=1}^N \left[\partial_j\omega_i(tx)x_j - \partial_i\omega_j(tx)x_j \right] dt,$$

we get the desired result. □

Remark 6.24. In a way, we see that the vector field V measures *how much ω is not exact.* There is a theoretical way to write the above formula, by using the notion of 2-forms, but we will not do it in here.

What is surprising, is that, for dimension $N = 3$ (and only for this dimension), the above vector field V can be written as the curl of another vector field.

Definition 6.25. Let $V : \mathbb{R}^3 \to \mathbb{R}^3$ be a C^1 vector field. We define its *curl*, $\operatorname{curl}(V) : \mathbb{R}^3 \to \mathbb{R}^3$, as

$$\operatorname{curl}(V) \coloneqq (\partial_2 V_3 - \partial_3 V_2,\, \partial_3 V_1 - \partial_1 V_3,\, \partial_1 V_2 - \partial_2 V_1).$$

An alternative notion for the curl of V is $\nabla \times V$.

Remark 6.26. Two important vector identities are the following:

(i) Let $f \in C^2(\mathbb{R}^N)$. Then,

$$\operatorname{curl}(\nabla f) = 0;$$

(ii) Let $V : \mathbb{R}^3 \to \mathbb{R}^3$ be a C^2 vector field. Then,

$$\operatorname{div}(\operatorname{curl}(V)) = 0.$$

What is the geometrical meaning of these identities?

The result by Helmholtz requires some regularity of the bounded domain we are in, or requires some integrability conditions of the coordinates of ω is we consider the entire space $\mathbb{R}^3$. we will state the result by using directly a vector field $F : \mathbb{R}^3 \to \mathbb{R}^3$, that we identify with a 1-form $\omega \in \Lambda^1(\mathbb{R}^N 0$.

Theorem 6.27 (Helmholtz decomposition Theorem)**.** *Let $\Omega \subset \mathbb{R}^3$ be an open set, and let $F : \Omega \to \mathbb{R}^3$ be a C^2 vector field. If Ω is bounded, we assume its boundary ∂S to be a regular surface. If Ω is unbounded (in particular, if $\Omega = \mathbb{R}^3$), we assume that there exists $R > 0$ and $C > 0$ such that*

$$|F(x)| \leq \frac{C}{|x|},$$

for all $|x| > R$. Then, there exist $\varphi \in C^3(\Omega)$, and $W \in C^3(\mathbb{R}^3; \mathbb{R}^3)$ such that $F = \nabla\varphi + \operatorname{curl} W$.

Remark 6.28. Helmholtz Decomposition Theorem has plenty of applications in physics, oceanology, geophysics, weather modeling, and computer graphics, since it allows to understand properties of a vector field like its vorticity (its curl) and incompressibility (its divergence).

Remark 6.29. With similar ideas, it is possible to study the topology of a set, by studying analytical properties of differential operators on the set. This brought to the development of the *Hodge Decomposition Theorem*, that can be seen as a generalization of the Helmholtz Decomposition Theorem (see Theorem 6.27).

7. Inverse Function Theorem

What type of information about the local behavior of a function can we get from properties of the differential at a point? This is the main question of this chapter. The basic idea can be understood in the one dimensional scalar case. Consider a C^1 map $f : \mathbb{R} \to \mathbb{R}$, and let $\bar{x} \in \mathbb{R}$ be a point such that $f'(\bar{x}) \neq 0$. By looking at Figure 23, it is geometrically clear that f is locally invertible at $\bar{x}$. Namely, that there exists $\delta > 0$ such that $f : (\bar{x} - \delta, \bar{x} + \delta) \to \mathbb{R}$ admits an inverse.

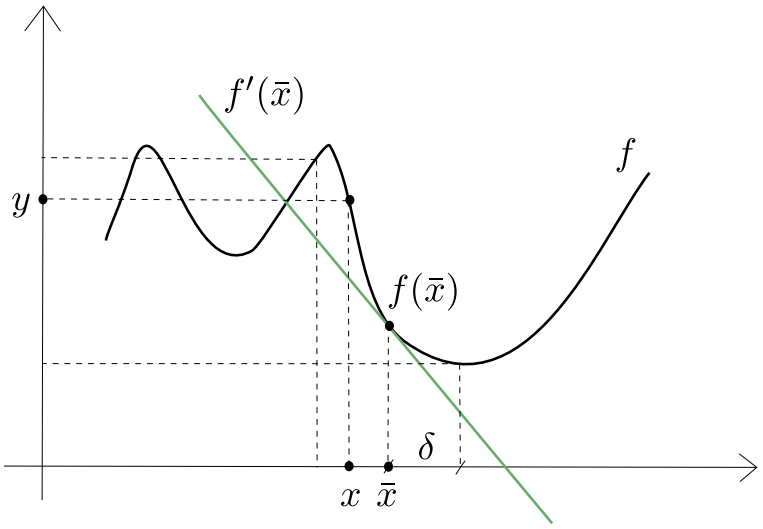

Figure 23. The geometric idea behind the inversion function theorem in the one dimensional scalar case.

Since

$$f(x) = f(\bar{x}) + f'(\bar{x})(x - \bar{x}) + o(|x - \bar{x}|),$$

we can write

$$x = \bar{x} + \frac{f(x) - f(\bar{x}) - o(|x - \bar{x}|)}{f'(\bar{x})} = \bar{x} + (f'(\bar{x}))^{-1}[f(x) - f(\bar{x}) - o(|x - \bar{x}|)].$$

Note that this expression is not explicit, since the *error* $o(|x - \bar{x}|)$ is not explicit, in general. Nevertheless, in many cases, it is enough to know that such inverse exists.

It is less clear, at least geometrically, that also the inverse is of class C^1. This will require more delicate investigations that will be carried out in this chapter, and that will also allow to treat the general dimension case.

In particular, we will investigate what can be said when the differential is injective at a point, when it is surjective, and how to combine the information we get from these two conditions to obtain a global characterization of a class of maps called diffeomorphisms, as well as a local one. This latter bears the name of *Inverse Function Theorem*. A diffeomorphisms is a C^1 bijection with C^1 inverse. In particular, it is a map that preserves the *differential structure of sets*.

In order to undertake such investigation, we first need to better understand the role of the differential as a map that sends tangent vectors to tangent vectors. This will be done in the first section.

7.1. The differential as a tangent application.

We want to study the behavior of the differential on tangent cones. First, we prove that a tangent cone to a set is sent by the differential inside the tangent cone of the image of the set.

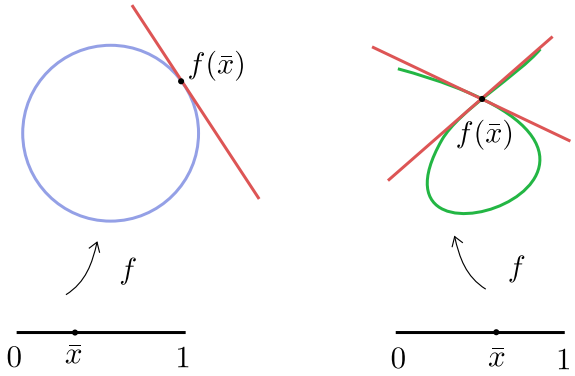

FIGURE 24. Two examples where the inclusion (7.1) is strict. In tangent cone of the image at $f(\bar{x})$ is depicted in red.

Proposition 7.1. *Let $\Omega \subset \mathbb{R}^N$ be an open set, and let $f : \Omega \to \mathbb{R}^M$ be differentiable at a point $\bar{x} \in \Omega$. Then,*

$$\mathrm{d}f(\bar{x})\left[\mathrm{Tan}(S, \bar{x})\right] \subset \mathrm{Tan}(f(S), f(\bar{x})), \tag{7.1}$$

for any set $S \subset \Omega$.

Proof. Let $v \in \mathrm{Tan}(S, \bar{x})$. Then, by definition of tangent cone, there exist $(x_n)_{n\in\mathbb{N}} \subset S$ and $(\lambda_n)_{n\in\mathbb{N}} \subset (0, 1)$ with

$$x_n \to \bar{x}, \qquad \lambda_n \to 0, \qquad \frac{x_n - \bar{x}}{\lambda_n} \to v,$$

as $n \to \infty$. Since f is continuous at $\bar{x}$, we have that $\lim_{n\to\infty} f(x_n) = f(\bar{x})$. Therefore, thanks to Theorem 5.38, we get that

$$\frac{f(x_n) - f(\bar{x})}{\lambda_n} \to \mathrm{d}f(\bar{x})[v].$$

This concludes the proof. □

Remark 7.2. The above inclusion might be strict (see Figure 24). For example, consider the map $f : [0, 2\pi] \to \mathbb{R}^2$ given by

$$f(t) := (\cos t, \sin t).$$

Then, the tangent cone of S^1 at $(1, 0)$ is the vertical line $\{x = 1\}$. Since

$$\mathrm{Tan}([0, 2\pi], 0) = \{t \geq 0\},$$

we get that

$$\mathrm{d}f(\bar{x})\left[\mathrm{Tan}(S, \bar{x})\right] = \{x = 0, y \geq 0\} \subset \mathrm{Tan}(f(S), f(\bar{x})),$$

with proper inclusion. What makes the inequality strict is that half of the tangent cone of $f(S)$ comes from the image of the tangent cone at $t = 0$, while the other half comes from the image of the tangent cone at $t = 2\pi$. This is because the same point in the image can be reached from two different points in the domain.

We therefore introduce the class of functions such that this pathology is avoided.

Definition 7.3. A map $f : \mathbb{R}^N \to \mathbb{R}^M$ is said to be a *homeomorphism* on a set $S \subset \mathbb{R}^N$, if it is continuous, and with a continuous inverse. Namely, if

$$x_n \to \bar{x} \quad \Leftrightarrow \quad f(x_n) \to f(\bar{x}),$$

for all $(x_n)_{n\in\mathbb{N}} \subset S$, and $\bar{x} \in S$.

Remark 7.4. Note that being a homeomorphism is stronger than having an inverse, since we are requiring this latter to be continuous. Can you find an example of a continuous function with a non continuous inverse?

Example 7.5. Consider the map $f : \mathbb{R}^2 \to \mathbb{R}^2$ defined as

$$f(x) \coloneqq \frac{|x|}{\|x\|_\infty} x$$

where $|x|$ denotes the Euclidean norm of x, and $\|x\|_\infty$ is the ∞-Minkowski norm of x (see Example 1.12). Then, f is a homeomorphism (check it!). In particular,

$$f(B(0,1)) = (-1,1)^2.$$

Namely, f is a homeomorphism between the unit circle and the unit square with sides parallel to the axes.

Remark 7.6. Another example when the inequality in (7.1) is strict is the following. Consider the map $f : \mathbb{R} \to \mathbb{R}$ defined as

$$f(t) \coloneqq t^3.$$

Then, $f'(0) = 0$, but $\mathrm{Tan}(\mathrm{f}(\mathbb{R}), \mathrm{f}(0)) = \mathbb{R}$. What makes the inequality strict in this case is that the differential of f at $\bar{x}$ is not injective.

Remark 7.7. Note that none of the two above conditions, namely being a homeomorphism or having injective differential, is sufficient by itself to ensure equality in (7.1).

7.2. **When the differential is injective.** We now want to understand what properties a function enjoys when the differential at a point is injective.

We start by showing that if the differential is injective, then the function is locally injective. We will prove something more, namely we will get an explicit modulus of continuity of the inverse.

Proposition 7.8. *Let $\Omega \subset \mathbb{R}^N$ be an open set, and let $f : \Omega \to \mathbb{R}^M$ be differentiable at a point $\bar{x} \in \Omega$. Assume that the differential $\mathrm{d}f(\bar{x})$ is injective on a linear subspace $V \subset \mathbb{R}^N$. Then, the map f restricted to $\bar{x} + V$ is locally injective. Namely, there exists $\delta > 0$ such that*

$$\delta \|x - \bar{x}\| \leq \|f(x) - f(\bar{x})\|, \tag{7.2}$$

for all $x \in B(\bar{x}, \delta) \cap (\bar{x} + V)$. Moreover, if f is C^1 in a neighborhood of $\bar{x}$, it holds

$$\delta \|x - y\| \leq \|f(x) - f(y)\|, \tag{7.3}$$

for all $x, y \in B(\bar{x}, \delta) \cap (\bar{x} + V)$.

Proof. We prove the first claim. The second follows by using a similar argument. The idea of the proof is geometrically clear: if there was a sequence of points $(x_n)_{n\in\mathbb{N}}$ approaching $\bar{x}$ with $f(x_n) = f(\bar{x})$, then the directional derivative of f along the direction identified by any limit of

$$\frac{x_n - \bar{x}}{\|x_n - \bar{x}\|}$$

will be zero. Since that direction is not zero, this is in contradiction with the injectivity of $\mathrm{d}f(\bar{x})$. Let's write mathematically this idea.

Assume by contradiction that the result is not true. Then, there would exist a sequence $(x_n)_{n\in\mathbb{N}} \subset \mathbb{R}^N$ with

$$x_n \in B\left(\bar{x}, \frac{1}{n}\right) \cap (\bar{x} + V)$$

such that

$$\frac{\|f(x_n) - f(\bar{x})\|}{\|x_n - \bar{x}\|} < \frac{1}{n}. \tag{7.4}$$

We now continue the argument with a typical *shortcut* that is commonly used in modern mathematics: instead of writing explicitly the subsequence that will give us the desired contradiction, we will just write as follows. *Up to a subsequence, we can assume that*

$$\frac{x_n - \bar{x}}{\|x_n - \bar{x}\|} \to v, \tag{7.5}$$

for some $v \in \mathbb{S}^{N-1} := \partial B(0,1)$. Note that, since the argument proceeds by contradiction, there is no need to specify the subsequence that we use, nor to make sure that the conclusion is independent of the chosen subsequence. This is why we can use such shortcut and have a lighter notation.

Let us continue with the proof. By definition of differentiability (or by Theorem 5.23), together with (7.4), and (7.5), we would get that

$$0 = \lim_{n\to\infty} \frac{\|f(x_n) - f(\bar{x})\|}{\|x_n - \bar{x}\|} = \lim_{n\to\infty} \mathrm{d}f(\bar{x}) \left[\frac{x_n - \bar{x}}{\|x_n - \bar{x}\|} \right] = \mathrm{d}f(\bar{x})[v].$$

But $\mathrm{d}f(\bar{x})[v] \neq 0$, since $v \neq 0$, and $\mathrm{d}f(\bar{x})$ is injective. This gives the desired contradiction. □

Remark 7.9. Is the opposite true? Namely, if the function is injective, can we conclude that the differential is injective?

Remark 7.10. The above proposition provides a *quantitative* version of the statement: the injectivity of the differential at a point implies that the function is locally injective around that point. In particular, note that (7.2) writes as

$$\|f^{-1}(p) - f^{-1}(\bar{x})\| \leq \frac{1}{\delta}\|p - \bar{x}\|,$$

for all $p \in f(B(\bar{x},\delta) \cap V)$. This proves that the inverse function is continuous, and with an explicit form of the modulus of continuity. Moreover, (7.3) writes as

$$\|f^{-1}(p) - f^{-1}(q)\| \leq \frac{1}{\delta}\|p - q\|,$$

for all $p \in f(B(\bar{x},\delta) \cap V)$. This implies that (see Remark 4.14)

$$[f^{-1}]_{\mathrm{Lip}} \leq \frac{1}{\delta}.$$

Note that, from this, we cannot conclude that, in general, f^{-1} is differentiable at $f^{-1}(\bar{x})$. We will see that, thanks to Theorem 7.17, this is though the case when $N = M$, and $V = \mathbb{R}^N$.

We now continue the investigation initiated in the previous section, by showing that the above ones are the only cases where things can go wrong for having the equality case in Proposition 7.1. Namely, that for a homeomorphism with injective differential at a point $\bar{x}$, the differential maps the tangent cone of any set S at $\bar{x}$ onto the tangent cone of $f(S)$ at $f(\bar{x})$, and that any element of this latter is the image of a tangent vector to S at $\bar{x}$.

Proposition 7.11. *Let $\Omega \subset \mathbb{R}^N$ be an open set, and let $f : \Omega \to \mathbb{R}^M$ be differentiable at a point $\bar{x} \in \Omega$, with $\mathrm{d}f(\bar{x})$ injective. Moreover, assume that f is a homeomorphism at $S \subset \Omega$. Then*

$$\mathrm{d}f(\bar{x})\left[\mathrm{Tan}(S,\bar{x})\right] = \mathrm{Tan}(f(S), f(\bar{x})),$$

Proof. Thanks to Proposition 7.1, we just need to prove the inclusion $\supset$. For, let $(x_n)_{n\in\mathbb{N}} \subset S$, $(\lambda_n)_{n\in\mathbb{N}} \subset (0,1)$, and $w \in \mathbb{R}^M$ be such that

$$f(x_n) \to f(\bar{x}), \qquad \lambda_n \to 0, \qquad \frac{f(x_n) - f(\bar{x})}{\lambda_n} \to w,$$

as $n \to \infty$. We claim that there exists $v \in \mathbb{R}^N$ such that

$$\frac{x_n - \bar{x}}{\lambda_n} \to v, \qquad \mathrm{d}f(\bar{x})[v] = w.$$

Note that the difficulty here lies in the fact that, by setting

$$v_n := \frac{x_n - \bar{x}}{\lambda_n},$$

it is not clear a-priori that the vectors v_n have uniformly bounded norm. Indeed, if that was the case, we could just conclude by extracting a converging subsequence and using the linearity of the differential to prove that each subsequence converges to the same vector v.

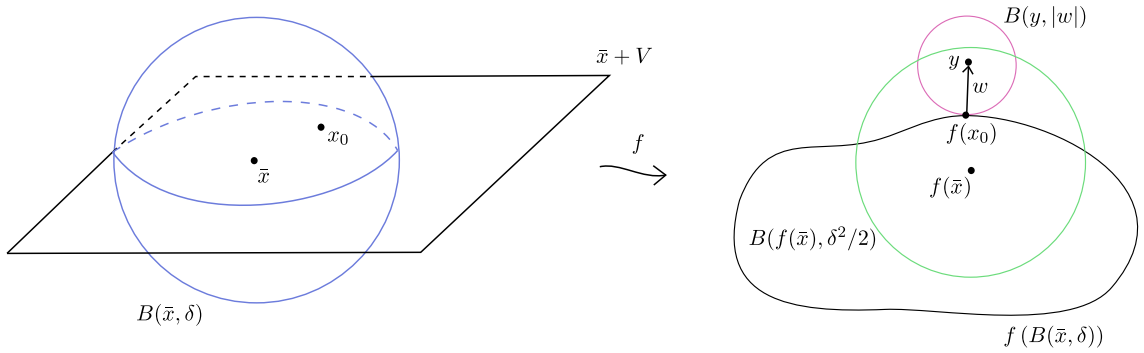

FIGURE 25. The geometric idea behind the proof of Proposition 7.12

.

First of all, we note that, if $x_n = \bar{x}$ for infinitely many indexes n's, then it follows that $w = 0$, and that $v = 0$ (Fill in the details yourself!).

Therefore, we can assume that there exists $\bar{n} \in \mathbb{N}$ such that $x_n \neq \bar{x}$ for all $n \geq \bar{n}$. Write

$$\mathrm{d}f(\bar{x})\left[\frac{x_n - \bar{x}}{\lambda_n}\right] = \frac{f(x_n) - f(\bar{x})}{\lambda_n} - \frac{f(x_n) - f(\bar{x}) - \mathrm{d}f(\bar{x})[x_n - \bar{x}]}{\|x_n - \bar{x}\|} \frac{\|x_n - \bar{x}\|}{\|f(x_n) - f(\bar{x})\|} \frac{\|f(x_n) - f(\bar{x})\|}{\lambda_n}. \quad (7.6)$$

Now, the first term converges to w by assumption. The second converges to zero by differentiability, and the latter is bounded by assumption. In order to estimate the previous to last term, we use Proposition 7.8 to get a $\delta > 0$ such that

$$\delta\|x_n - \bar{x}\| \leq \|f(x_n) - f(\bar{x})\|,$$

for all $x_n \in B(\bar{x}, \delta)$. Thus, from (7.6) we get that

$$\lim_{n\to\infty} \mathrm{d}f(\bar{x})\left[\frac{x_n - \bar{x}}{\lambda_n}\right] = w.$$

Thus, w is the limit of a sequence of images of vectors via the map $\mathrm{d}f(\bar{x})$. Since the image of the linear map $\mathrm{d}f(\bar{x})$ is a closed set, there exists $v \in \mathbb{R}^N$ such that

$$\mathrm{d}f(\bar{x})[v] = w.$$

The continuity and the injectivity of $\mathrm{d}f(\bar{x})$ then imply that

$$\frac{x_n - \bar{x}}{\lambda_n} \to v,$$

as desired. □

7.3. **When the differential is surjective.** When the differential of a map $f : \mathbb{R}^N \to \mathbb{R}^N$ is surjective at a point $\bar{x}$, it means that, locally, the map *sends points in all directions*. In particular, this means that the image is open. We can prove this statement in a quantitative way, and also for maps $f : \mathbb{R}^N \to \mathbb{R}^M$, with a general dimension M.

Proposition 7.12. *Let $\Omega \subset \mathbb{R}^N$ be an open set, and let $f : \Omega \to \mathbb{R}^M$ be differentiable at a point $\bar{x} \in \Omega$, with $\mathrm{d}f(\bar{x})$ surjective. Then, the image of f is locally open. Namely, there exists $\delta > 0$ such that*

$$B(f(\bar{x}), \delta^2/2) \subset f(B(\bar{x}, \delta)).$$

Proof. Since $\mathrm{d}f(\bar{x})$ is surjective, we get that $M \leq N$. Moreover, there exists a linear subspace $V \subset \mathbb{R}^N$ of dimension M where $\mathrm{d}f(\bar{x})$ is injective. Thus, from Proposition 7.8, we get the existence of a $\delta > 0$ such that

$$\delta \|x - \bar{x}\| \leq \|f(x) - f(\bar{x})\|, \tag{7.7}$$

for all $x \in \overline{B(\bar{x}, \delta)} \cap (\bar{x} + V)$. Note that we consider the closure of the ball, since the map f is continuous. We claim that this $\delta > 0$ does the job. Let $y \in B(f(\bar{x}), \delta^2/2)$. We will show that there exists $x_0 \in B(\bar{x}, \delta)$ such that $f(x_0) = y$. Note that we are not claiming any uniqueness of such a point. Let $x_0 \in \mathbb{R}^N$ be such that

$$\|f(x_0) - y\| = \min\left\{ \|f(x) - y\| : x \in \overline{B(\bar{x}, \delta)} \cap (\bar{x} + V) \right\}.$$

Note that such a point exists thanks to Weierstraß Theorem (see Theorem 3.15). Indeed, the function $x \mapsto \|f(x) - y\|$ is continuous, and the set $\overline{B(\bar{x}, \delta)} \cap (\bar{x} + V)$ is compact, since it is closed and bounded (see Bolzano-Weierstraß Theorem 2.23).

First of all, we show that $x_0 \in B(\bar{x}, \delta) \cap (\bar{x} + V)$. Indeed, from (7.7), we have that

$$\begin{aligned} \delta \|x_0 - \bar{x}\| &\leq \|f(\bar{x}) - f(x_0)\| \leq \|f(\bar{x}) - y\| + \|y - f(x_0)\| \\ &\leq 2\|f(\bar{x}) - y\| < \delta^2, \end{aligned} \tag{7.8}$$

where the previous to last step follows from the choice of x_0, while last step from the fact that $y \in B(f(\bar{x}), \delta^2/2)$.

We claim that $f(x_0) = y$. Assume not. Then, $w \coloneqq y - f(x_0) \neq 0$. Thus, by the definition of x_0, we have that

$$B(y, |w|) \cap f\left(B(\bar{x}, \delta) \cap (\bar{x} + V)\right) = \emptyset.$$

In particular, this implies that (see Figure 25)

$$w \notin \mathrm{Tan}\left(f\left(B(\bar{x}, \delta) \cap (\bar{x} + V)\right), f(\bar{x})\right). \tag{7.9}$$

On the other hand, since (7.7) implies that f is a homeomorphism on $\overline{B(\bar{x}, \delta)} \cap (\bar{x} + V)$, we can apply Proposition 7.11 to get that

$$\mathrm{Tan}\left(f\left(B(\bar{x}, \delta) \cap (\bar{x} + V)\right), f(\bar{x})\right) = \mathrm{d}f(\bar{x})\left[\mathrm{Tan}(B(\bar{x}, \delta) \cap (\bar{x} + V), \bar{x})\right].$$

Since $\bar{x}$ is an internal point to $B(\bar{x}, \delta)$, we have that

$$\mathrm{Tan}(B(\bar{x}, \delta) \cap (\bar{x} + V)) = V,$$

which has dimension M. Thus, by using the fact that $\mathrm{d}f(\bar{x})$ is surjective, we get that

$$\mathrm{Tan}\left(f\left(B(\bar{x}, \delta) \cap (\bar{x} + V)\right), f(\bar{x})\right) = \mathbb{R}^M.$$

This is in contradiction with (7.9). □

7.4. **Diffeomorphisms.** We now want to study a class of regular transformation of an object into another. Since there is no differentiability requirement, an homeomorphism does not have to maintain the *differential structure of a set*. Namely, if $f : \mathbb{R}^N \to \mathbb{R}^N$ is a homeomorphism, and we consider a set $E \subset \mathbb{R}^N$, we have no control on the relation between the tangent cone to E at a point $x \in \mathbb{R}^N$, and the tangent cone to $f(E)$ at $f(x)$. In particular, homeomorphisms can *regularize* a set, or (equivalently), add *singularities* to it. For instance, the homeomorphism considered in Example 7.5 transforms a square into a circle. But the former has corner, while the latter does not!

In order to get information about the tangent cone to a set that are properly mapped by the differential to the tangent cone to the image of the set, we introduce a subclass of homeomorphisms. The idea is to require that both f and all of its first order derivatives to be homeomorphisms.

Definition 7.13. Let $\Omega \subset \mathbb{R}^N$ be an open set. A function $f : \Omega \to \mathbb{R}^N$ is said to be a *diffeomorphism* on Ω, if it is a homeomorphism of class C^1 from Ω to $f(\Omega)$, and also $f^{-1} : f(\Omega) \to \mathbb{R}^N$ is of class C^1.

Remark 7.14. Note that, for diffeomorphisms, the domain and the target have the same dimension N. Moreover, note that we talk about diffeomorphisms on *open* sets, since we need to have the differential defined.

We first investigate properties of diffeomorphisms.

Proposition 7.15. *Let $\Omega \subset \mathbb{R}^N$ be an open set, and let $f : \Omega \to \mathbb{R}^N$ be a diffeomorphism. Then:*

(i) f is a homeomorphism;
(ii) For all $x \in \Omega$
$$\det(Jf(x)) \neq 0;$$
(iii) For all $x \in \Omega$
$$J(f^{-1})(f(x)) = [Jf(x)]^{-1};$$
(iv) For all $x \in \Omega$ and $S \subset \Omega$, it holds that
$$\mathrm{Tan}(f(S), f(x)) = \mathrm{d}f(x)[\mathrm{Tan}(S, x)].$$
In particular, if $\mathrm{Tan}(S, x)$ is a vector space, also $\mathrm{Tan}(f(S), f(x))$ is.

Proof. Proof of (i). It follows directly from the definition of diffeomorphism.

Proof of (ii) and (iii). Fix $x \in \Omega$. By using the identity $f^{-1}(f(x)) = x$, and the fact that both f and f^{-1} are differentiable at x and $f(x)$ respectively, we get that
$$J(f^{-1})(f(x)) = [Jf(x)]^{-1},$$
and
$$\det(Jf(x)) \det(Jf^{-1}(f(x))) = 1,$$
proving what we wanted.

Proof of (iv). From (ii) we get that $\mathrm{d}f(x)$ is injective for all $x \in \Omega$. Thus, from Proposition 7.11 we get the desired conclusion. □

We are now in position to prove an important characterization of diffeomorphisms.

Theorem 7.16. *Let $\Omega \subset \mathbb{R}^N$ be an open set, and let $f : \Omega \to \mathbb{R}^N$. Then, f is a diffeomorphism in Ω if and only if*

(i) f is of class C^1 in Ω;
(ii) f is injective in Ω;
(iii) For all $x \in \Omega$, $\det(Jf(x)) \neq 0$

Proof. Thanks to Proposition 7.15 we just need to prove that (i), (ii), and (iii) are sufficient in order for f to be a diffeomorphism.

Step 1. We claim that $f(\Omega)$ is open. Indeed, since $\det(Jf(x)) \neq 0$, the differential $\mathrm{d}f(x)$ is surjective at each point $x \in \Omega$. Thus, the conclusion follows from Proposition 7.12.

Step 2. We claim that f is a homeomorphism. Indeed, f is continuous because it is differentiable at each point. Moreover, f^{-1} is well defined in $f(\Omega)$, because by assumption f is injective in Ω. To prove that f^{-1} is continuous, we reason as follows. Since $\det(Jf(x)) \neq 0$, the differential $\mathrm{d}f(x)$ is injective at each point $x \in \Omega$. Fix $x \in \Omega$. From Proposition 7.8 we get that there exists $\delta > 0$ such that
$$\delta\|x - y\| \leq \|f(x) - f(y)\|,$$
for all $y \in \overline{B(x, \delta)}$. This writes as
$$\delta\|f^{-1}(p) - f^{-1}(q)\| \leq \|p - q\|,$$

for all $p, q \in f(\overline{B(x,\delta)})$. This proves the continuity of f^{-1}.

Step 3. We claim that f^{-1} is differentiable in $f(\Omega)$. Indeed, note that, by step 1, f is a homeomorphism, and by assumption the differential $\mathrm{d}f(x)$ is injective at all points $x \in \Omega$. We want to use the characterization of differentiability provided by Theorem 5.38 to conclude that f^{-1} is differentiable. For, fix $\bar{y} \in \mathbb{R}^N$. We want to prove that there exists an $N \times N$ matrix A with the following property: for any $(y_n)_{n\in\mathbb{N}} \subset f(\Omega 0$ with $y_n \to \bar{y}$, and any $(\lambda_n)_{n\in\mathbb{N}} \subset (0,1)$ with $\lambda_n \to 0$ such that

$$\lim_{n\to\infty} \frac{y_n - \bar{y}}{\lambda_n} = w. \tag{7.10}$$

it holds

$$\lim_{n\to\infty} \frac{f^{-1}(y_n) - f^{-1}(\bar{y})}{\lambda_n} = A[w]. \tag{7.11}$$

Then, by Theorem 5.38 we get that f^{-1} is differentiable, and $\mathrm{d}(f^{-1})(\bar{y}) = A$.

Since f is a homeomorphism, for each $n \in \mathbb{N}$ there exists $x_n \in \mathbb{R}^N$ such that $y_n = f(x_n)$, and a point $\bar{x} \in \mathbb{R}^N$ such that $f(\bar{x}) = \bar{y}$. Moreover $x_n \to \bar{x}$. Therefore, (7.10) writes as

$$\lim_{n\to\infty} \frac{f(x_n) - f(\bar{x})}{\lambda_n} = w, \tag{7.12}$$

while (7.11) writes as

$$\lim_{n\to\infty} \frac{x_n - \bar{x}}{\lambda_n} = A[w]. \tag{7.13}$$

Since $f(\bar{x})$ is injective, we can use Proposition 7.8 to get $\delta > 0$ such that

$$\delta\|x - \bar{x}\| \leq \|f(x) - f(\bar{x})\|,$$

for all $x \in B(\bar{x}, \delta)$. Since $x_n \to \bar{x}$, let $\bar{n} \in \mathbb{N}$ be such that $x_n \in B(\bar{x}, \delta)$ for all $n \geq \bar{n}$. Thus

$$\delta\|x_n - \bar{x}\| \leq \|f(x_n) - f(\bar{x})\|, \tag{7.14}$$

for all $n \geq \bar{n}$. Note that, without loss of generality, we can assume $y_n \neq \bar{y}$ for all $n \in \mathbb{N}$. Indeed, if $y_n = \bar{y}$ for infinitely many indexes n's, then we would get $w = 0$, and thus there is nothing to prove, since this would mean, by injectivity of f that $x_n = \bar{x}$ for all $n \in \mathbb{N}$. Thus, (7.13) would hold for any matrix A. So, assume $y_n \neq \bar{y}$ for all $n \in \mathbb{N}$. Write

$$\mathrm{d}f(\bar{x})\left[\frac{x_n - \bar{x}}{\lambda_n}\right] = \frac{f(x_n) - f(\bar{x})}{\lambda_n} - \frac{f(x_n) - f(\bar{x}) - \mathrm{d}f(\bar{x})[x_n - \bar{x}]}{\|x_n - \bar{x}\|} \frac{\|x_n - \bar{x}\|}{\|f(x_n) - f(\bar{x})\|} \frac{\|f(x_n) - f(\bar{x})\|}{\lambda_n}.$$

Note that, thanks to (7.12) and (7.14), from the above writing we get that

$$\lim_{n\to\infty} \mathrm{d}f(\bar{x})\left[\frac{x_n - \bar{x}}{\lambda_n}\right] = w.$$

Since the image of a linear transformation is a (sequentially) closed space (see Definition 2.21), we get that there exists a vector $v \in \mathbb{R}^N$ such that

$$\mathrm{d}f(\bar{x})[v] = w. \tag{7.15}$$

By using (iii) we get that $Jf(\bar{x})$ is an invertible matrix, and define

$$A := [\, Jf(\bar{x}) \,]^{-1}.$$

Therefore, by applying A to both sides of (7.15),we get

$$v = A[w].$$

Thus, the matrix A satisfies (7.13), which proves that f^{-1} is differentiable. Moreover, we proved that

$$Jf^{-1}(f(\bar{x})) = [\, Jf(\bar{x}) \,]^{-1}.$$

Step 4. Finally, we claim that f^{-1} is of class C^1. Indeed, from the equality

$$J(f^{-1})(f(x)) = [Jf(x)]^{-1},$$

we just need to prove that the map $x \mapsto [Jf(x)]^{-1}$ is continuous. Note that

$$[Jf(x)]^{-1} = \frac{[\mathrm{cof}(Jf(x))]^T}{\det(Jf(x))},$$

where $\mathrm{cof}(Jf(x))$ is the cofactor matrix of $Jf(x)$. Since this latter is a polynomial in the $\partial_i f_j$'s, and these latter are continuous because by assumption f is C^1, we conclude that also $x \mapsto J(f^{-1})(f(x))$ is continuous. □

As a corollary, we get a *local* characterization of diffeomorphisms, known as the Inverse Function Theorem. This is an important result because it allows to get *local* information on the behavior of a function by having a *pointwise* information on its differential. In particular, it is a local version of Theorem 7.16, that does not require to check whether or not the function is injective. Note that a drawback of all of the theorems presented is that the neighborhood in which everything works (namely, the radius of the ball where we get the inverse of the function) is not explicit! This is usually not a problem for many applications.

Theorem 7.17 (Inverse Function Theorem)**.** *Let $\Omega \subset \mathbb{R}^N$ be an open set, and let $f : \Omega \to \mathbb{R}^N$ be a function of class C^1. Assume that $\bar{x} \in \Omega$ is such that*

$$\det(Jf(\bar{x})) \neq 0.$$

Then, there exists $r > 0$ such that f restricted to $B(\bar{x}, r)$ is a diffeomorphism.

Proof. Since the function f is of class C^1, the function

$$x \mapsto \det(Jf(x))$$

is continuous. Thus, there exists $r_1 > 0$ such that

$$\det(Jf(x)) \neq 0$$

for all $x \in B(\bar{x}, r_1)$. Moreover, since $\mathrm{d}f(\bar{x})$ is injective, from we get that there exists $\delta > 0$ and $r_2 > 0$ such that

$$\delta \|x - \bar{x}\| \leq \|f(x) - f(\bar{x})\|,$$

for all $x \in B(\bar{x}, r_2)$. In particular, this implies that f is injective in $B(\bar{x}, r_2)$. Set $r := \min\{r_1, r_2\}$. Then, f is a C^1 map that is injective in $B(\bar{x}, r)$ and with $\det(Jf(x)) \neq 0$ for all $x \in B(\bar{x}, r)$. the result then follows from Theorem 7.16. □

Remark 7.18. Note that the assumption of f being C^1 is necessary for Theorem 7.17 to hold. Indeed, the function $f : \mathbb{R} \to \mathbb{R}$ defined as

$$f(x) := \begin{cases} \dfrac{x}{2} + x^2 \sin \dfrac{1}{x} & \text{if } x \neq 0, \\ 0 & \text{else,} \end{cases}$$

is such that $f \in C^1(\mathbb{R} \setminus \{0\})$, but $f \notin C^1(\mathbb{R})$, and it is *not* locally injective around $x = 0$.

Example 7.19. Theorem 7.17 is very useful when using *change of coordinates*. Indeed, when you have to compute an integral, or rewrite a differential equation by using different coordinates, you need to make sure that the map that you use to go from one set of coordinates to the other is a diffeomorphism.

A change of coordinates that is widely used when the equation or integral that you are studying has a rotational symmetry is that of *spherical coordinates* (also called *polar coordinates* in dimension $N = 2$). They write in a complicated way for higher dimension. Thus, for the goal

of illustrating an example, we will only consider the case $N = 2$. In this case, we consider the map $f : \mathbb{R}^2 \to \mathbb{R}^2$ defined as

$$f(r, \theta) \coloneqq (r \cos\theta, r \sin\theta).$$

It is possible to see that f is a diffeomorphism from A to B, where

$$A \coloneqq (0, +\infty) \times (0, 2\pi), \qquad B \coloneqq \mathbb{R}^2 \setminus \{(x, y) \in \mathbb{R}^2 : x \geq 0,\ y = 0\}.$$

Indeed, f is of class C^1, it is injective (this is the reason why we consider the set A), and

$$Jf(r, \theta) = \begin{pmatrix} \cos\theta & -r \sin\theta \\ \sin\theta & r \cos\theta \end{pmatrix},$$

and thus $\det(Jf(r, \theta)) = r \neq 0$ on A.

We conclude by stating a result saying that there cannot be diffeomorphisms between spaces of different dimensions.

Corollary 7.20. *Let $f : \mathbb{R}^N \to \mathbb{R}^M$ be a C^1 map. If $M < N$, then f cannot be a diffeomorphism, not even locally. Indeed, f cannot be locally injective.*

On the other hand, if $M > N$, then it might be that f is a diffeomorphism from $\mathbb{R}^N$ onto $f(\mathbb{R}^N) \subset \mathbb{R}^M$.

Remark 7.21. A similar result holds for homeomorphisms: if $f : \mathbb{R}^N \to \mathbb{R}^M$ is a homeomorphism, then $N = M$. To prove such a result, you need tools from *Topology* that you'll learn next semester.

On the other hand, if no regularity is required on the map $f : \mathbb{R}^N \to \mathbb{R}^M$, then f can be a bijection for any $N, M \in \mathbb{N} \setminus \{0\}$. This was a result that shocked mathematicians. Indeed, this was a question that puzzled George Cantor: every person he talked to, was surprised about the question, since it was *evident* that to identify a point in $\mathbb{R}^M$ you need M coordinates, while to determine a point on $[0, 1]$ you just need one. Therefore, it is *evident* that there cannot be a bijection between the two sets. However, Cantor was not satisfied with a justification that relied on the *evidence* of such a fact. He then proved that such a bijection always exist (with an astonishingly simple proof!) that he wrote on 25 June 1877 to Richard Dedekind to get feedback. Four days after, having gotten not reply yet (yes, he was a bit anxious!), he wrote again to Dedekind writing[7] a sentence that became famous:

"Je le vois, mais je ne le crois pas." ["*I see it, but I don't believe it.*"]

This result called, once again, for a better foundation of Analysis, and for the need of rigorous proofs when dealing with mathematical objects with infinitely many points.

Moreover, the above result means that the number of points in $[0, 1]$ is the same as the number of points in any $\mathbb{R}^M$. As surprising as such result might seem at first sight (and even after years!), this means that, when dealing with sets with *infinitely many points*, our intuition can be misleading. Moreover, we get that the number of points is not really a good *measure* of the *size* of an object! We will see a proper definition of length, area, volume, and their higher dimensional versions in the last section devoted to Measure Theory.

[7] Note that the letter was in German, but this sentence was in French

8. Implicit Function Theorem

The goal of this section is to study the shape of the set of solutions of a system of equations. Namely, given C^1 functions $\Phi_i : \mathbb{R}^N \to \mathbb{R}$, for $i = 1, \dots, k$, with[8] $k \leq N$, we consider the system

$$\begin{cases} \Phi_1(x) = 0, \\ \Phi_2(x) = 0, \\ \vdots \\ \Phi_k(x) = 0. \end{cases} \tag{8.1}$$

We want to understand properties of the set

$$S \coloneqq \{x \in \mathbb{R}^N \,:\, \Phi_i(x) = 0 \text{ for all } i = 1, \dots, k\}$$

of its solutions. Note that, geometrically, the set S is the *intersection* of the sets

$$S_i \coloneqq \{x \in \mathbb{R}^N \,:\, \Phi_i(x) = 0\},$$

for $i = 1, \dots, k$. For instance, if we consider the case of $N = 3$, $k = 2$, and

$$\Phi_1(x, y, z) = x^2 + y^2 + z^2 - 1, \qquad \Phi_2(x, y, z) = 3x^2 + 4y^2 + 9z^2 - 1,$$

we want to understand the intersection of a sphere and an ellipsoid in $\mathbb{R}^3$. Note that, if we want to describe analytically the *union* of S_1 and S_2, we could do it as follows:

$$T \coloneqq \{x \in \mathbb{R}^N \,:\, \Phi_1(x)\Phi_2(x) = 0\}.$$

Indeed, $\Phi_1(x)\Phi_2(x) = 0$ if at least one of the two equations $\Phi_1(x) = 0$ or $\Phi_2(x) = 0$ is satisfied.

Let us start by reviewing a known case. Assume that all of the functions Φ_i's (also called *constraints*) are linear. Thus, for $i = 1, \dots, k$, we can write

$$\Phi_i(x) = \langle v_i, x \rangle,$$

for some $v_i \in \mathbb{R}^N$. Then, you know from *Linear Algebra* that each of the sets S_i's is an hyperplane, and that the set S is a linear subspace of $\mathbb{R}^N$. Moreover, the *dimension* of the linear space S is $N - d$, where d is the number of independent vectors in the set $\{v_1, \dots, v_k\}$.

When the functions are not linear, things are more complicated. We mention that the case of the Φ_i's being polynomials is studied in the branch of Mathematics called *Algebraic Geometry*, also for more general ambient spaces than $\mathbb{R}^N$. Let us try to get some heuristics for what we expect to happen in the general case of non-linear constraints. Consider the following example: $N = 2$, $k = 1$, and $\Phi_1 : \mathbb{R}^2 \to \mathbb{R}$ defined as

$$\Phi_1(x, y) \coloneqq x^2 + y^2 - 1.$$

Then, we know that S is the unit circle centered at the origin. In particular, we have that

(i) S is a *one dimensional* object;
(ii) S has a tangent line at all of its points;
(iii) It is possible to locally see S as the graph of a function defined over one of the two coordinate axes. This means that it is possible, for points $(x, y) \in S$ to locally express one of the two coordinates in terms of the other;
(iv) It is possible to locally see S as the graph of a function defined over its tangent line.

Does the same happen in the general case? Consider a general dimension N, and general k constraints. Let $\bar{x} \in S$. We want to understand the behavior of S around $\bar{x}$. Since, for all $i = 1, \dots, k$,

$$\Phi_i(x) = \Phi_i(\bar{x}) + \langle \nabla \Phi_i(\bar{x}), (x - \bar{x}) \rangle + o(\|x - \bar{x}\|),$$

[8] Note that, when the number k of equations is higher than the dimension N of the space, the system (8.1) does not have, in general, a solution.

we would expect S to look like (at first order) the set of solutions to the *linearization* of the system (8.1), namely, close to the affine space given by the set of solutions to the system

$$\begin{cases} \langle \nabla \Phi_1(\bar{x}), (x - \bar{x}) \rangle + \Phi_1(\bar{x}) = 0, \\ \langle \nabla \Phi_2(\bar{x}), (x - \bar{x}) \rangle + \Phi_2(\bar{x}) = 0, \\ \vdots \\ \langle \nabla \Phi_k(\bar{x}), (x - \bar{x}) \rangle + \Phi_k(\bar{x}) = 0. \end{cases} \tag{8.2}$$

Thus, we expect S to locally look like an affine space. In particular, this means that S can locally be written as a graph of a function defined on the affine subspace determined by (8.2). Namely, it is possible, for points $x \in S$ to express some of the coordinates in terms of the other. Sets that can be locally described as the solutions to a system of equations are called *submanifolds*.

The path we will use to prove rigorously the above result is by establishing the so called *Implicit Function Theorem* (see Theorem 8.1), which states that it is possible to express some of the coordinates of points $x \in S$ in terms of the other coordinates.

Before continuing with the rigorous mathematics, let's see what can go wrong with the heuristics detailed above. Consider the case of a single equation, $\Phi_1 : \mathbb{R}^2 \to \mathbb{R}$ defined as

$$\Phi_1(x) \coloneqq x^2 - y^2.$$

Then, the set S is the union of the two lines $\{y = x\}$ and $\{y = -x\}$. This set has a *singularity* at the point $(0, 0)$, since it is not possible to represent S as the graph of a function, or, equivalently, the tangent cone to the set S at the origin is not a one dimensional linear space. The problem of the point $(0, 0)$ is that $\nabla \Phi_1(0, 0) = (0, 0)$. Thus, the gradient does not give us any information on the local behavior of the function Φ_1, and, in turn, no local information on the shape of the set S. This is where the previous heuristics fails.

Another example, that generalized the one above, where things go wrong is the following. Consider the constraints $\Phi_1, \Phi_2 : \mathbb{R}^3 \to \mathbb{R}$ defined as

$$\Phi_1(x, y, z) \coloneqq x^2 + y^2 - 1, \qquad \Phi_2(x, y, z) \coloneqq x^2 + z^2 - 1.$$

Then, the set S is the intersection of two cylinders. To understand what happens, we rewrite the constraints in the following way:

$$S = \{(x, y, z) \in \mathbb{R}^3 \, : \, x^2 + y^2 - 1 = 0, \, z^2 - y^2 = 0\}.$$

In this way, it is easier to see that S is the union of two one-dimensional ellipses. Indeed, from $\Phi_1(x, y, z) = 0$ we get $x^2 = 1 - y^2$; by substituting this in the expression for Φ_2, we get

$$0 = z^2 - y^2 = (z - y)(z + y).$$

Thus, we see that points in S are points $(x, y, z) \in \mathbb{R}^3$ for which $\Phi_1(x, y, z) = 0$ and either $z = y$ or $z = -y$. This means that $S = S_1 \cup S_2$, where

$$S_1 \coloneqq \{(x, y, z) \in \mathbb{R}^3 \, : \, \Phi_1(x, y, z) = 0, z = y\},$$

$$S_2 \coloneqq \{(x, y, z) \in \mathbb{R}^3 \, : \, \Phi_1(x, y, z) = 0, z = -y\}.$$

Since the set $\coloneqq \{(x, y, z) \in \mathbb{R}^3 \, . \, \Phi_1(x, y, z) = 0\}$ is a cylinder, the sets S_1 and S_2 are one-dimensional ellipses, being the intersection of a cylinder and a plane.

Now, the set S has two singular points, $(\pm 1, 0, 0)$. What goes wrong at these points? Well, let us look at the gradients of the constraints. We have

$$\nabla \Phi_1(x, y, z) = (2x, 2y, 0), \qquad \nabla \Phi_2(x, y, z) = (2x, 0, 2z).$$

The two gradients fail to be *linearly independent* for points $(x, y, z) \in \mathbb{R}^3$ such that $y = z = 0$. Of these points, those that belong to the set S are $(\pm 1, 0, 0)$. This reflects on the fact that, at those points, the gradients only provide us with *just one* information at the singular points. This allows the set S to behave wildly, and this is exactly what happens: there are two independent directions that are tangent to S at the singular points, despite the fact that we expect S to be

one dimensional, since we have two equations in $\mathbb{R}^3$.

Therefore, by using the information given by the gradients of the constraints, we will be able to locally describe the set of solutions far from the singularities. The study of the behavior of the set S at a singular point is extremely delicate, and goes way beyond the scope of the course. The only information we can provide is that the set of singular points is contained (note that it might be a strict inclusion!) in the set where the gradients are not linearly independent. Unfortunately, this information is very weak to extract any useful insight on the local behavior of S at a singular point.

Before stating the main result of this section, we need to work a bit on the notation. We first rewrite the system (8.1) as a single equation as follows: let $f : \mathbb{R}^N \to \mathbb{R}^k$ be defined as

$$f(x) \coloneqq (\Phi_1(x), \dots, \Phi_k(x)) .$$

Moreover, since we want to ask the gradients of the constraints to be linearly independent, we naturally[9] require $k \leq N$. Thus, we can write the domain $\mathbb{R}^N$ of the function f as

$$\mathbb{R}^{N-k} \times \mathbb{R}^k,$$

and denote variables as $(x, y) \in \mathbb{R}^{N-k} \times \mathbb{R}^k$, namely

$$x = (x_1, \dots, x_{N-k}) \in \mathbb{R}^{N-k}, \qquad y = (y_1, \dots, y_k) \in \mathbb{R}^k.$$

Then, we need some notation for the Jacobian matrix of f to separate the two sets of variables. We write

$$Jf = \left(\begin{array}{ccc|ccc} \partial_{x_1} f_1 & \cdots & \partial_{x_{N-k}} f_1 & \partial_{y_1} f_1 & \cdots & \partial_{y_k} f_1 \\ \vdots & \dots & \vdots & \vdots & \dots & \vdots \\ \partial_{x_1} f_k & \cdots & \partial_{x_{N-k}} f_k & \partial_{y_1} f_k & \cdots & \partial_{y_k} f_k \end{array} \right) = \left(\begin{array}{c|c} \frac{\partial(f_1, \cdots f_k)}{\partial x_1 \cdots \partial x_{N-k}} & \frac{\partial(f_1, \cdots f_k)}{\partial y_1 \dots \partial y_k} \end{array} \right).$$

Note that the matrix

$$\left(\frac{\partial(f_1, \cdots f_k)}{\partial y_1 \dots \partial y_k} \right).$$

is a $k \times k$ matrix. We will prove the result for a general *level set* $c \in \mathbb{R}^k$ of f.

We are now in position to prove the Implicit Function Theorem[10]. Actually, the proof will give us a stronger result: namely we will be able to get a local linearization of $\mathbb{R}^{N-k} \times \mathbb{R}^k$ by using a *foliation by the level sets* of the function f.

Theorem 8.1 (Implicit Function Theorem)**.** *Let $f : \mathbb{R}^{N-k} \times \mathbb{R}^k \to \mathbb{R}^k$ be a C^1 function. Let $(\bar{x}, \bar{y}) \in \mathbb{R}^{N-k} \times \mathbb{R}^k$ be such that $f(\bar{x}, \bar{y}) = c$, for some $c \in \mathbb{R}^k$, and*

$$\det \left(\frac{\partial(f_1, \cdots f_k)}{\partial y_1 \dots \partial y_k}(\bar{x}, \bar{y}) \right) \neq 0. \tag{8.3}$$

Namely, the vectors

$$\left(\begin{array}{c} \partial_{y_1} f_1(\bar{x}, \bar{y}) \\ \vdots \\ \partial_{y_1} f_k(\bar{x}, \bar{y}) \end{array} \right), \dots, \left(\begin{array}{c} \partial_{y_k} f_1(\bar{x}, \bar{y}) \\ \vdots \\ \partial_{y_k} f_k(\bar{x}, \bar{y}) \end{array} \right)$$

are linearly independent. Then, there exist open sets $X \subset \mathbb{R}^{N-k}$, $Y \subset \mathbb{R}^k$, and a C^1 function $g : X \to Y$ such that $\bar{x} \in X$, $\bar{y} \in Y$ and

$$\{f = c\} \cap (X \times Y) = \{(x, g(x)) : x \in X\}.$$

[9] Note that, when the number k of equations is higher than then dimension N of the space, the system (8.1) does not have, in general, a solution.

[10] Ulisse Dini generalized the result by Augustin-Louis Cauchy to the case of a function of several variables. This is the reason why, in Italy, the Implicit Function Theorem is known as (one of) the Dini's Theorem(s).

Proof. We start by explaining the idea of the proof with an example. Consider the function $f : \mathbb{R}^2 \times \mathbb{R} \to \mathbb{R}$ given by

$$f(x_1, x_2, y_1) = x_1^2 + x_2^2 + y_1^2,$$

and the point $P = (\bar{x}_1, \bar{x}_2, \bar{y}_1) = (0, 0, 1) \in \{f = 1\}$. Note that, in this example, $c = 1$. We have that

$$Jf(x_1, x_2, y_1) = (2x_1, 2x_2, 2y_1).$$

In particular, condition (8.3) writes as

$$2\bar{y}_1 \neq 0. \tag{8.4}$$

We have that $\{f = 1\}$ is a sphere of radius 1. Moreover, for every $c \in (-\varepsilon, +\varepsilon)$, for $|\varepsilon| < 1$, we have that the set $\{f = c\}$ is a sphere. In particular, we know that we can locally around P describe *every* set $\{f = c\}$ as the graph of a function g that depends on the first two coordinates x_1, x_2 and on the level c. Namely, we can *foliate* the space $\mathbb{R}^2 \times \mathbb{R}$ around the point P by the level sets of the function f, and we can *linearize* this neighborhood by using the map g. To be more precise, we consider the function $F : [-1, 1]^2 \times \mathbb{R} \to [-1, 1]^2 \times [-\varepsilon, \varepsilon]$ defined as

$$F(x_1, x_2, y_1) \coloneqq (x_1, x_2, f(x_1, x_2, y_1)).$$

Note that the codomain of F is a *rectangle*. This means that we are linearizing a neighborhood of $\mathbb{R}^3$ around P. By using (8.4), we get that

$$\det JF(x_1, x_2, y_1) = \partial_{y_1} f(x_1, x_2, y_1) = 2y_1 \neq 0,$$

for (x_1, x_2, y_1) close to P. Since F is of class C^1 and clearly injective, by using the Inverse Function Theorem (see Theorem 7.17), we get that the map F is a diffeomorphism, and thus we can use it as a change of coordinates. In particular, the inverse of the last component of F will give us the desired inverse function g. In this case, we know that we can write $\{f = c\}$ around P as the graph of the map

$$\varphi(x_1, x_2, c) \coloneqq \sqrt{c - x_1^2 - x_2^2},$$

and that the function we want corresponds to the function φ at the level $c = 1$. Namely, the implicit function we were looking for is given by

$$g(x_1, x_2) = \varphi(x_1, x_2, 1) = \sqrt{1 - x_1^2 - x_2^2}.$$

Note that φ is the inverse of the last component of F.

Let us now use this strategy for the general case. Since f is of class C^1, we have that

$$(x, y) \mapsto \det \left(\tfrac{\partial(f_1, \cdots f_k)}{\partial y_1 \dots \partial y_k}(x, y) \right)$$

is continuous, being a polynomial function of the partial derivatives. Therefore, by using (8.3) together with the continuity of the determinant, there exist an open neighborhood $A \subset \mathbb{R}^{N-k}$ of $\bar{x}$, and an open neighborhood $B \subset \mathbb{R}^k$ of $\bar{y}$ such that

$$\det \left(\tfrac{\partial(f_1, \cdots f_k)}{\partial y_1 \dots \partial y_k}(x, y) \right) \neq 0,$$

for all $(x, y) \in A \times B$. Define the function $F : A \times B \to \mathbb{R}^{N-k} \times \mathbb{R}^k$ as

$$F(x, y) \coloneqq (x, f(x, y)).$$

We claim there exist open sets $X \subset A$, and $Y \subset B$ such that $F : X \times Y \to \mathbb{R}^{N-k} \times \mathbb{R}^k$ is a diffeomorphism. Indeed,

$$\det(JF(x, y)) = \det \left(\begin{array}{c|c} \mathrm{Id}_{\mathrm{N-k}} & 0 \\ \hline \frac{\partial(f_1, \cdots f_k)}{\partial x_1 \cdots \partial x_{N-k}}(x, y) & \frac{\partial(f_1, \cdots f_k)}{\partial y_1 \dots \partial y_k}(x, y) \end{array} \right) = \det \left(\tfrac{\partial(f_1, \cdots f_k)}{\partial y_1 \dots \partial y_k}(x, y) \right) \neq 0,$$

for all $(x,y) \in X \times Y$. Here $\mathrm{Id_{N-k}}$ denotes the $(N-k)\times(N-k)$-identity matrix. Therefore, from the Inverse Function Theorem (see Theorem 7.17), we get that there exist open sets $X \subset A$, and $Y \subset B$ such that $F : X \times Y \to \mathbb{R}^{N-k} \times \mathbb{R}^k$ is a diffeomorphism. In particular, there exists its inverse $h : F(X \times Y) \to X \times Y$, and it is of class C^1. Note that the inverse is of the form

$$h(x,y) = (x, \varphi(x,y)),$$

for some C^1 function $\varphi : X \times Y \to Y$. As discussed above, we are interested in the level set c of the function f. This is why we define $g : X \to Y$ as

$$g(x) \coloneqq \varphi(x,c).$$

First of all, we note that the function g is of class C^1 We now have to check that $\{f = c\}$ is locally the graph of the function g. Namely, we have to make sure that

$$f(x, g(x)) = c,$$

for all $x \in X$. Let $\pi : X \times Y \to Y$ be the projection on the second coordinate, namely $\pi(x,y) \coloneqq y$. Note that we can write $f = \pi \circ F$. We have that

$$\begin{aligned} f(x,g(x)) &= f(x, \varphi(x,c)) \\ &= (f \circ h)(x,c) \\ &= ((\pi \circ F) \circ h)(x,c) \\ &= (\pi \circ (F \circ h))(x,c) \\ &= \pi(x,c) \\ &= c, \end{aligned}$$

where in the previous to last step we used the fact that $F \circ h$ is the identity. This concludes the proof of the theorem. □

Remark 8.2. Note that, by the way we defined it, the function $g : X \to Y$ is an homeomorphism. Moreover, the vectors

$$\partial_{x_1} g(\bar{x}), \dots, \partial_{x_{N-k}} g(\bar{x})$$

are linearly independent. Thus, g plays the role of the local *parametrization* of the set S of solutions to the systems of equations. Indeed, we can express *every* point in

$$\{f = c\} \cap (X \times Y)$$

in a unique way as $(x, g(x))$, for some *parameter* $x \in X$. In this case, the parameter is given by the first $N-k$ coordinates of the point $(x,y) \in \{f = c\}$.

Remark 8.3 (The gradient of the implicit function)**.** The implicit function provided by the above theorem is differentiable. Can we compute its gradient? Recall that $f : \mathbb{R}^{N-k} \times \mathbb{R}^k \to \mathbb{R}^k$, and that $g : X \to Y$, with $X \subset \mathbb{R}^{N-k}$, and $Y \subset \mathbb{R}^k$. For any $i = 1, \dots, k$, and $j = 1, \dots, N-k$, consider the equality

$$f_i(x, g(x)) = 0.$$

We can differentiate both sides of it with respect to x_j. By using the chain rule (see Proposition 5.39), we get

$$\partial_{x_j} f_i(x, g(x)) + \sum_{r=1}^{k} \partial_{y_r} f_i(x, g(x)) \partial_{x_j} g_r(x) = 0. \tag{8.5}$$

Thus, we get a system of $k(N-k)$ affine equations in $k(N-k)$ unknowns $\partial_1 g(x), \dots, \partial_k g(x)$. Note that it is possible to write the system of equations (8.5) as follows:

$$\left(\frac{\partial(f_1, \cdots f_k)}{\partial y_1 \dots \partial y_k}(x, g(x)) \right) \cdot Jg(x) = - \left(\frac{\partial(f_1, \cdots f_k)}{\partial x_1 \cdots \partial x_{N-k}}(x, g(x)) \right).$$

Thus, by assumption (8.3), we get that the system admits a unique solution, given by

$$Jg(x) = -\left(\frac{\partial(f_1, \cdots f_k)}{\partial y_1 \dots \partial y_k}(x, g(x))\right)^{-1} \cdot \left(\frac{\partial(f_1, \cdots f_k)}{\partial x_1 \cdots \partial x_{N-k}}(x, g(x))\right).$$

Note that the expressions for the partial derivatives $\partial_j g_r(x)$ are written in terms of $g(x)$ itself, so in an implicit form. This is unavoidable. Sometimes, these expressions can be made explicit.

Example 8.4. Let us consider the following example: let $N = 2$, $k = 1$, and $\Phi_1 : \mathbb{R}^2 \to \mathbb{R}$ given by

$$\Phi_1(x_1, x_2) \coloneqq x_1^2 + x_2^2 - 1.$$

Then, $f : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is Φ_1 itself, where we now denote x_2 by y_1 (this is a choice; of course, you can choose to invert the first variable as well.). Since

$$\partial_{x_1} f(x_1, x_2) = 2x_1, \qquad \partial_{x_2} f(x_1, x_2) = 2x_2,$$

we get that it is possible to apply the Implicit Function Theorem (see Theorem 8.1) at every point

$$(x_1, y_1) \in S \setminus \{(\pm 1, 0)\},$$

At such a point, we get that there exists a C^1 function $g : X \to \mathbb{R}$, for a certain interval X containing the point x_1, such that

$$f\left(x_1, g(x_1)\right) = 0.$$

Thus, by taking the derivative with respect to x_1, we get

$$0 = \partial_{x_1} f\left(x_1, g(x_1)\right) + \partial_{y_1} f\left(x_1, g(x_1)\right) g'(x_1) = 2x_1 + 2g(x_1)g'(x_1),$$

from which we get

$$g'(x_1) = -\frac{x_1}{g(x_1)}.$$

This is the implicit expression for the gradient of g. You can solve it explicitly by using the theory of *Ordinary Differential Equations*, and get that

$$g(x_1) = \pm \frac{x_1}{\sqrt{1 - x_1^2}},$$

and choose the sign of g in accordance with the position of the point (x_1, y_1) where you initially wanted to locally describe S.

We now show that the Implicit and the Inverse Function Theorem are equivalent.

Proposition 8.5. *The Implicit Function Theorem (see Theorem 8.1) and the Inverse Function Theorem (see Theorem 7.17) are equivalent.*

Proof. The proof of the Implicit Function Theorem (see Theorem 8.1) used the Inverse Function Theorem (Theorem 7.17). On the other hand, if we assume the Implicit Function Theorem to hold, then it is possible to prove the Inverse Function Theorem as follows: let $h : \mathbb{R}^N \to \mathbb{R}^N$ be a function of class C^1. Assume that $\bar{x} \in \Omega$ is such that $\det(Jh(\bar{x})) \neq 0$. Take $k = N$, and let $f : \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}^N$ defined as

$$f(x, y) \coloneqq h(x) - y.$$

Then, f satisfies the assumption of the Implicit Function Theorem. Indeed,

$$Jf(x, y) = \left(Jh(x, y) \mid -\mathrm{Id}\right).$$

Thus, since $\det(-\mathrm{Id}) \neq 0$, it is possible to write the variables y in terms of the variable x; but this is of no interest for us, since we simply get $y = h(x)$. What is interesting, is that, since

$$\det(Jh(\bar{x})) \neq 0,$$

it is locally possible to write the variable x in terms of the variable y. More precisely, there exist open sets $X, Y \subset \mathbb{R}^N$ with $\bar{x} \in X$, and $h(\bar{x}) \in Y$, and a C^1 function $g : X \to Y$ such that

$$f\left(g(y), y\right) = 0,$$

for all $y \in Y$. Namely, $h(g(y)) = y$ for all $y \in Y$. Thus, g is the inverse function we were looking for. $\square$

9. Analysis on submanifolds

9.1. **Submanifolds in $\mathbb{R}^N$.** Let us consider an example that will illustrate the geometric meaning of the main results that will be presented in this section. Consider the unit circle in $\mathbb{R}^2$ centered at the origin. What we know is the following:

(i) Around each point it is possible to describe the circle as the set of solutions to a system of equations (actually, in this case, just one equation, and the same equation works for all the points in the circle);
(ii) Around each point it is possible to locally flattened the circle;
(iii) It is possible to describe the circle by using polar coordinates; namely, the circle is a one-dimensional object;
(iv) Around each point, it is possible to describe the circle as a graph over one of the coordinate axes;
(v) At each point it has a tangent line.

We want to study objects that are like the circle. Namely, sets that can be locally flattened like a space $\mathbb{R}^d$, that can be locally described by using d parameters, that can be locally described as the set of solutions to a system of $N-d$ equations, or by a graph over some coordinate axes, and that possess a tangent space that is a linear space at each point. All such properties are all connected with each other, and all follow from the Implicit Function Theorem (see Theorem 8.1), the Inverse Function Theorem (see Theorem 7.17) together with properties of the differential of a diffeomorphism (see Proposition 7.15(iv)).

Definition 9.1. Let $S \subset \mathbb{R}^N$, and $\bar{x} \in S$. We say that S is a *submanifold (of class C^1) of dimension d at $\bar{x}$* if and only if there exist an open set $U \subset \mathbb{R}^N$ with $\bar{x} \in U$, and C^1 functions $\Phi_1, \dots, \Phi_{N-d} : U \to \mathbb{R}$ such that

$$S \cap U = \{x \in U : \Phi_1(x) = 0, \dots, \Phi_{N-d}(x) = 0\},$$

and such that

$$\nabla \Phi_1(x), \dots, \nabla \Phi_{N-d}(x)$$

are linearly independent for all $x \in U$.

Remark 9.2. Note that it is possible to check the condition of linear independence of the gradients only at the point $\bar{x}$. Indeed, suppose that the set S is locally described in an open neighborhood $U \subset \mathbb{R}^N$ of $\bar{x}$ as the set of solutions to a system of equations as in the definition above. Assume that

$$\nabla \Phi_1(\bar{x}), \dots, \nabla \Phi_{N-d}(\bar{x})$$

are linearly independent . Then, by the continuity of the gradients, we get that there exists an open set $A \subset \mathbb{R}^N$ with $\bar{x} \in A$ such that the gradients

$$\nabla \Phi_1(x), \dots, \nabla \Phi_{N-d}(x)$$

are linearly independent for all $x \in A$. Thus, by taking $U' := U \cap A$, we get that S is a d-dimensional submanifold at the point $\bar{x}$.

Example 9.3. Let's consider the unit sphere of $\mathbb{R}^N$ centered at the origin. We claim that S is a submanifold of dimension $d = N-1$ at each of its points. Indeed, it is possible to describe

$$S = \{x \in \mathbb{R}^N : \|x\|^2 = 1\}.$$

Since the function $\Phi_1(x) := \|x\|^2 - 1$ is a C^1 function, and

$$\nabla \Phi_1(x) = 2x \neq 0,$$

for all $x \in S$, we have the desired result.

Note that, in this case, we do not need an open set set U to localize the description of S as the set of solutions to an equation.

Example 9.4. Let S be the boundary of the square of side 2 in $\mathbb{R}^2$ centered at the origin. We claim that S is a submanifold of dimension 1 at all of its points, except at the corners. Indeed, let $\bar{x} = (\bar{x}_1, \bar{x}_2) \in S$. Assume that it is a point on one of the horizontal sides, but not a vertex. The case where it is a point on the vertical sides is treated similarly. Consider the open set $U \coloneqq B(x, r)$, where $r \coloneqq 1 - |\bar{x}_1|$. Define the function $\Phi_1 : \mathbb{R}^2 \to \mathbb{R}$ as

$$\Phi_1(x_1, x_2) \coloneqq 1 - |x_2|^2.$$

Then, we have that

$$S \cap U = \{(x_1, x_2) \in U \,:\, \Phi_1(x_1, x_2) = 0\}.$$

Moreover, thanks to the assumption on the point $\bar{x}$ and on the set U, we also have that

$$\nabla \Phi_1(x_1, x_2) = -2x_2 \neq 0,$$

and thus the desired result follows.

Remark 9.5. Note that the property of being a submanifold is *local*, not pointwise. Indeed, let S be a submanifold of dimension d at the point $\bar{x} \in S$. Then, there exists an open set $U \subset \mathbb{R}^N$ containing $\bar{x}$ such that S is a submanifold of class d at all points $x \in S \cap U$.

On the other hand, as we have seen in the previous examples, the set of points where a set S is not a submanifold, can also be closed (in the precious example, the four corners of the square).

Remark 9.6. A trivial case of a submanifold, is when $d = N$: these are open sets of $\mathbb{R}^N$. We are not interested in such a case. In particular, we only care about submanifolds of dimension $d = 0, 1, 2, \ldots, N-1$. Note that, in the case $d = 0$, the set S is made by isolated points (prove it!).

Remark 9.7. In particular, a submanifold of dimension d in $\mathbb{R}^N$ is *locally* the intersection of $N - d$ submanifolds of dimension $N - 1$.

By using the same argument, we obtain that the intersection of a d_1-dimensional submanifold S_1 and a d_2-dimensional submanifold S_2 of $\mathbb{R}^N$ is a[11] $2N - (d_1 + d_2)$ submanifold in all of the points where the gradients of the functions describing locally S_1 and S_2 around the intersection point are linearly independent.

As an example, consider the intersection of two spheres S_1 and S_2 in $\mathbb{R}^N$. We can describe them as

$$S_i \coloneqq \left\{x \in \mathbb{R}^N \,:\, \|(x - P_i)\| - r_i = 0\right\}.$$

for some $P_i \in \mathbb{R}^N$, and some $r_i > 0$. At points at which they are not tangent (and those are the points at which the gradients of the functions describing them are not linearly independent), the set $S_1 \cap S_2$ is a submanifold of dimension $N - 2$, and it is given by

$$S_1 \cap S_2 = \left\{x \in \mathbb{R}^N \,:\, \|(x - P_1)\| - r_1 = 0, \quad \|(x - P_2)\| - r_2 = 0\right\}.$$

You can check that this is an $(N-2)$-dimensional sphere in $\mathbb{R}^N$.

Remark 9.8. In case the functions Φ_i's describing locally the set S are of class C^r, for $r \geq 1$, we say that the submanifold is a submanifold of class C^r. We usually omit this specification in the case $r = 1$.

Remark 9.9. The reason why the terminology we use for S is *sub*manifold and not just manifold, is because we are looking at S as an object contained in $\mathbb{R}^N$. This latter plays the role of what is called a manifold. In the course *Manifolds*, you'll develop the general theory for abstract manifolds. This goes beyond the scope of this course, since it requires a higher level of abstraction. Here, we would like just to mention some results that could be of interest for the curious reader. When you'll study manifolds, you'll see that the set

$$S \coloneqq \{(x, y) \in \mathbb{R}^2 \,:\, y - |x| = 0\}$$

is a manifold, but not a submanifold of $\mathbb{R}^2$ at the origin (since it has a corner at that point).

[11] Note that a submanifold of dimension d in $\mathbb{R}^N$ is locally described by a set of $N - d$ equations.

Finally, it is worth mention a very important result, the *Nash's embedding theorem*, stating that every abstract manifold can be embedded (namely, it is diffeomorphic) to a submanifold of $\mathbb{R}^N$. On the one hand, this result says that all abstract notions can be reduced to known cases; on the other hand, this does not diminish the importance of abstract manifolds, since they are able to provide a structure of sets in very general situations.

Remark 9.10. Note the following: consider the set

$$S \coloneqq \{x \in \mathbb{R}^N \,:\, \|x\| < 1\}.$$

Then, S is an open set, and thus, thanks to Remark 9.6 above, it is a submanifold of $\mathbb{R}^N$ of dimension N. Its topological boundary ∂S is

$$\partial S = \left\{x \in \mathbb{R}^N \,:\, \|x\| = 1\right\}.$$

By using similar computations to those of Example 9.3, we have that ∂S is a submanifold of $\mathbb{R}^3$ of dimension $d = 2$. Moreover, its topological boundary is empty. In particular, these are the sets for which you know how to apply theorems that you saw in *Calculus*, such as the Divergence Theorem, and the Gauss-Green Theorem.

The above mentioned situation is typical for submanifolds that are regular enough: the submanifold S with dimension d has a topological boundary ∂S that is a submanifold of dimension $d-1$, and its topological boundary is empty. You'll study the general setup for such situation in the course *Manifold*.

An important example of submanifolds are graphs of C^1 functions $f : \mathbb{R}^N \to \mathbb{R}^M$.

Proposition 9.11. *Let $f : \mathbb{R}^N \to \mathbb{R}^M$ be a function of class C^1. Then,*

$$\operatorname{graph}(f) \coloneqq \left\{\, (x, f(x)) \in \mathbb{R}^N \times \mathbb{R}^M \,\right\}$$

is a submanifold of $\mathbb{R}^{N+M}$ of dimension N at each of its points.

Proof. Denote by (x, y) a point in $\mathbb{R}^N \times \mathbb{R}^M$, and, for $i = 1, \dots, M$, denote by $f_i(x)$ the i^{th} component of $f(x)$. Define, for $i = 1, \dots, M$, the function $\Phi_i : \mathbb{R}^N \times \mathbb{R}^M \to \mathbb{R}$ as

$$\Phi_i(x, y) \coloneqq y_i - f_i(x).$$

Then, each Φ_i is a function of class C^1, and

$$\operatorname{graph}(f) = \{(x, y) \in \mathbb{R}^N \times \mathbb{R}^M \,:\, \Phi_1(x, y) = 0, \dots, \Phi_M(x, y) = 0\}.$$

Moreover, at each point $(x, y) \in \mathbb{R}^N \times \mathbb{R}^M$, we have that

$$\nabla \Phi_i(x, y) = -(\nabla f_i(x), e_i),$$

where e_i is the i^{th} vector of the canonical bases of $\mathbb{R}^M$. Thus, the vectors

$$\nabla \Phi_1(x, y), \dots, \nabla \Phi_M(x, y)$$

are linearly independent at all points $(x, y) \in \mathbb{R}^N \times \mathbb{R}^M$. This proves the desired result. □

As mentioned at the beginning of this section, S being a submanifold means that it can be locally flattened to a space $\mathbb{R}^d$. In particular, we have the following rewriting of the definition of a manifold.

Proposition 9.12. *Let $S \subset \mathbb{R}^N$, and let $\bar{x} \in S$. Then, S is a d-dimensional submanifold of $\mathbb{R}^N$ at the point $\bar{x}$ if and only if it is possible to find an open set $U \subset \mathbb{R}^N$ with $\bar{x} \in U$, and a C^1 diffeomorphism $\varphi : U \to \mathbb{R}^N$ such that*

$$\varphi\left(S \cap U\right) = A \times \{0\},$$

where $A \subset \mathbb{R}^d$ is an open set, and $0 \in \mathbb{R}^{N-d}$.

Proof. Step 1. Let S be a d-dimensional submanifold of $\mathbb{R}^N$ at the point $\bar{x}$. By definition, we can find an open set $U \subset \mathbb{R}^N$ with $\bar{x} \in U$, and C^1 functions $\Phi_1, \dots, \Phi_{N-d} : U \to \mathbb{R}$ such that

$$S \cap U = \{x \in U : \Phi_1(x) = 0, \dots, \Phi_{N-d}(x) = 0\},$$

and such that

$$\nabla\Phi_1(\bar{x}), \dots, \nabla\Phi_{N-d}(\bar{x})$$

are linearly independent. Consider the matrix

$$\begin{pmatrix} \nabla\Phi_1(\bar{x}) \\ \vdots \\ \nabla\Phi_{N-d}(\bar{x}) \end{pmatrix}.$$

This is a $(N-d)\times N$ matrix. Without loss of generality (namely, up to renaming the coordinates), we can assume that the last $(N-d)$ columns are linearly independent. Now, write a point $x \in \mathbb{R}^N$ as

$$x = (x_1, \dots, x_d, x_{d+1}, \dots x_N) =: (x', x'') \in \mathbb{R}^d \times \mathbb{R}^{N-d}.$$

Thus, we can write the above matrix as follows

$$\begin{aligned} \begin{pmatrix} \nabla\Phi_1(\bar{x}) \\ \vdots \\ \nabla\Phi_{N-d}(\bar{x}) \end{pmatrix} &= \left(\begin{array}{c|c} \frac{\partial(\Phi_1,\cdots\Phi_{N-d})}{\partial x_1 \cdots \partial x_d}(\bar{x}) & \frac{\partial(\Phi_1,\cdots\Phi_{N-d})}{\partial x_{d+1}\cdots\partial x_N}(\bar{x}) \end{array} \right) \\ &= \left(\begin{array}{c|c} \frac{\partial(\Phi_1,\cdots\Phi_d)}{\partial x'}(\bar{x}) & \frac{\partial(\Phi_1,\cdots\Phi_{N-d})}{\partial x''}(\bar{x}) \end{array} \right). \end{aligned}$$

By what we said above, the $(N-d)\times(N-d)$ matrix

$$\left(\frac{\partial(\Phi_1, \cdots \Phi_{N-d})}{\partial x''}(\bar{x})\right)$$

has linearly independent columns, and thus

$$\det\left(\frac{\partial(\Phi_1, \cdots \Phi_{N-d})}{\partial x''}(\bar{x})\right) \neq 0.$$

Define the function $\varphi : \mathbb{R}^d \times \mathbb{R}^{N-d} \to \mathbb{R}^N$ as

$$\varphi(x', x'') \coloneqq (x', \Phi_1(x), \dots, \Phi_{N-d}(x)).$$

Then, φ is a function of class C^1, and

$$\varphi(S \cap U) = A \times \{0\},$$

where A is the projection of the open set U on the first d coordinates. Moreover,

$$\begin{aligned} \det(J\varphi)(\bar{x}) &= \det \left(\begin{array}{c|c} \mathrm{Id_d} & 0 \\ \hline \frac{\partial(\Phi_1,\cdots\Phi_{N-d})}{\partial x'}(\bar{x}) & \frac{\partial(\Phi_1,\cdots\Phi_{N-d})}{\partial x''}(\bar{x}) \end{array} \right) \\ &= \det \left(\frac{\partial(\Phi_1,\cdots\Phi_{N-d})}{\partial x''}(\bar{x}) \right) \\ &\neq 0. \end{aligned}$$

Thus, from the Inverse Function Theorem (see Theorem 7.17), we get that φ is locally a diffeomorphism, as desired.

Step 2. Assume that S is locally described as

$$\varphi(S \cap U) = A \times \{0\},$$

where $A \subset \mathbb{R}^d$ is an open set, $0 \in \mathbb{R}^{N-d}$, and $\varphi : \mathbb{R}^N \to \mathbb{R}^N$ is a diffeomorphism. Then, for each $i = 1, \dots, N-d$, define the function $\Phi_i : \mathbb{R}^N \to \mathbb{R}$ as the $(d+i)^{th}$ component of the function φ.

Then, each Φ_i is of class C^1, and the gradients $\nabla\Phi_1(x), \dots, \nabla\Phi_{N-d}(x)$ are linearly independent for all $x \in U$, since φ is a diffeomorphism. This gives the desired result. □

Remark 9.13. The above result clarifies what we mean by '*a submanifold can be locally flattened to look like a linear space* $\mathbb{R}^d$'.

As we saw in the introduction of this section, locally, the unit circle $S \subset \mathbb{R}^2$ centered at the origin can also be given by using parametrization. Namely

$$S = \{(\cos\theta, \sin\theta) \,:\, \theta \in (\theta_0, \theta_1)\},$$

for some $0 < \theta_0 < \theta_1 < 2\pi$. We would like to show that this is the case for any submanifold; namely, we now provide a characterization of submanifolds by using a local *parametrization*.

Proposition 9.14. *Let $S \subset \mathbb{R}^N$, and $\bar{x} \in S$. Then, S is a submanifold of dimension d at $\bar{x}$ if and only if there exist an open set $U \subset \mathbb{R}^N$ with $\bar{x} \in U$, an open set $A \subset \mathbb{R}^d$ with $\bar{\lambda} \in A$, such that*

$$S \cap U = \Psi(A),$$

where $\Psi : A \to \mathbb{R}^N$ is a C^1 function such that:

(i) Ψ is an homeomorphism;

(ii) $\partial_1\Psi(\bar{\lambda}), \dots, \partial_d\Psi(\bar{\lambda})$ are linearly independent;

and $\Psi(\bar{\lambda}) = \bar{x}$.

Proof. Step 1. Assume S is a submanifold of $\mathbb{R}^N$ of dimension d at the point $\bar{x}$. Define the function $f : \mathbb{R}^d \times \mathbb{R}^{N-d} \to \mathbb{R}^{N-d}$ as

$$f(x) \coloneqq (\Phi_1(x), \dots, \Phi_{N-d}(x)).$$

Arguing as in the proof of Proposition 9.12, we can assume that

$$\det\left(\tfrac{\partial(f_1, \cdots f_{N-d})}{\partial x''}(\bar{x}) \right) \neq 0.$$

Thus, the function f satisfies by using the Implicit Function Theorem (see Theorem 8.1). Therefore, there exist open sets $A \subset \mathbb{R}^d$, and $B \subset \mathbb{R}^{N-d}$, a C^1 function $\Phi : A \times \{0\}$ (note that in the theorem this function is called g, and the sets A and B are x and Y respectively), where $0 \in \mathbb{R}^{N-d}$ such that (see Remark 8.2)

(i) Ψ is a homeomorphism;

(ii) the vectors

$$\partial_{x_1}\Psi(\bar{\lambda}), \dots, \partial_{x_d}\Psi(\bar{\lambda})$$

are linearly independent, where $\bar{x} = \Psi(\bar{\lambda})$;

(iii) $S \cap (A \times B) = \Psi(A)$.

This is the parametrization we wanted.

Step 2. Assume that $\Psi : A \to \mathbb{R}^N$ is a function satisfying the assumptions stated in the result. in particular, we get that the $N \times d$ matrix

$$J\Psi(\bar{\lambda})$$

has rank d. Without loss of generality (namely, up to renaming the coordinates), we can assume that the first d columns are linearly independent. Write a point $x \in \mathbb{R}^N$ as

$$x = (x', x'') \in \mathbb{R}^d \times \mathbb{R}^{N-d},$$

and define $\alpha : A \to \mathbb{R}^d$ and $\beta : A \to \mathbb{R}^{N-d}$ by

$$\Psi(\lambda) = (\alpha(\lambda), \beta(\lambda)).$$

Namely, $\alpha(\lambda)$ and $\beta(\lambda)$ are the first d and the last $N - d$ coordinates of $\Psi(\lambda)$ respectively. Thanks to what we assumed above on the columns of $J\Psi(\bar{\lambda})$, we have that

$$\det J\alpha(\bar{\lambda}) \neq 0.$$

Thus, by using the Inverse Function Theorem (see Theorem 7.17), we have that α is locally a diffeomorphism. Set $U \coloneqq \alpha(\mathbb{R}^d)$, which is an open set. Define the function $f : U \to \mathbb{R}^N$ as

$$f(x) \coloneqq \beta(\alpha^{-1}(x)).$$

Then, f is a function of class C^1, and that

$$S \cap U = \text{graph}(f) \cap U.$$

By using Proposition 9.11, we get that S is a submanifold of dimension d at $\bar{x}$, as desired. □

Remark 9.15. The above result justifies the terminology 'a submanifold S *of dimension* d.'. Indeed, the dimension refers to the *linear* dimension, namely the number of coordinates we need in order to move on the set S.

Remark 9.16. Note, however, that when a set S is given in a parametric form, it is more difficult to check whether or not it is a submanifold, compared to when a set S is given as the set of solutions to a system of equations. Indeed, in the latter case we need to check (see Proposition 9.12) that the gradients of the constraints are linearly independent at a point. This is a *local* condition, and it requires the computation of $N - d$ gradients, and Linear Algebra. On the other hand, if the set S is given in a parametric form, other than the *local* condition of having the partial derivatives that are linear independent at a specific point, you need to check that the parametrizing function Ψ is invertible (in particular, injective), and the the inverse is continuous. To understand the difficulty, consider the case of the set S defined as

$$S \coloneqq \left\{ \left(16\left(\sin(\theta)\right)^3, 16\cos(\theta) - 5\cos(2\theta) - 2\cos(3\theta) - \cos(4\theta) \,:\, \theta \in (-\pi, \pi)\right) \right\}.$$

Try to prove that S is a submanifold at all of its points, except two singular points.

We now investigate the tangent space to a submanifold. There are two equivalent ways to compute the tangent space to a submanifold at a point, depending on the given local description of the submanifold. If this is given as the set of solutions to a system of equations, the tangent space is given by the vectors that are tangent to all gradients of the functions in the system.

Proposition 9.17. *Let $S \subset \mathbb{R}^N$ be a submanifold of dimension d at a point $\bar{x}$. Let $U \subset \mathbb{R}^N$ be an open neighborhood of $\bar{x}$, and write*

$$S \cap U = \{x \in A \,:\, \Phi_i(x) = 0, \text{ for all } i = 1, \dots, N - d\},$$

for some C^1 functions $\Phi_i : \mathbb{R}^N \to \mathbb{R}$. Then,

$$\text{Tan}(S, \bar{x}) = \{\, v \in \mathbb{R}^N \,:\, \langle \nabla \Phi_i(\bar{x}), v \rangle = 0, \text{ for all } i = 1, \dots, N - d \,\}.$$

In particular, $\text{Tan}(S, \bar{x})$ is a linear space of dimension d.

Remark 9.18. Note that the previous result is saying that a submanifold S of dimension d in $\mathbb{R}^N$ is locally the intersection of $N - d$ submanifolds S_i of class $N - 1$ (see Remark 9.7), and also the tangent space to S at a point $\bar{x}$ is the intersection of the tangent spaces at the S_i's at the point $\bar{x}$, and that it is a linear space itself.

Example 9.19. Consider the set

$$S \coloneqq \{(x, y, z) \in \mathbb{R}^3 \,:\, x^2 + y^2 + z^2 - 1 = 0\}.$$

Then, S is a submanifold of dimension $d = 2$ at each of its points. By using Proposition 9.17, we get that the tangent space to S at the point $(x, y, z) \in S$ is given by

$$\text{Tan}(S, (x, y, z)) = \{v = (v_x, v_y, v_z) \in \mathbb{R}^3 \,:\, 2xv_x + 2yv_y + 2zv_z = 0\}.$$

Remark 9.20. When it is difficult to explicitly compute the gradient of the functions Φ_i's, we can use an equivalent way to identify the tangent space to the submanifold at a point. We know that $\text{Tan}(S, \bar{x})$ is a d dimensional linear subspace of $\mathbb{R}^N$. To identify the vectors $v \in \mathbb{R}^N$ that belong to $\text{Tan}(S, \bar{x})$ we argue as follows: consider the curve $\gamma : [-1, 1] \to \mathbb{R}^N$ given by

$$\gamma(t) \coloneqq \bar{x} + tv.$$

Then, the above result states that $v \in \mathrm{Tan}(S, \bar{x})$ if and only if

$$\partial_v \Phi_i(\bar{x}) = 0,$$

for all $i = 1, \dots, N - d$. Namely, if the curve satisfies the system of equation at first order. Despite being equivalent, sometimes this point of view is more natural for certain situations, as the example below shows.

Example 9.21. Consider the case where S is the set of $N \times N$ matrices A with determinant 1. Thus, the set S is the zero level set of the function

$$\Phi_1(A) \coloneqq \det(A) - 1.$$

Then, it is difficult to compute the gradient of the function Φ_1. Nevertheless, we know (it was done in an exercise of the homework) that

$$\partial_B \Phi_1(A) = \sum_{i=1}^{N} \det(A^i B_i),$$

where the matrix $A^i B_i$ is the matrix A with its i^{th} column substituted by the i^{th} column of B.

If the submanifold is locally given as a parametrization, the tangent space is given by the directional derivatives of the parametrization function.

Proposition 9.22. *Let $S \subset \mathbb{R}^N$ be a submanifold of dimension d at a point $\bar{x}$. Let $U \subset \mathbb{R}^N$ is an open neighborhood of $\bar{x}$, and write*

$$S \cap U = \Psi(A),$$

where $A \subset \mathbb{R}^d$ is an open set, and $\Psi : A \to \mathbb{R}^N$ is a C^1 function such that:

(i) Ψ is an homeomorphism,

(ii) $\partial_1 \Psi(\bar{\lambda}), \dots, \partial_d \Psi(\bar{\lambda})$ are linearly independent, where $\Psi(\bar{\lambda}) = \bar{x}$.

Then,

$$\mathrm{Tan}(S, \bar{x}) = \left\{ \partial_v \Psi(\bar{\lambda}) : v \in \mathbb{R}^d \right\}.$$

In particular, $\mathrm{Tan}(S, \bar{x})$ is a linear space of dimension d.

The proof is left as an exercise to the reader.

Example 9.23. Consider the set

$$S \coloneqq \{(\sin\theta\cos\varphi, \sin\theta\sin\varphi, \cos\theta) \in \mathbb{R}^3 \ : \ \theta, \varphi \in (0, \pi/2)\}.$$

Then, the function $\Psi : (0, \pi/2)^2 \to \mathbb{R}^3$ given by

$$\Psi(\theta, \varphi) \coloneqq (\sin\theta\cos\varphi, \sin\theta\sin\varphi, \cos\theta)$$

satisfies the assumption of Proposition 9.14; thus, we get that S is a submanifold of dimension $d = 2$ at each of its points. Moreover, by using Proposition 9.22, we get that the tangent space to S at a point $(x, y, z) = \Psi(\bar{\theta}, \bar{\varphi}) \in S$ is given by

$$\mathrm{Tan}(S, (x, y, z)) = \{\partial_v \Psi(\bar{\theta}, \bar{\varphi}) \ : \ v = (v_x, v_y) \in \mathbb{R}^2\}.$$

We would like to write in a more explicit form the right-hand side. Since

$$\partial_v \Psi(\bar{\theta}, \bar{\varphi}) = J\Psi(\bar{\theta}, \bar{\varphi}) \cdot v^T,$$

and

$$J\Psi(\bar{\theta}, \bar{\varphi}) = \begin{pmatrix} \cos\bar{\theta}\cos\bar{\varphi} & \cos\bar{\theta}\cos\bar{\varphi} \\ -\sin\bar{\theta}\sin\bar{\varphi} & \sin\bar{\theta}\cos\bar{\varphi} \\ -\sin\bar{\theta} & 0 \end{pmatrix}$$

we get that $\mathrm{Tan}(S, (x, y, z))$ is given by

$$\left\{ (v_x \cos\bar{\theta}\cos\bar{\varphi} + v_y \cos\bar{\theta}\cos\bar{\varphi}, -v_x \sin\bar{\theta}\sin\bar{\varphi} + v_y \sin\bar{\theta}\cos\bar{\varphi}, -v_x \sin\bar{\theta}) \ : \ v_x, v_y \in \mathbb{R} \right\}.$$

Remark 9.24. There are two other equivalent ways to compute the tangent space to a submanifold at a point, depending on the given local description of the submanifold. We recall that the tangent space to a submanifold of dimension k is a linear space of dimension d.

Assume the submanifold to be locally described by *linearization*, namely by using the diffeomorphism given by Proposition 9.12. Namely, if S is a d-dimensional submanifold of $\mathbb{R}^N$ at the point $\bar{x}$, and $U \subset \mathbb{R}^N$ is an open set with $\bar{x} \in U$, and $\varphi : \mathbb{R}^N \to \mathbb{R}^N$ is a C^1 diffeomorphism such that

$$\varphi(S \cap U) = A \times \{0\},$$

where $A \subset \mathbb{R}^d$ is an open set. Then,

$$\mathrm{Tan}(S, \bar{x}) = (\,\mathrm{d}\varphi(\bar{x}))^{-1}\,[\mathbb{R}^d \times \{0\}].$$

Now, assume the submanifold to be locally parametrized by a function $\Psi : \mathbb{R}^d \to \mathbb{R}^N$ satisfying the assumptions in Proposition 9.14. Then, in order to compute $\partial_v \Psi(\bar{\lambda})$, where $v \in \mathbb{R}^d$, we can take a C^1 curve $\gamma : [-1, 1] \to \mathbb{R}^d$ such that

$$\gamma(0) = \bar{\lambda}, \qquad \gamma'(0) = v.$$

Then, the curve $\Psi \circ \gamma : [-1, 1] \to S$ is of class C^1 and, by using the chain rule (see Proposition 5.39), we get that

$$\partial_v \Psi(\bar{\lambda}) = (\Psi(\gamma))'\,(0).$$

Thus, if we take $v_1, \dots, v_d \in \mathbb{R}^d$ that are linearly independent, and C^1 curves $\gamma_i : [-1, 1] \to \mathbb{R}^d$, for $i = 1, \dots, d$, such that

$$\gamma_i(0) = \bar{\lambda}, \qquad \gamma_i'(0) = v_i,$$

we get that $\mathrm{Tan}(S, \bar{x})$ is the linear space generated by the vectors $(\Psi(\gamma_1))'\,(0), \dots, (\Psi(\gamma_d))'\,(0)$.

Finally, we study submanifold in $\mathbb{R}^N$ of dimensions $N - 1$, and their relation to graphs of C^1 functions, both over the coordinate axes, and over the tangent space to the submanifold.

What the Implicit Function Theorem yields, is that also the opposite is true: submanifold of $\mathbb{R}^N$ of dimension $N - 1$ are locally the graph of a function over the coordinate axes.

Proposition 9.25. *Let $S \subset \mathbb{R}^N$ be a submanifold of $\mathbb{R}^N$ of dimension $N-1$ at the point $\bar{x} \in S$. Then, there exist an open set $U \subset \mathbb{R}^N$ with $\bar{x} \in U$, and (up to reshuffling the coordinates) a function $f : \mathbb{R}^{N-1} \to \mathbb{R}$ such that*

$$S \cap U = \{(x', f(x')) \,:\, x' \in A\},$$

for some open set $A \subset \mathbb{R}^{N-1}$.

Remark 9.26. Note that a set described as the graph of a function $f : \mathbb{R}^N \to \mathbb{R}^M$ is a special case of a parametrization $\Psi : \mathbb{R}^N \to \mathbb{R}^{N+M}$ given by

$$\Psi(x) \coloneqq (x, f(x)),$$

where the first N coordinates are precisely the parameters. On the other hand, a set described in a parametric way is not always the graph of a function, as the example below shows.

We now prove a modified version of the above results, that states that, locally, $(N-1)$-dimensional submanifolds in $\mathbb{R}^N$ can be seen as graphs over their tangent hyperplane. This is useful when dealing with local properties of a manifold.

Proposition 9.27. *Let $S \subset \mathbb{R}^N$ be a submanifold of dimension $N-1$ at the point $\bar{x}$. Let $\nu \in \mathbb{R}^N$ be a unit normal vector to $\mathrm{Tan}(S, \bar{x})$. Then, there exists an open set $U \subset \mathbb{R}^N$ containing $\bar{x}$, and a C^1 function $h : \bar{x} + \mathrm{Tan}(S, \bar{x}) \to \mathbb{R}$ such that*

$$S \cap U = \{\, \pi(x) + h(\pi(x))\nu \,:\, x \in S \cap U \,\},$$

where $\pi : \mathbb{R}^N \to (\bar{x} + \mathrm{Tan}(S, \bar{x}))$ is the orthogonal projection on $\bar{x} + \mathrm{Tan}(S, \bar{x})$.

For lower dimensional submanifold, it is in general not possible to write them as graphs over their tangent space. The reason being that there are too many directions that are orthogonal to the tangent space. Think, for example, to the one-dimensional submanifold given by

$$S := \{(\cos t, \sin t, t) \,:\, t \in (0,1)\}.$$

Nevertheless, it is possible to prove that, locally, the submanifold intersects a normal direction in just one point.

Proposition 9.28. *Let $S \subset \mathbb{R}^N$ be a submanifold of dimension d at the point $\bar{x}$. Let $V \subset \mathbb{R}^N$ be a linear space such that*

$$V \cap \mathrm{Tan}(S, \bar{x}) = \{0\}.$$

Then, that there exists an open set $A \subset \mathbb{R}^N$ containing $\bar{x}$ such that

$$x - y \notin V,$$

for all $x \neq y \in S \cap A$.

The proof is left as an exercise to the reader.

9.2. **Critical points on submanifolds: Lagrange multipliers.** We now want to consider the minimization (or the maximization) of a function over a submanifold. Namely, we are given a C^1 function $f : \mathbb{R}^N \to \mathbb{R}$, and a d-dimensional submanifold $S \subset \mathbb{R}^N$, and we want to study the problem

$$\min\{f(x) \,:\, x \in S\}. \tag{9.1}$$

This is called *constrained optimization*, since the set S plays the role of the constraints we impose on our variables $x \in \mathbb{R}^N$. What we would like to do is to derive some *first order necessary conditions* for minimum points of f over S. First, consider the unconstrained minimization problem, namely if S was the entire $\mathbb{R}^N$. Assume that there exists a point $\bar{x} \in \mathbb{R}^N$ of minimum. Namely,

$$f(\bar{x}) \leq f(x), \tag{9.2}$$

for all $x \in \mathbb{R}^N$. Then, you know that

$$\nabla f(\bar{x}) = 0. \tag{9.3}$$

The above condition is extremely useful in order to identify the point $\bar{x}$, because it turns condition (9.2), which is an inequality to be tested over *all* points of $\mathbb{R}^N$, into an equation, namely (9.3). Of course, you know that (9.3) is only *necessary*, but not *sufficient* for minimality, in that a solution of (9.3) is not necessarily a minimum of f. Indeed, think about the function $f : \mathbb{R} \to \mathbb{R}$ given by $f(x) := x^3$.

We now want to derive a similar condition in the case of the constrained minimization problem (9.1). We will indeed derive a condition that works for *general sets* S, not necessarily submanifolds, and then we'll deduce a stronger condition in the case S is a submanifold. The idea is simple, and comes from the way condition (9.3) is derived: if $\bar{x} \in S$ is a point of local minimum, then if I move along a tangent direction to S at $\bar{x}$, the function f cannot decrease. Namely, take $v \in \mathrm{Tan}(S, \bar{x})$; let $(x_n)_{n\in\mathbb{N}} \subset S$, and $(\lambda_n)_{n\in\mathbb{N}} \subset (0,1)$ be such that $\lambda_n \to 0$ and

$$\frac{x_n - \bar{x}}{\lambda_n} \to v.$$

Then, since

$$f(\bar{x}) \leq f(x_n),$$

and $\lambda_n > 0$, we get

$$0 \leq \frac{f(x_n) - f(\bar{x})}{\lambda_n}.$$

Thanks to Theorem 5.23, we get that the right-hand side converges to $\partial_v f(\bar{x})$. Note that we did not use any structure of S in order to deduce the above condition. Namely, we get the following result

Lemma 9.29. *Let $f : \mathbb{R}^N \to \mathbb{R}$ be of class C^1, and let $S \subset \mathbb{R}^N$ be a set. Assume that $\bar{x} \in S$ is a point of (local) minimum of f over S. Then,*

$$\langle \nabla f(\bar{x}), v \rangle \geq 0,$$

for all $v \in \mathrm{Tan}(S, \bar{x})$.

This is the mathematical writing of the sentence 'if I move along a tangent direction to S at $\bar{x}$, the function f cannot decrease', that we wrote above.

Remark 9.30. Note that Lemma 9.29 allows us to deduce condition (9.3). Indeed, since the problem is unconstrained, namely $S = \mathbb{R}^N$, we have that $\mathrm{Tan}(S, \bar{x}) = \mathbb{R}^N$. Thus,

$$\langle \nabla f(\bar{x}), v \rangle \geq 0,$$

for all $v \in \mathbb{R}^N$. This implies that $\nabla f(\bar{x}) = 0$ (prove it!).

Let us now consider the case of a d-dimensional submanifold $S \subset \mathbb{R}^N$. The first order necessary condition that we obtain is called *Lagrange multipliers rule*.

Theorem 9.31 (Lagrange multipliers)**.** *Let $f : \mathbb{R}^N \to \mathbb{R}$, and let $S \subset \mathbb{R}^N$ be a d-dimensional submanifold. Write*

$$S = \left\{ x \in \mathbb{R}^N \,:\, \Phi_1(x) = \cdots = \Phi_{N-d}(x) = 0 \right\},$$

for some C^1 functions $\Phi_i : \mathbb{R}^N \to \mathbb{R}$, for $i = 1, \dots, N-d$. Assume that $\bar{x} \in S$ is a point of (local) minimum of f over S. Then,

$$\nabla f(\bar{x}) = \sum_{i=1}^{N-d} \lambda_i \nabla \Phi_i(\bar{x}),$$

for some coefficients $\lambda_1, \dots, \lambda_{N-d} \in \mathbb{R}$.

Proof. By using Lemma 9.29, we have that

$$\langle \nabla f(\bar{x}), v \rangle \geq 0, \tag{9.4}$$

for all $v \in \mathrm{Tan}(S, \bar{x})$. Now, since $\mathrm{Tan}(S, \bar{x})$ is a linear space, we have that $v \in \mathrm{Tan}(S, \bar{x})$ satisfies (9.4), then also $-v$ does. Thus, from (9.4) we get that

$$\langle \nabla f(\bar{x}), v \rangle = 0,$$

for all $v \in \mathrm{Tan}(S, \bar{x})$. This means that $\nabla f(\bar{x})$ lies in the orthogonal space to $\mathrm{Tan}(S, \bar{x})$. We now want to describe such orthogonal space. By using Proposition 9.17 we know that

$$\mathrm{Tan}(S, \bar{x}) = \{ v \in \mathbb{R}^N \,:\, \langle \nabla \Phi_i(\bar{x}), v \rangle = 0, \text{ for all } i = 1, \dots, N-d \}.$$

Therefore, the orthogonal space to $\mathrm{Tan}(S, \bar{x})$ is generated by the vectors $\nabla \Phi_1(\bar{x}), \dots, \nabla \Phi_{N-d}(\bar{x})$. This concludes the proof of the theorem. □

Remark 9.32. Note that above result does *not* provide the existence of a minimum point! This is something that you have to prove separately (typically, by using Weierstraß Theorem 3.15 and the compactness of the set S, or of the sub-level sets of the function f).

Remark 9.33. Of course, the result holds also in the case where the submanifold S is locally described by a set of equations. The choice of writing the result as above is only for the sake of notation.

Remark 9.34. The same result holds also for a point $\bar{x} \in S$ of (local) maximum of f over S.

Example 9.35. Let us consider the function $f : \mathbb{R}^3 \to \mathbb{R}$ given by $f(x, y, z) \coloneqq x^2 - 2y + z^2$, and let S be the unit sphere centered at the origin, namely

$$S \coloneqq \{(x, y, z) \in \mathbb{R}^3 \,:\, \Phi(x, y, z) = 0\},$$

where $\Phi(x, y, z) \coloneqq x^2 + y^2 + z^2 - 1$. Then, we know that (at least) a minimum point of f over S exists, since f is continuous, and S is compact. We would like to find such points (note that

we do not know whether there is one or more than one). By using Theorem 9.31, we know that minimum points are solutions to the equation

$$\nabla f(x,y,z) = \lambda \nabla \Phi(x,y,z).$$

This condition writes explicitly as

$$(2x, -2, 2z) = \lambda(2x, 2y, 2z), \tag{9.5}$$

which is a system of three equations in four unknowns (x, y, z, and λ). Actually, we have a fourth equation, given by the fact that the point (x, y, z) has to be in S, namely it has to satisfy $\Phi(x, y, z) = 0$. Thus, we have to solve the system

$$\begin{cases} 2x = 2\lambda x \\ -2 = 2\lambda y \\ 2z = 2\lambda z \\ x^2 + y^2 + z^2 - 1 = 0. \end{cases} \tag{9.6}$$

First of all, we note that (9.5) implies that λ cannot be zero. Therefore, from the second equation in (9.6) we get that

$$y = -\frac{1}{\lambda}. \tag{9.7}$$

If both x and z are zero, from the last equation in (9.6) we get that $y = \pm 1$, and thus, from (9.7) we obtain two solutions

$$P_1 = (0, 1, 0), \qquad \lambda = -1, \tag{9.8}$$

$$P_2 = (0, -1, 0), \qquad \lambda = 1, \tag{9.9}$$

On the other hand, if either x or z are not zero, from either the first or the third equation in (9.6) we get that $\lambda = 1$, and thus $y = -1$. From the fourth equation in (9.6) this would imply that both x and z are zero, which is a contradiction. Thus, the two solutions to the system (9.6) are given by (9.8) and (9.9).

Since we know that there exists at least a minimum point, it has to be one of the two above. Note that, since we also know that the function f has a maximum over S, and that also for points of (local) maximum the Lagrange multipliers conditions is valid, we have that there is only a point a minimum of f over S, and only one point of maximum of f over S, and they are $(0, 1, 0)$ and $(0, -1, 0)$. To tell which is which, we just compute f at those two points, and compare the values. We have that

$$f(P_1) = -2, \qquad f(P_2) = 2.$$

Thus, P_1 is the point of minimum of f over S, and P_2 is the point of maximum of f over S.

Remark 9.36. In previous example, you could have started using another method to find the minimum of f on S. Since $x^2 + y^2 + z^2 = 1$, we know that $x^2 + z^2 = 1 - y^2$. By substituting this in the expression of f, we get the function

$$g(y) := 1 - y^2 - 2y.$$

One could think that by studying the minimum of such function. The problem is that this function is *unbounded*, and thus the infimum is $-\infty$!!! But we know that f has a finite minimum over S. The power of Theorem 9.31 lies in giving a condition also in cases where the above substitution, even if in theory possible, cannot be explicitly written, or, even if it can be explicitly written, gives rise to an unbounded function.

Remark 9.37. Note that the system that we have to solve when looking for critical points of a C^1 function on a d-dimensional submanifold, is a system of $2N - d$ equations (N given by the

first condition, and $N-d$ by the vanishing of the constraints Φ_i's) in $2N-d$ unknowns (N are the coordinates of the gradient of f, and $N-d$ the coefficients λ_i's), which writes as

$$\begin{cases} \nabla f(\bar{x}) = \sum_{i=1}^{N-d} \lambda_i \nabla \Phi_i(\bar{x}) \\ \Phi_1(\bar{x}) = 0 \\ \vdots \\ \Phi_{N-d}(\bar{x}) = 0, \end{cases}$$

where the first condition is actually N equations. Note that, in the case S is compact, the above system always have at least *two* solutions: one identifying a point of maximum of f over S, and one identifying a point of minimum of f over S.

Remark 9.38. We now would like to give a geometric interpretation of the Lagrange multipliers. The gradient $\nabla f(x)$ of a function $f : \mathbb{R}^N \to \mathbb{R}$ at a point $x \in \mathbb{R}^N$, if not null, points at the direction of maximal increase of the function f. Indeed, among all unit vectors $v \in \mathbb{S}^{N-1}$,we have that

$$\partial_v f(x) = \langle \nabla f(x), v \rangle \leq |\nabla f(x)|,$$

where the last inequality follows from the Cauchy-Schwarz inequality (see Proposition 1.9). By taking

$$w := \frac{\nabla f(x)}{|\nabla f(x)|},$$

we have that that inequality becomes an equality. Thus, the direction w is the one that maximizes the growth of f at the point x. In particular, if we consider a direction $v \in \mathbb{R}^N$ such that

$$\langle v, \nabla f(x) \rangle \geq 0,$$

we have that f is locally increasing if we move along the direction v. Similarly, we can say that if we consider a direction $v \in \mathbb{R}^N$ such that

$$\langle v, \nabla f(x) \rangle \leq 0,$$

we have that f is locally decreasing if we move along the direction v. Note that, if f is locally increasing in a direction v, then f is locally decreasing in the direction $-v$, and viceversa.

Now, let us consider a submanifold S, locally described by one C^1 equation

$$S = \left\{ x \in \mathbb{R}^N \, : \, \Phi_1(x) = 0 \right\}.$$

Namely, S is the zero level set of the function Φ_1. Let $\bar{x} \in S$ be a point of minimum for f over S. We know from Proposition 9.17 that

$$\mathrm{Tan}(S, \bar{x}) = \{ v \in \mathbb{R}^N \, : \, \langle \nabla \Phi_1(\bar{x}), v \rangle = 0 \}.$$

at all points $\bar{x} \in S$. Thus, locally - and up to a translation -, the set S looks like the linear set of vectors that are orthogonal to $\nabla \Phi_1(\bar{x})$'s. Therefore, if $\nabla \Phi_1(\bar{x})$ and $\nabla f(\bar{x})$ were not linearly dependent, then we could find a vector $v \in \mathrm{Tan}(S, \bar{x})$ such that

$$\langle v, \nabla f(\bar{x}) \rangle < 0.$$

Thus, f would locally decrease if we move along a direction v; this means that we could find points $x \in S$ close to $\bar{x}$ with

$$f(x) < f(\bar{x}),$$

contradicting the minimality of the point $\bar{x}$ for f over S.

The same argument gives the similar geometric interpretation also in the case where S is locally described by the intersection of the zero level sets of $N-d$ equations.

10. The Darboux-Riemann integral and the Peano-Jordan content

What is the Riemann integral? What type of functions are Riemann integrable? For what functions do the Fundamental Theorem of Calculus hold? These are the questions that troubled mathematicians for about a century, before being completely understood at the beginning of the 20th century. It is a story of struggles, fights, and failed attempts that eventually led to the beautiful and powerful theory of Lebesgue integration that is used nowadays. But let's start from the beginning.

10.1. **The problems of the Cauchy-Riemann integration.** The development of Calculus led to the notion of the derivative of a function. This allowed to consider differential equations that model physical phenomena, and to study such situations by using mathematical tools. This was one of the most important revolutions for human kind. At that time, though, things were not really clear. First of all, the notion of *function* was really vague. Every mathematician had a different notion, that was not clearly stated, if not by sentences like '*a function is given by a single equation*'. By this, they meant that it is possible to write an explicit expression of the function by using a single expression. For instance,

$$f(x) \coloneqq x^2 + 1$$

was a (continuous) function, while

$$f(x) \coloneqq \begin{cases} x & \text{if } x \leq 0, \\ x^2 & \text{if } x > 0, \end{cases}$$

was either not a function, or considered as a discontinuous function (note that, by using the modern notion of continuity, such function is continuous). In particular, this meant that a very limited class of functions were those considered by mathematicians at that time. Thus, when in 1829 Dirichlet introduced the function

$$f(x) \coloneqq \begin{cases} 1 & \text{if } x \in \mathbb{Q}, \\ 0 & \text{if } x \in \mathbb{R} \setminus \mathbb{Q}, \end{cases} \tag{10.1}$$

mathematicians were really puzzled of whether to consider this strange object a function or not. Indeed, it is not possible to write it as *a single equation*. Moreover, the Dirichlet function is discontinuous *at every point*! Is it then an object worth considering or not? This problem is intimately connected with the *representation of functions*: how is it possible to write a generic function? This was one of the main problems of interest of the 19th century after Joseph Fourier published his *Théorie analytique de la chaleur* (Analytic theory of heat) in 1822, where he established a Partial Differential Equation governing heat diffusion and used *infinite* series of trigonometric functions to solve it. Are these latter functions? Are the computations used by Fourier (or, later, by Lagrange), where the equality

$$\int \sum_{n\in\mathbb{N}} f_n(x)\,dx = \sum_{n\in\mathbb{N}} \int f_n(x)\,dx$$

is repeatedly used, justified?

To answer these questions, we first have to understand what the meaning of

$$\int f(x)\,dx$$

is. At that time, it was meant as the *antiderivative* of f. Namely, it was the function F such that $F' = f$ (this is the Fundamental Theorem of Calculus). How to construct such a function? Well, first of all, assume that it exists! We follow the work of Augustin Cauchy of his *Course d'analyse* (Course of Analysis) of 1821 (culmination of his efforts from 1814). Let us take an interval $[a, b] \subset \mathbb{R}$, and a *partition*

$$a = x_0 < x_1 < \cdots < x_N = b$$

of it. Then, we can say that

$$f(x_i) = F'(x_i) \sim \frac{F(x_i) - F(x_{i-1})}{x_i - x_{i-1}},$$

from which we deduce that

$$F(x_i) \sim F(x_{i-1}) + f(x_i)(x_i - x_{i-1}).$$

By using the same argument to write $F(x_{i-1})$, and thus $F(x_{i-2}), \dots F(x_0)$, we get that

$$F(x_i) \sim \sum_{j=1}^{i} F(x_{j-1}) + \sum_{j=1}^{i} f(x_j)(x_j - x_{j-1}).$$

Now, the first sum is just a constant, and we will call it C. Thus, up to a constant, we have that

$$F(x_i) \sim \sum_{j=1}^{i} f(x_j)(x_j - x_{j-1}). \tag{10.2}$$

Note that, by using the fact that F' is continuous, since f is, it is possible to use the exact writing

$$f(x_i) = F'(\xi_i) = \frac{F(x_i) - F(x_{i-1})}{x_i - x_{i-1}},$$

for some $\xi_i \in (x_{i-1}, x_i)$. The expression that we then obtain for F is

$$F(x_i) = \sum_{j=1}^{i} f(\xi_j)(x_j - x_{j-1}), \tag{10.3}$$

for $\xi_i \in (x_{i-1}, x_i)$. The sum on the right-hand side of either (10.2) or (10.3) is called *Cauchy-Riemann sum*.

Definition 10.1. A function $f : [a, b] \to \mathbb{R}$ is said to be *Cauchy-Riemann integrable* if the Cauchy-Riemann sums converge as the maximum length of a subinterval (x_{i-1}, x_i) goes to zero.

Cauchy proved that, for a *continuous* function f, the Cauchy-Riemann sum converges. This was a revolutionary proof at that time, since it used the *refinement of a partition*, that you have seen in Analysis 1. Moreover, he proved that $F' = f$, and that, if G is a function such that $G' = f$, then $G = C + F$, for some constant $C \in \mathbb{R}$. Finally, he proved conditions for the term by term integration

$$\int_{[a,b]} \sum_{n\in\mathbb{N}} f_n(x)\, dx = \sum_{n\in\mathbb{N}} \int_{[a,b]} f_n(x)\, dx$$

to hold. Note that all the functions involved have to be continuous: not only each f_n, but also the function $\sum_{n\in\mathbb{N}} f_n$. The theory developed by Cauchy is great, but not powerful enough to treat general discontinuous functions.

Riemann, in his habilitation thesis of 1854 studied the problem of representation of a function by trigonometric functions, and investigated conditions under which the Cauchy-Riemann sum converge, without having to assume f continuous. We will state the result in the following section (see Theorem 10.43), since we first need to introduce a new notion.

Riemann showed an example of a discontinuous function that is Riemann integrable. For $x \in \mathbb{R}$, let $I(x) \in \mathbb{Z}$ be the closest integer to x. Consider the function

$$g(x) \coloneqq \begin{cases} x - I(x) & \text{if } x \neq n/2, \text{ with } n \text{ odd}, \\ 0 & \text{else.} \end{cases}$$

Define

$$f(x) \coloneqq \sum_{n\geq 1} \frac{g(nx)}{n^2}.$$

Then, it is possible to show that f is well defined (namely, the series converges[12] for all $x \in \mathbb{R}$). Moreover, f is discontinuous at all points $x \in \mathbb{R}$ of the form

$$x = \frac{m}{2n},$$

with $m, n \in \mathbb{Z}$ co-prime (namely, with no common divisor, other than ± 1). Finally, this function is Riemann integrable!

Another example of a function that is discontinuous at countably many points, but still Riemann integrable was given by Thomae in 1875, as a modification of the Dirichlet function (10.1): let $f : [0,1] \to \mathbb{R}$ be defined as

$$f(x) := \begin{cases} \dfrac{1}{q} & \text{if } x = \dfrac{p}{q} \text{ with } p, q \text{ co-prime,} \\ \\ 0 & \text{else.} \end{cases} \tag{10.4}$$

Thus, there are functions with *countably* many discontinuities that are Riemann integrable.

This looks promising for the Riemann integral. Nevertheless, there are still some drawbacks. First of all, the function (10.1) is *not* Riemann integrable. This is a problem, because we heuristically expect the following. Let $\{q_n\}_{n \geq 1}$ be an enumeration of the rational numbers in $[0,1]$. Consider the function $g_0 \equiv 0$. Then, g_0 is clearly Riemann integrable, and

$$\int_{[0,1]} g_0(x)\, dx = 0.$$

Then, consider the function

$$g_1(x) := \begin{cases} 1 & \text{if } x = q_1, \\ 0 & \text{else.} \end{cases}$$

Then, g_1 is clearly Riemann integrable, and

$$\int_{[0,1]} g_1(x)\, dx = 0.$$

We now define g_{n+1} as follows:

$$g_{n+1}(x) := \begin{cases} 1 & \text{if } x = q_{n+1}, \\ g_n(x) & \text{else.} \end{cases}$$

Namely, g_{n+1} has just one more point where it is one, with respect to g_n. Then, g_{n+1} is clearly Riemann integrable, and

$$\int_{[0,1]} g_{n+1}(x)\, dx = 0.$$

The function (10.1) introduced by Dirichlet is just the limit of these functions g_n. Why would it not be Riemann integrable? What fails? Well, what happens is that the only result for Riemann integrability of a sequence of functions is the following:

Theorem 10.2. *Let $(f_n)_{n \in \mathbb{N}}$ be a sequence of functions $f_n : [a,b] \to \mathbb{R}$ which is Riemann integrable. Assume that $f_n \to f$ uniformly. Then, f is Riemann integrable, and*

$$\lim_{n \to \infty} \int_{[a,b]} f_n(x)\, dx = \int_{[a,b]} f(x)\, dx.$$

Remark 10.3. What makes the above result weak is the requirement of *uniform* convergence. Indeed, consider the sequence of functions $f_n : (0,1) \to \mathbb{R}$ defined as

$$f_n(x) := x^n.$$

[12] Indeed, the n^{th} term of the series is bounded by $1/n^2$.

Then, $f_n \to f$ pointwise, where $f \equiv 0$. We have that f is Riemann integrable, and that

$$\lim_{n\to\infty} \int_{(0,1)} f_n(x)\,dx = \int_{(0,1)} f(x)\,dx.$$

Unfortunately, we *cannot* infer the Riemann integrability of f, neither the limit above from the previous theorem, since the convergence of the f_n's is not uniform.

Remark 10.4. Also the function (10.4) can be obtained as pointwise limit of a sequence of Riemann integrable functions as we did above for the function (10.1). Also in this case, its Riemann integrability cannot be deduced from the theorem above.

This is the same issue in justifying rigorously the

$$\int_{[a,b]} \sum_{n\in\mathbb{N}} f_n(x)\,dx = \sum_{n\in\mathbb{N}} \int_{[a,b]} f_n(x)\,dx$$

used in Fourier series. Indeed,

$$\sum_{n\in\mathbb{N}} f_n(x) = \lim_{k\to\infty} g_k \coloneqq \sum_{n=1}^{k} f_n.$$

Thus, if the series converges *uniformly*, we have that the series is Riemann integrable and that it is possible to exchange the series and the integral. Unfortunately, uniform convergence is not always (basically, never!) true.

10.2. **The Peano-Jordan content.** The first breakthrough in developing a more powerful theory of integration was to see the integral of a function $f : \mathbb{R}^N \to \mathbb{R}$ as the *signed area/volume under the graph*. This, of course, requires the notion of *area/volume of a set*. We will see that it is precisely in the definition of the area used that all the issues of the Riemann integral are contained, and it is precisely by modifying the idea of area that the powerful Lebesgue integration was born.

Let us start with a consideration. Whatever reasonable notion of area/volume[13] we assign to a set $E \subset \mathbb{R}^N$, the following has to be true:

$$\begin{gathered}\text{The volume of a rectangle } R \coloneqq [a_1, b_1] \times \cdots \times [a_N, b_N] \\ \text{is } (b_1 - a_1)(b_2 - a_2)\dots(b_N - a_N).\end{gathered}$$

The question is how to *extend* the notion of volume to the case of a general set E. It is in the way the extension is done, that two important notions were developed: the Peano-Jordan content, and the Lebesgue measure. In this section we will discuss the former, while the latter in the following section.

What Peano and Jordan did (in 1887 and 1892, respectively, and independently) is to use an idea similar to the *exhaustion method* of Archimedes: approximate a set from inside and from outside with union of rectangles. If the two processes lead to the same number, that will be the number assigned to the set. There are cases, though, where the two processes lead to different numbers. In these cases, we won't be able to assign a number to the set.

Definition 10.5. A set $R \subset \mathbb{R}^N$ is called a *pluri-rectangle* if it is possible to write it as

$$R = \bigcup_{i=1}^{k} R^i,$$

where each $R^i \subset \mathbb{R}^N$ is a rectangle.

Remark 10.6. Note that we are using *closed* rectangles.

[13] We will work in an arbitrary dimension, while the terminology *area* and *volume* refer to the two and three dimensional case, respectively.

Remark 10.7. Note that the writing of a pluri-rectangle is not unique. Nevertheless, it is always possible to write (again, not in a unique way!) a pluri-rectangle as a finite union of rectangles with *pairwise disjoint* interiors.

Remark 10.8. Note that a pluri-rectangle can have multiple connected components.

Simple (but extremely tedious to prove!) properties of pluri-rectangle are the followings

Lemma 10.9. *Let $R_1, \dots, R_m \subset \mathbb{R}^N$ be pluri-rectangles. Then,*

$$\bigcup_{i=1}^m R_i \qquad \bigcap_{i=1}^m R_i$$

are also pluri-rectangles.

Remark 10.10. Note that *countable* union of pluri-rectangles might fail to be a pluri-rectangle. Indeed, every open set of $\mathbb{R}^N$ can be written as countable union of rectangles.

Definition 10.11. Let $R \subset \mathbb{R}^N$ be a rectangle

$$R = [a_1, b_1] \times \cdots \times [a_N, b_N],$$

with $0 < a_i < b_i$ for all $i = 1, \dots, N$. We define the *Peano-Jordan content* of R as

$$\mathcal{PJ}(R) \coloneqq (b_1 - a_1)(b_2 - a_2) \dots (b_N - a_N),$$

Definition 10.12. Let $R \subset \mathbb{R}^N$ be a pluri-rectangle. Write R as

$$R = \bigcup_{i=1}^k R^i$$

where the R^i's have pairwise disjoint interiors. We define the *Peano-Jordan content* of R as

$$\mathcal{PJ}(R) \coloneqq \sum_{i=1}^k \mathcal{PJ}(R_i).$$

Remark 10.13. It is easy to check that the above is a well-defined number. Namely, it does not depend on the way we write the pluri-rectangle R as the union of disjoint rectangles.

It is now time to define how to extend the Peano-Jordan content to more general sets.

Definition 10.14. Let $E \subset \mathbb{R}^N$. We define the *inner Peano-Jordan content* of E as

$$\mathcal{PJ}^-(E) \coloneqq \sup \{ \mathcal{PJ}(R) \,:\, R \subset E, R \text{ pluri-rectangle} \},$$

and the *outer Peano-Jordan content* of E as

$$\mathcal{PJ}^+(E) \coloneqq \inf \{ \mathcal{PJ}(R) \,:\, E \subset R, R \text{ pluri-rectangle} \}.$$

If

$$\mathcal{PJ}^-(E) = \mathcal{PJ}^+(E),$$

we say that E is *Peano-Jordan measurable*, and we denote the common value by $\mathcal{PJ}(E)$.

Remark 10.15. First of all, we note that the definition of the Peano-Jordan content given in Definition 10.14 is consistent with those given in Definitions 10.11 and 10.12 for rectangles and pluri-rectangles respectively. Indeed, if $R \subset \mathbb{R}^N$ is a rectangle, it holds that

$$\mathcal{PJ}(R) = \sup\{\mathcal{PJ}(S) \,:\, S \subset R, S \text{ pluri-rectangle}\} = \inf \{ \mathcal{PJ}(S) \,:\, R \subset S, S \text{ pluri-rectangle} \}.$$

The same is true in the case R is a pluri-rectangle.

It is possible to see that we can equivalently define the Peano-Jordan content by using grids. The idea is to consider finer and finer grids of cubes, and, at each scale, to identify the inner and the outer Peano-Jordan content of a set by counting how many cubes are inside and how many cubes intersect the set, respectively. We make this more precise.

Definition 10.16. Let $r > 0$. We define the *r-grid* $\mathcal{G}(r)$ as

$$\mathcal{G}(r) \coloneqq \{\, Q(z_i, r) \,:\, z_i \in r\mathbb{Z} \,\},$$

where, for $x \in \mathbb{R}^N$, and $r > 0$,

$$Q(x, r) \coloneqq \{y \in \mathbb{R}^N \,:\, |x_i - y_i| < r/2 \text{ for all } i = 1, \dots, N\}$$

denotes the open cube centered at x with sides of length r parallel to the orthogonal axes.

Proposition 10.17. *Let $E \subset \mathbb{R}^N$. It holds that*

$$\mathcal{PJ}^-(E) = \sup_{r>0} r^N \#\{\, i \in \mathbb{N} \,:\, Q(z_i, r) \subset E,\, Q(z_i, r) \in \mathcal{G}(r) \,\},$$

and

$$\mathcal{PJ}^+(E) = \inf_{r>0} r^N \#\{\, i \in \mathbb{N} \,:\, Q(z_i, r) \cap E \neq \emptyset,\, Q(z_i, r) \in \mathcal{G}(r) \,\}$$

The proof is left as an exercise to the reader.

We now state some basic properties of Peano-Jordan measurable sets. The proofs are left as exercises to the reader.

Lemma 10.18. *Let $E \subset \mathbb{R}^N$. Then,*

(i) $\mathcal{PJ}^-(E) = 0$ if and only if E has empty interior;
(ii) Assume that $\mathcal{PJ}^+(E) < \infty$. Then, E is bounded;
(iii) $\mathcal{PJ}^+(E) = \mathcal{PJ}^+(\overline{E})$.

The Peano-Jordan content is finitely additive on pairwise disjoint sets.

Lemma 10.19. *Let $E_1, \dots, E_m \subset \mathbb{R}^N$ be Peano-Jordan measurable sets. Then,*

$$\bigcup_{i=1}^m E_i \qquad \bigcap_{i=1}^m E_i$$

are also Peano-Jordan measurable. Moreover, if $E_1, \dots, E_m \subset \mathbb{R}^N$ are pairwise disjoint, then

$$\mathcal{PJ}\left(\bigcup_{i=1}^m E_i\right) = \sum_{i=1}^m \mathcal{PJ}(E_i).$$

Remark 10.20. Note that neither the inner, nor the outer Peano-Jordan content are finitely additive. Indeed, consider the sets $E_1 \coloneqq \mathbb{Q} \cap [0, 1]$, and $E_2 \coloneqq [0, 1] \setminus \mathbb{Q}$. Then,

$$\mathcal{PJ}^-(E_1) = \mathcal{PJ}^-(E_2) = 0, \qquad \mathcal{PJ}^-(E_1 \cup E_2) = 1.$$

and

$$\mathcal{PJ}^+(E_1) = \mathcal{PJ}^+(E_2) = \mathcal{PJ}^+(E_1 \cup E_2) = 1,$$

even if E_1 and E_2 are disjoint.

An important example of a set with Peano-Jordan measure zero is the graph of a continuous function over a compact set.

Proposition 10.21. *Let $f : \mathbb{R}^N \to \mathbb{R}$ be a continuous function. Then,*

$$\mathcal{PJ}\left(\{(x, f(x)) \,:\, x \in K\,\}\right) = 0,$$

for all compact sets $K \subset \mathbb{R}^N$.

The proof is left to the reader. Finally, we give a characterization of Peano-Jordan measurability.

Theorem 10.22. *Let $E \subset \mathbb{R}^N$ be a bounded set. Then,*

$$\mathcal{PJ}^+(E) - \mathcal{PJ}^-(E) = \mathcal{PJ}^+(\partial E).$$

In particular, E is Peano-Jordan measurable if and only if $\mathcal{PJ}^+(\partial E) = 0$.

Proof. Step 1. We prove that

$$\mathcal{PJ}^+(E) - \mathcal{PJ}^-(E) \geq \mathcal{PJ}^+(\partial E).$$

Fix $\varepsilon > 0$. Let R_1, R_2 be pluri-rectangles such that

$$R_1 \subset E, \qquad \mathcal{PJ}^-(E) - \varepsilon \leq \mathcal{PJ}(R_1), \tag{10.5}$$

$$E \subset R_2, \qquad \mathcal{PJ}^+(E) + \varepsilon \geq \mathcal{PJ}(R_2). \tag{10.6}$$

Let $R := R_2 \setminus R_1$. Then, by Lemma 10.9, R is a pluri-rectangle, and $\partial E \subset R$. Therefore, by using the definition of $\mathcal{PJ}^+(\partial E)$ we have that

$$\mathcal{PJ}^+(\partial E) \leq \mathcal{PJ}(R) = \mathcal{PJ}(R_2) - \mathcal{PJ}(R_1) \leq \mathcal{PJ}^+(E) - \mathcal{PJ}^-(E) + 2\varepsilon,$$

where in the second step we used Lemma 10.19 while last step follows from (10.5), and (10.6). Since $\varepsilon > 0$ is arbitrary, we get the desired conclusion.

Step 2. We now prove that

$$\mathcal{PJ}^+(E) - \mathcal{PJ}^-(E) \leq \mathcal{PJ}^+(\partial E). \tag{10.7}$$

Fix $\varepsilon > 0$. By using Proposition 10.17, we can find $r > 0$ such that

$$\mathcal{PJ}^+(\partial E) + \varepsilon \geq r^N \# \{ i \in \mathbb{N} : Q(z_i, r) \cap \partial E \neq \emptyset,\, Q(z_i, r) \in \mathcal{G}(r) \}.$$

Note that if a cube $Q(z_i, r) \in \mathcal{G}(r)$ is such that $Q(z_i, r) \cap E \neq \emptyset$, then either $Q(z_i, r) \cap \partial E \neq \emptyset$, or $Q(z_i, r) \subset E$. Therefore,

$$\begin{aligned} &\# \{ i \in \mathbb{N} : Q(z_i, r) \cap E \neq \emptyset,\, Q(z_i, r) \in \mathcal{G}(r) \} \\ &\quad = \# \{ i \in \mathbb{N} : Q(z_i, r) \subset E \neq \emptyset,\, Q(z_i, r) \in \mathcal{G}(r) \} \\ &\qquad + \# \{ i \in \mathbb{N} : Q(z_i, r) \cap \partial E \neq \emptyset,\, Q(z_i, r) \in \mathcal{G}(r) \}. \end{aligned} \tag{10.8}$$

Thus, from (10.7) (10.8), we get that

$$\begin{aligned} \mathcal{PJ}^+(\partial E) + \varepsilon &\geq r^N \# \{ i \in \mathbb{N} : Q(z_i, r) \cap \partial E \neq \emptyset,\, Q(z_i, r) \in \mathcal{G}(r) \} \\ &= r^N \# \{ i \in \mathbb{N} : Q(z_i, r) \cap E \neq \emptyset,\, Q(z_i, r) \in \mathcal{G}(r) \} \\ &\qquad - r^N \# \{ i \in \mathbb{N} : Q(z_i, r) \subset E \neq \emptyset,\, Q(z_i, r) \in \mathcal{G}(r) \} \\ &\geq \mathcal{PJ}^+(E) - \mathcal{PJ}^-(E). \end{aligned}$$

Since $\varepsilon > 0$ is arbitrary, we get the desired conclusion. □

We now see some examples.

Example 10.23. Let E be the unit ball of $\mathbb{R}^2$ centered at the origin. Then, E is Peano-Jordan measurable. Indeed, it is easy to see that $\mathcal{PJ}^+(\partial E) = 0$. Thus, the result follows from Theorem 10.22. Moreover, we have that $\mathcal{PJ}(E) = \pi$.

Example 10.24. Let $E := \mathbb{N}$. Then, E is *not* Peano-Jordan measurable, since it is not bounded (see Lemma 10.18).

Example 10.25. Let $E := \{1/n : n \geq 1\}$. Then, E is Peano-Jordan measurable, and $\mathcal{PJ}(E) = 0$ (prove it!).

Example 10.26. Let $E := \mathbb{Q} \cap [0, 1]$. Then, E is *not* Peano-Jordan measurable. Indeed, $\mathcal{PJ}^+(\partial E) = \mathcal{PJ}^+([0, 1]) \neq 0$, and thus the result follows from Theorem 10.22.

Remark 10.27. The above three examples show something very important. All of the sets considered are countable. In the first and in the latter case, the set is not Peano-Jordan measurable, while in the second it is, and the Peano-Jordan content is zero. Note that, since a countable set $E \subset \mathbb{R}^N$ has empty interior, thanks to Lemma 10.18, we get that $\mathcal{PJ}^-(E) = 0$. Therefore, if the set E is dense or unbounded, then it is not Peano-Jordan measurable. Otherwise it is, and $\mathcal{PJ}(E) = 0$. Thus, the *topological* property of being dense or not determines, for a countable set whether it is Peano-Jordan measurable or not.

Remark 10.28. Another important fact about the Peano-Jordan content, that Examples 10.24 and 10.26 show, is that *countable* union of Peano-Jordan measurable sets might fail to be Peano-Jordan measurable.

10.3. **The Darboux-Riemann integral.** We now give a definition of the integral as the area under the graph, by following the same idea of the inner and outer Peano-Jordan content. This is what Darboux did in 1875. The idea is to define the Darboux integral for piecewise constant functions, and then to extend it to more general functions by using similar notions to the inner and the outer Peano-Jordan content.

We will then show the connection of this Darboux integral with the Riemann integral, and with the Jordan content of the sub-graph. Note that the sets where we are integrating our function are rectangles. This will be extended later to more general sets. Thus, in this section, $D \subset \mathbb{R}^N$ will always denote a rectangle.

Definition 10.29. Let $f : D \to \mathbb{R}$. We say that f is a *piecewise-constant* function if

$$f(x) = \sum_{i=1}^{k} c_i \mathbb{1}_{E_i}(x), \tag{10.9}$$

for some $c_i \in \mathbb{R}$, and some rectangles $E_1, \dots, E_k \subset \mathbb{R}^N$ with pairwise disjoint interiors, such that $E_1 \cup \cdots \cup E_k = D$.

Definition 10.30. Let $f : D \to \mathbb{R}$ be a piecewise-constant function as in (10.9). We define its *Darboux integral* of f on D as

$$\int_D f(x)\,dx := \sum_{i=1}^{k} c_i \mathcal{PJ}(E_i).$$

Remark 10.31. Note that the definition of the Darboux integral of a piecewise constant function is independent of the way the function is written. Moreover, we have that if $f : D \to \mathbb{R}$ is piecewise constant, then

$$\int_D f(x)\,dx := \sum_{i=1}^{k} a_i \mathcal{PJ}(A_i).$$

for any writing

$$f(x) = \sum_{i=1}^{m} a_i \mathbb{1}_{A_i}(x),$$

where $A_1, \dots, A_m$ are rectangles, even if not with pairwise disjoint interiors.

Definition 10.32. Let $f : D \to \mathbb{R}$ be a function, where $D \subset \mathbb{R}^N$ is a rectangle. Then, we define the *lower Darboux integral* as

$$\underline{\int_D} f(x)\,dx := \sup\left\{ \int_D g(x)\,dx \,:\, g \le f,\ g \text{ piecewise constant} \right\},$$

and the *upper Darboux integral* as

$$\overline{\int_D} f(x)\,dx := \inf\left\{ \int_D g(x)\,dx \,:\, g \ge f,\ g \text{ piecewise constant} \right\}$$

If the two values coincide, we denote the common value by

$$\int_D f(x)\,dx,$$

and we say that f is *Darboux integrable*.

Remark 10.33. Note that we are not assuming any continuity on f.

The reason why we didn't use the same symbol for the Darboux integral and for the Riemann integral, is because they are the same! Indeed, it is possible to see that the definition of Darboux integral is equivalent to the definition of Riemann integral[14] as the limit of the Cauchy-Riemann sums.

Theorem 10.34. *Let $f : D \to \mathbb{R}$. Then, f is Cauchy-Riemann integrable if and only if it is Darboux integrable. In such a case, the values of the two integrals coincide.*

Remark 10.35. The advantage of the Darboux definition lies in its simplicity for computations and proofs. In the following, we will just say that a function is Riemann integrable.

Next, we state that the notion of Darboux integral is the same as that obtained by considering Riemann sums, namely by restricting the class of simple functions we consider. The proof is immediate from the definition of Darboux lower and upper integrals.

Theorem 10.36. *Let $f : D \to \mathbb{R}$ be a function. Let $\mathcal{R}$ denote the family of finite partitions of D into pairwise disjoint rectangles. Then,*

$$\underline{\int_D} f(x)\,dx = \sup\left\{ \sum_{i=1}^k m_i \mathcal{PJ}(R_i) \,:\, (R_i)_{i=1}^k \in \mathcal{R},\, m_i := \inf_{R_i} f \right\},$$

and

$$\overline{\int_D} f(x)\,dx = \inf\left\{ \sum_{i=1}^k M_i \mathcal{PJ}(R_i) \,:\, (R_i)_{i=1}^k \in \mathcal{R},\, M_i := \sup_{R_i} f \right\}$$

In particular, f is Riemann integrable, if and only if, for every $\varepsilon > 0$ it is possible to find $(R_i)_{i=1}^k \in \mathcal{R}$ such that

$$\sum_{i=1}^k |M_i - m_i| \mathcal{PJ}(R_i) < \varepsilon.$$

Remark 10.37. What the above result says is that, given a partition $(R_i)_{i=1}^k \in \mathcal{R}$, the best you can do for the lower and the upper Darboux integral is to take the infimum and the supremum of f in R_i, respectively.

As anticipated at the beginning of this section, the definition of the Darboux integral makes it easier to see that the Riemann integral of a function is the signed Peano-Jordan content of its subgraph.

Theorem 10.38. *Let $f : D \to \mathbb{R}$. Then, f is Riemann integrable if and only if the sets*

$$E^+ := \{(x, y) \in D \times [0, \infty) \,:\, 0 \le y \le f(x)\},$$

and

$$E^- := \{(x, y) \in D \times (-\infty, 0] \,:\, f(x) \le y \le 0\}$$

are Peano-Jordan measurable. In this case, it holds

$$\int_D f(x)\,dx = \mathcal{PJ}(E^+) - \mathcal{PJ}(E^-).$$

10.4. **Lebesgue's characterization of Riemann integrability.** Finally, we discuss a result of fundamental importance for Riemann integration: the Lebesgue's characterization of Riemann integrable functions. The idea is the following: thanks to Theorem 10.36, we have that the lower and the upper Darboux integral have a chance to coincide, if the supremum and the infimum of f on R_i are close enough. This does not have to happen in all R_i's, but on sufficiently many. We now give the definitions needed to make the above heuristics clear.

[14] Note that the definition of Cauchy-Riemann integrability we gave, Definition 10.1 can be generalized to higher dimension.

Definition 10.39. Let $f : D \to \mathbb{R}$. We define the *oscillation* of f at the point $x \in D$ as

$$\omega_f(x) \coloneqq \inf_{r>0} \{ \sup |f(y) - f(z)| \, : \, y, z \in B(x,r) \} .$$

We prove two important properties of the oscillation. The first one says that the notion of oscillation is a *pointwise* notion.

Lemma 10.40. *Let $f : D \to \mathbb{R}$. Then,*

$$\omega_f(x) = \lim_{r \to 0} \left| \sup_{B(x,r)} f - \inf_{B(x,r)} f \right| .$$

The proof is left as an exercise to the reader.

Definition 10.41. Let $f : R \to \mathbb{R}$, and $\varepsilon > 0$. Denote by

$$\Omega(f, \varepsilon) \coloneqq \{ x \in D \, : \, \omega_f(x) \geq \varepsilon \} .$$

The second properties concerns the sup and sub level sets of the oscillation.

Lemma 10.42. *Let $f : D \to \mathbb{R}$, and $\varepsilon > 0$. Then, the set $\Omega(f, \varepsilon)$ is closed.*

Proof. We will show that the set

$$D \setminus \Omega(f, \varepsilon) = \{ x \in D \, : \, \omega_f(x) < \varepsilon \}$$

is open. Let $x \in D$ be such that $\omega_f(x) < \varepsilon$. Then, thanks to Lemma 10.40, there exists $r > 0$ such that

$$\left| \sup_{B(x,r)} f - \inf_{B(x,r)} f \right| < \varepsilon.$$

Thus, for every $y, z \in B(x,r)$, it holds

$$|f(y) - f(z)| < \varepsilon. \tag{10.10}$$

We claim that $B(x,r) \subset \{ p \in D \, : \, \omega_f(p) < \varepsilon \}$. Fix $y \in B(x,r)$, and let $r_0 \coloneqq r - \|y - x\|$. Then, from (10.10) we get that

$$\left| \sup_{B(y,r_0)} f - \inf_{B(y,r_0)} f \right| < \varepsilon.$$

Since

$$\sup_{B(y,s)} f \leq \sup_{B(y,r_0)} f, \qquad \inf_{B(y,s)} f \geq \inf_{B(y,r_0)} f,$$

for all $s \leq r_0$, we infer that

$$\left| \sup_{B(y,s)} f - \inf_{B(y,s)} f \right| \leq \left| \sup_{B(y,r_0)} f - \inf_{B(y,r_0)} f \right| < \varepsilon,$$

for all $s \leq r_0$. Thus, by using again Lemma 10.40, we get that $y \in \left\{ x \in \mathbb{R}^N \, : \, \omega_f(x) < \varepsilon \right\}$. This concludes the proof. □

We are now in position to prove the main result of this section. It is a characterization of Riemann integrability based on the *size* of the sets where the oscillations of f is large.

Theorem 10.43 (Lebesgue's characterization of Riemann integrability)**.** *Let $f : D \to \mathbb{R}$. Then, f is Riemann integrable if and only if it is bounded and $\mathcal{PJ}^+(\Omega(f, \varepsilon)) = 0$, for all $\varepsilon > 0$.*

Proof. Step 1. Assume that f is Riemann integrable. Without loss of generality, we can assume $f \geq 0$. Then, it is easy to see that f is bounded. We now prove that $\mathcal{PJ}^+(\Omega(f, \varepsilon)) = 0$, for all $\varepsilon > 0$. Assume by contradiction that there exists $\varepsilon > 0$ such that

$$\mathcal{PJ}^+(\Omega(f, \varepsilon)) = \delta > 0.$$

We want to show that, given any partition $(R_i)_{i=1}^k \in \mathcal{R}$,

$$\sum_{i=1}^k M_i \mathcal{PJ}(R_i) - \sum_{i=1}^k m_i \mathcal{PJ}(R_i) \geq \varepsilon\delta.$$

Here, $m_i \coloneqq \inf_{R_i} f$, and $M_i \coloneqq \sup_{R_i} f$. Thanks to Theorem 10.36, this implies that f is not Riemann integrable, contradicting our assumption.

First of all, since by our absurd hypothesis $\mathcal{PJ}^+(\Omega(f,\varepsilon)) > 0$, there exists $\delta > 0$ such that, for any $R \in \mathcal{R}$ with $\Omega(f,\varepsilon) \subset R$, it holds

$$\mathcal{PJ}^+(R) \geq \delta. \tag{10.11}$$

Let $(R_i)_{i=1}^k \in \mathcal{R}$, and assume that, up to renaming the indexes,

$$\Omega(f,\varepsilon) \subset \bigcup_{i=1}^m R_i, \tag{10.12}$$

for some $m \leq k$. Then, recalling that $f \geq 0$, we get

$$\begin{aligned}
\sum_{i=1}^k M_i \mathcal{PJ}(R_i) - \sum_{i=1}^k m_i \mathcal{PJ}(R_i) &= \sum_{i=1}^k \left[\sup_{R_i} f - \inf_{R_i} f \right] \mathcal{PJ}(R_i) \\
&\geq \sum_{i=1}^m \left[\sup_{R_i} f - \inf_{R_i} f \right] \mathcal{PJ}(R_i) \\
&\geq \varepsilon \sum_{i=1}^m \mathcal{PJ}(R_i) \\
&\geq \varepsilon\delta,
\end{aligned}$$

where in the last step we used (10.12) together with (10.11). This gives the desired contradiction.

Step 2. Assume that

$$0 \leq f \leq M, \tag{10.13}$$

for some $M < \infty$, and that $\mathcal{PJ}^+(\Omega(f,\varepsilon)) = 0$ for all $\varepsilon > 0$. Fix $\varepsilon > 0$. Then, it is possible to find $R \in \mathcal{R}$ such that

$$\Omega(f,\varepsilon) \subset \mathring{R}, \qquad \mathcal{PJ}(R) < \varepsilon, \tag{10.14}$$

where $\mathring{R}$ denotes the interior of R. Let

$$S \coloneqq D \setminus \mathring{R}.$$

Then, S is compact, and, by using (10.14), we have that

$$S \subset \{x \in D : \omega_f(x) < \varepsilon\}.$$

Thanks to Lemma 10.42, since the set on the right-hand side is open, for each $x \in S$ it is possible to find $r(x) > 0$ such that the open cube $Q(x, r(x))$ centered at x, with sides of length $r(x)$ parallel to the orthogonal axes, such that

$$Q(x, r(x)) \subset \{x \in D : \omega_f(x) < \varepsilon\}. \tag{10.15}$$

Therefore, the family

$$\{Q(x, r(x))\}_{x \in S}$$

is an open covering of the set S. Since S is compact, by using Theorem 2.35, it is possible to find a finite family that covers S, say $Q(x_1, r(x_1)), \dots, Q(x_k, r(x_k))$. Since it is possible to write a finite union of cubes as a disjoint union of rectangles, we can assume, without loss of generality (in order not to use a heavy notation), that

$$Q(x_1, r(x_1)), \dots, Q(x_k, r(x_k))$$

are pluri-rectangles with pairwise disjoint interior. Consider the partition of D given by

$$\widetilde{R},\, Q(x_1, r(x_1)), \dots, Q(x_k, r(x_k)),$$

where

$$\widetilde{R} \coloneqq R \setminus \bigcup_{i=1}^{k} Q(x_i, r(x_i)).$$

By using (10.15), we get that

$$\begin{aligned}\left[\sup_{\widetilde{R}} f - \inf_{\widetilde{R}} f\right] \mathcal{PJ}(\widetilde{R}) + \sum_{i=1}^{k} [M_i - m_i] \mathcal{PJ}(Q(x_i, r(x_i))) \\ \leq M\mathcal{PJ}(R) + \varepsilon \sum_{i=1}^{k} \mathcal{PJ}(Q(x_i, r(x_i))) \\ \leq M\varepsilon + \varepsilon \mathcal{PJ}(D),\end{aligned}$$

where in the last step we used (10.13) to bound the first term, and (10.14) to bound the second. Since $\varepsilon > 0$ is arbitrary, we conclude that f is Riemann integrable thanks to Theorem 10.36. □

Remark 10.44. In particular, what the above result says, is that it is the *size* of the set where the oscillation is large that determines whether a function is integrable or not. We will see in the next chapter how to relate such information to another notion of size of the discontinuity set of f.

Finally, we extend the integral of a function to a general domain $E \subset \mathbb{R}^N$.

Definition 10.45. Let $E \subset \mathbb{R}^N$ be a bounded set, and let $D \subset \mathbb{R}^N$ be a rectangle such that $\bar{E} \subset D$. We say that a function $f : E \to \mathbb{R}$ is *Riemann integrable*, if the function $\widetilde{f} : D \to \mathbb{R}$ defined as

$$\widetilde{f}(x) \coloneqq \begin{cases} f(x) & \text{if } x \in E, \\ 0 & \text{else,} \end{cases}$$

is Riemann integrable.

As a consequence of Theorem 10.43, we have the following characterization of functions that are Riemann integrable over a general set.

Corollary 10.46. *Let $E \subset \mathbb{R}^N$ be a bounded set. Then, a function $f : E \to \mathbb{R}$ is Riemann integrable if and only if it is bounded, and $\mathcal{PJ}^+(\Omega(\widetilde{f}, \varepsilon)) = 0$, for all $\varepsilon > 0$.*

We now have a complete characterization of Riemann integrable functions, and the relation of the Cauchy-Darboux-Riemann integral with the Peano-Jordan measure of the subgraph of the function. The question is: how is it possible to develop a new theory of integration that allows to treat more general functions, like (10.1), and that requires less strict assumptions for having the identity

$$\lim_{n\to\infty} \int_D f_n(x)\, dx = \int_D f(x)\, dx$$

in force?

11. Lebesgue measure

In this section we present one of the most important cornerstone of modern mathematics: the Lebesgue measure. This was developed by Lebesgue in his PhD thesis of 1901 with the goal of building a theory of integration that is able to overcome the difficulties of the Cauchy-Darboux-Riemann integral (task that was completed in 1902). This work, together with those of Borel, Carathéodory, Luzin, and Radon laid the foundations of *Measure Theory*, that you will develop more in details in the homonym course.

To understand the idea behind the outer Lebesgue measure, consider the following example. Let $f_n : (0,1) \to \mathbb{R}$ be defined as $f_n(x) \coloneqq x^n$. As we have seen in Remark 10.3, the sequence $\{f_n\}_{n\in\mathbb{N}}$ converges *pointwise* to the function $f \equiv 0$. Since the convergence is only pointwise, by the sole knowledge that each f_n is a Riemann integrable function, we cannot conclude that f is Riemann integrable. How would we prove that f is Riemann integrable? Well, we need to bound from above and from below f with two piece-wise constant functions whose integrals are sufficiently close to each other. An idea to do that is the following: Since all functions f_n's are bounded by one, we can consider the sets

$$E^n_{k,i} \coloneqq \left\{ x \in (0,1) \,:\, \frac{i}{k} \le f_n(x) < \frac{i+1}{k} \right\},$$

for all $k \ge 1$, and $i = 0, \dots, k-1$. Note that each set $E^n_{k,i}$ is Peano-Jordan measurable. Let $g_n, h_n : (0,1) \to \mathbb{R}$ be the piecewise constant functions defined as

$$g_n(x) = \frac{i}{k}, \qquad h_n(x) = \frac{i+1}{k}, \qquad \text{on } E^n_{k,i}.$$

Then, by definition, $g_n \le f_n \le h_n$. Moreover,

$$\int_{(0,1)} g_n(x)\,dx \le \int_{(0,1)} f_n(x)\,dx \le \int_{(0,1)} h_n(x)\,dx,$$

and

$$\int_{(0,1)} h_n(x)\,dx - \int_{(0,1)} g_n(x)\,dx \le \frac{1}{k}.$$

We would expect that the limiting function f would be bounded above and below by the limits g and h of g_n, and h_n, respectively. These limiting functions would, hopefully, be defined as

$$g(x) = \frac{i}{k}, \qquad h(x) = \frac{i+1}{k}, \qquad \text{on } E_{k,i},$$

where $E_{k,i}$ is the *limit* of the sequence of sets $\{E^n_{k,i}\}_{n\in\mathbb{N}}$. The question is then: what is the limit of the Peano-Jordan measurable sets $E^n_{k,i}$? Is it a set $E_{k,i}$ Peano-Jordan measurable? The issue is that we are dealing with a *sequence* of Peano-Jordan measurable sets. The limiting set is not ensured to be Peano-Jordan measurable, even in the special case above where the sets $E^n_{k,i}$ are decreasing (or increasing). The problem is that the Peano-Jordan content, being defined by using *finite* partitions outside and inside a set, does not behave well with respect to *sequences* of sets, and, in particular, with respect to the *countable* union or intersection of measurable sets.

11.1. Definition and relation to the Peano-Jordan content.

The idea of Lebesgue is to allow for a general *covering* of a set by *countably many* cubes (or rectangles).

Definition 11.1. Let $E \subset \mathbb{R}^N$. We define the *outer Lebesgue measure* of E as

$$\mathcal{L}^N(E) \coloneqq \inf \left\{ \sum_{i=0}^{\infty} r_i^N \,:\, E \subset \bigcup_{i=0}^{\infty} Q(x_i, r_i),\, x_i \in \mathbb{R}^N,\, r_i \ge 0 \right\}.$$

Remark 11.2. The fact that, in the above definition, we can allow some of the r_i's to be zero, simply means that we can also take a finite covering of E. Note that the cubes need not to be disjoint. Moreover, in the definition, we can also use rectangles, or pluri-rectangles. This is

because every rectangle and pluri-rectangle can be written as a finite union of cubes, while a cube is a special case of a rectangle or of a pluri-rectangle.

First of all, we note that the outer Lebesgue measure is consistent with the basic notion of measure for rectangles.

Lemma 11.3. *Let $R \subset \mathbb{R}^N$ be the rectangle*

$$R = (a_1, b_1) \times \cdots \times (a_N, b_N).$$

Then, $\mathcal{L}^N(R) = \mathcal{L}^N(\overline{R}) = \mathcal{PJ}(R) = (b_1 - a_1) \cdots (b_N - a_N)$.

The proof is left as an exercise to the reader. Note that the inequality

$$\mathcal{L}^N(\overline{R}) \leq (b_1 - a_1) \cdots (b_N - a_N)$$

follows from the definition of the outer Lebesgue measure. To prove the other inequality, it has to be shown that, given *any* covering $\{Q(x_i, r_i)\}_{i \in \mathbb{N}}$ of R, it holds

$$\sum_{i=0}^{\infty} r_i^N \leq (b_1 - a_1) \cdots (b_N - a_N).$$

This requires a bit of care.

By using the same ideas, we get the following result.

Lemma 11.4. *Let $R, S \subset \mathbb{R}^N$ be two rectangles (open, closed, or anything in between). Then,*

$$\mathcal{L}^N(R) = \mathcal{L}^N(R \cap S) + \mathcal{L}^N(R \setminus S).$$

The outer Lebesgue measure satisfies some basic properties that will be useful later.

Lemma 11.5. *The followings hold:*

(i) $\mathcal{L}^N(\emptyset) = 0$;
(ii) If $E \subset F$, then, $\mathcal{L}^N(E) \leq \mathcal{L}^N(F)$;
(iii) For any $\{E_n\}_{n \in \mathbb{N}} \subset \mathbb{R}^N$, it holds

$$\mathcal{L}^N\left(\bigcup_{n=0}^{\infty} E_n\right) \leq \sum_{n=0}^{\infty} \mathcal{L}^N(E_n);$$

(iv) $\mathcal{L}^N(x + E) = \mathcal{L}^N(E)$, for all $x \in \mathbb{R}^N$, and all $E \subset \mathbb{R}^N$.

Remark 11.6. Properties (ii), (iii), and (iv) are called *monotonicity* and the *countable sub-additivity*, and *translation invariance*, respectively. Note that, by combining (iii) with (ii), we get

$$\mathcal{L}^N(E_1 \cup \cdots \cup E_k) \leq \mathcal{L}^N(E_1) + \cdots + \mathcal{L}^N(E_k),$$

for all $E_1, \ldots, E_k \subset \mathbb{R}^N$.

By using Lemma 11.3 together with Lemma 11.5(ii), it is possible to compute the outer Lebesgue measure of a countable union of cubes with pairwise disjoint interiors[15].

Lemma 11.7. *It holds*

$$\mathcal{L}^N\left(\bigcup_{i \in \mathbb{N}} Q(z_i, r_i)\right) = \mathcal{L}^N\left(\bigcup_{i \in \mathbb{N}} \overline{Q(z_i, r_i)}\right) = \sum_{i \in \mathbb{N}} \mathcal{L}^N(Q(z_i, r_i)) = \sum_{i \in \mathbb{N}} r_i^N,$$

whenever the cubes $Q(z_i, r_i)$ have pairwise disjoint interiors.

[15] Sets $\{A_i\}_{i \in \mathbb{N}}$ are called *pairwise disjoint* if $A_i \cap A_j = \emptyset$ for any $i \neq j$. Note that this is stronger than asking for the sets to be *disjoint*, which means that $\cap_{i \in \mathbb{N}} A_i = \emptyset$.

Proof. Let

$$E \coloneqq \bigcup_{i\in\mathbb{N}} Q(z_i, r_i).$$

By countable sub-additivity (see Lemma 11.5(iii)), we get that

$$\mathcal{L}^N(E) \leq \sum_{i\in\mathbb{N}} \mathcal{L}^N(Q(z_i, r_i)) = \sum_{i\in\mathbb{N}} r_i^N,$$

where the last step follows from Lemma 11.3. Now, since for every $n \in \mathbb{N}$

$$\bigcup_{i=1}^{n} Q(z_i, r_i) \subset E,$$

using again monotonicity, we get that

$$\sum_{i=1}^{n} r_i^N = \mathcal{L}^N\left(\bigcup_{i=1}^{n} Q(z_i, r_i)\right) \leq \mathcal{L}^N(E).$$

By taking the limit as $n \to \infty$, we get the opposite inequality. The same proof also holds for the case where we take the closed cubes. □

We now investigate the relation between the outer Lebesgue measure and the Peano-Jordan content. First of all, you might wonder why one is called Peano-Jordan *content*, while the other Lebesgue *outer measure*. The reason is that the latter is not what mathematicians call an *measure*, and not even what is called an *outer measure*: those are functions defined on $\mathcal{P}(X)$ and on a subfamily of it, respectively, satisfying certain properties, that the Peano-Jordan content does not obey.

Remark 11.8. Another difference, is that *unbounded* sets can have finite outer Lebesgue measure. Indeed, it is possible to prove that the unbounded set

$$E \coloneqq \bigcup_{n\geq 1} \left(n - \frac{1}{2n^2}, n + \frac{1}{2n^2}\right)$$

is such that $\mathcal{L}^N(E) = \pi^2/6$.

As for the Peano-Jordan content, one would expect to have a definition of *inner* Lebesgue measure of a set E, by using countable cubes contained in the interior of E. Contrary to the outer Lebesgue measure, allowing countably many cubes does not change anything for the inner Peano-Jordan content. This is because of the following result.

Lemma 11.9. *Let $E \subset \mathbb{R}^N$ be an open set. Then, E can be written as a countable union of closed cubes with pairwise disjoint interiors.*

Proof. The idea is to consider union of diadic cubes, namely a particular sequence for r-grids (see Definition 10.16). We will write

$$E = \bigcup_{n\in\mathbb{N}} U_n, \tag{11.1}$$

where each U_n is a countable union of closed cubes with pairwise disjoint interiors. To define the sets U_n's, we proceed as follows. Let

$$I_0 \coloneqq \{i \in \mathbb{N} : Q(z_i, 1) \subset E,\, z_i \in \mathbb{Z}\},$$

and let

$$U_0 \coloneqq \bigcup_{i\in I_0} \overline{Q(z_i, 1)}$$

For each $n \in \mathbb{N} \setminus \{0\}$, define recursively the sets I_n, and U_n as follows:

$$I_n \coloneqq \left\{i \in \mathbb{N} : \overline{Q\left(z_i, \frac{1}{2^n}\right)} \subset (E \setminus \mathring{U}_{n-1}),\, z_i \in \frac{1}{2^n}\mathbb{Z}\right\},$$

and let

$$U_n := \bigcup_{i\in I_n} \overline{Q\left(z_i, \frac{1}{2^n}\right)}.$$

We now claim that (11.1) holds. Note that, by definition, the inclusion $\supset$ is in force. To prove the opposite, we argue as follows. Since E is open, for each $x \in E$ there exists $r > 0$ such that $Q(x, r) \subset E$. In particular, there exists $n \in \mathbb{N}$, such that

$$x \in Q\left(z_i, \frac{1}{n}\right) \subset E.$$

for some $z_i \in \frac{1}{n}\mathbb{Z}$. Therefore, either $i \in I_n$, and thus in U_n, or, if $i \notin I_n$, contained in some U_m, for $m < n$.

Finally, note that

$$Q\left(z_i, \frac{1}{2^n}\right) \cap Q\left(z_j, \frac{1}{2^n}\right) = \emptyset,$$

for all $n \in \mathbb{N}$, and all $i \neq j \in I_n$. Moreover, if $m \neq n$, then,

$$Q\left(z_i, \frac{1}{2^n}\right) \cap Q\left(z_j, \frac{1}{2^m}\right) = \emptyset,$$

for all $i \in I_n$, and all $j \in I_m$. Thus, all cubes in (11.1) have pairwise disjoint interiors. □

Remark 11.10. As a consequence of the previous result, we have that

$$\mathcal{PJ}^-(E) = \sup\left\{ \sum_{i=0}^{\infty} r_i^N \ : \ \bigcup_{i=0}^{\infty} Q(x_i, r_i) \subset E,\ x_i \in \mathbb{R}^N,\ r_i \geq 0 \right\},$$

for all sets $E \subset \mathbb{R}^N$.

Remark 11.11. Note that we are using *closed* cubes in Lemma 11.9. Therefore, the unit cube $Q = (0, 1)^N$ requires countably many cubes to be written in such a way.

Remark 11.12. There is, however, a notion of *inner* Lebesgue measure, that can be used to define the notion of Lebesgue measurable sets. It reads as follows: let $E \subset \mathbb{R}^N$ be a bounded set, and let $R \subset \mathbb{R}^N$ be a rectangle containing E. We define the *inner* Lebesgue measure of E as

$$\mathcal{L}^N(R) - \mathcal{L}^N(R \setminus E).$$

It can be shown that this definition does not depend on the containing rectangle R. Since the notion of measurability by using this concept of inner Lebesgue measure is not intuitive, we will not discuss it further.

We then show that the Lebesgue outer measure extends the notion of the Peano-Jordan content, and we investigate the relation among the two notions.

Theorem 11.13. *Let $E \subset \mathbb{R}^N$ be bounded. Then,*

$$\mathcal{PJ}^-(E) = \mathcal{L}^N(\mathring{E}), \qquad \mathcal{PJ}^+(E) = \mathcal{L}^N(\overline{E}).$$

In particular, if E is Peano-Jordan measurable, then $\mathcal{PJ}(E) = \mathcal{L}^N(E) = \mathcal{L}^N(\mathring{E}) = \mathcal{L}^N(\overline{E})$.

Proof. Step 1. We prove that $\mathcal{PJ}^-(E) = \mathcal{L}^N(\mathring{E})$. First of all, note that, by definition, and by using Lemma 11.5(ii), we get

$$\mathcal{PJ}^-(E) \leq \mathcal{L}^N(\mathring{E}).$$

To prove the opposite inequality, by using Lemma 11.9, we have that it is possible to write

$$\mathring{E} = \bigcup_{i=0}^{\infty} Q(x_i, r_i),$$

where the cubes $Q(x_i, r_i)$ have pairwise disjoint interior. For each $n \in \mathbb{N}$ let

$$R_n := \bigcup_{i=0}^{n} Q(x_i, r_i).$$

Then, by using Lemma 11.7, we get

$$\mathcal{L}^N(R_n) = \sum_{i=0}^{n} r_i^N,$$

Moreover,

$$\mathcal{PJ}^-(\mathring{E}) \geq \sum_{i=0}^{n} r_i^N,$$

for all $n \in \mathbb{N}$. Therefore,

$$\mathcal{PJ}^-(\mathring{E}) \geq \sum_{i=0}^{\infty} r_i^N = \mathcal{L}^N(\mathring{E}),$$

where in the last step we used Lemma 11.7 again. This concludes this step.

Step 2. We now prove that $\mathcal{PJ}^+(E) = \mathcal{L}^N(\overline{E})$. First, we show that $\mathcal{PJ}^+(E) \leq \mathcal{L}^N(\overline{E})$. First of all, note that since E is bounded, $\mathcal{L}^N(E) < \infty$. Let $\varepsilon > 0$, and let

$$\{Q(x_i, r_i)\}_{i \in \mathbb{N}}$$

be such that

$$\mathcal{L}^N(\overline{E}) + \varepsilon \geq \sum_{i=0}^{\infty} r_i^N.$$

Since E is compact, it is possible to cover is with a finite number of the open cubes $Q(x_i, r_i)$'s. Up to renaming the sets, we can assume that

$$\overline{E} \subset \bigcup_{i=0}^{k} Q(x_i, r_i), \tag{11.2}$$

for some $k \in \mathbb{N}$. Therefore, from the definition of $\mathcal{PJ}^+(\bar{E})$ we get

$$\mathcal{PJ}^+(E) \leq \mathcal{PJ}^+(\overline{E}) \leq \sum_{i=0}^{k} r_i^N \leq \sum_{i=0}^{\infty} r_i^N \leq \mathcal{L}^N(\overline{E}) + \varepsilon,$$

where in the last step we used (11.2). Since $\varepsilon > 0$ is arbitrary, we get the desired inequality.

Then, we show that $\mathcal{PJ}^+(E) \geq \mathcal{L}^N(\overline{E})$. This follows directly from the definition, since for the outer Lebesgue measure we are allowed to take countably many cubes.

Step 3. The last claim of the result follows from the definition of Peano-Jordan measurability, and the previous two steps. □

11.2. **Measurable sets.** We now investigate how the outer Lebesgue measure behaves with respect to union of sets. This is where things become tricky. A property that we expect to be true is for the Lebesgue measure to be *finitely additive*: let $E, F \subset \mathbb{R}^N$ be disjoint. Then,

$$\mathcal{L}^N(E \cup F) = \mathcal{L}^N(E) + \mathcal{L}^N(F). \tag{11.3}$$

Note that, by sub-additivity (see Lemma 11.5 (ii)), the inequality

$$\mathcal{L}^N(E \cup F) \leq \mathcal{L}^N(E) + \mathcal{L}^N(F)$$

is always true. Surprisingly, the opposite inequality

$$\mathcal{L}^N(E \cup F) \geq \mathcal{L}^N(E) + \mathcal{L}^N(F)$$

might fail to be true! The fact that it is possible to construct sets for which the above, apparently innocent, equality does not hold, was first discovered by Vitali in 1905 for the case $N = 1$ (see Theorem 11.38). The construction was later extended to the case of a general dimension. The problem is the following: if we want (11.3) in force, we need to *separate* E and F in such a way that countable coverings of the two do not interact. This is possible if the sets are *far apart* from each other.

Proposition 11.14. *Let $E, F \subset \mathbb{R}^N$ be such that*

$$\mathrm{d}(E,F) := \inf\{\, \|x-y\| \,:\, x \in E,\, y \in F \,\} > 0.$$

Then, $\mathcal{L}^N(E \cup F) = \mathcal{L}^N(E) + \mathcal{L}^N(F)$.

The proof is left as an exercise to the reader. Nevertheless, if E and F are a general pair of disjoint sets, it can happen that they are so intertwined that every covering for E interacts with every covering of F. Thus, we will always have overlaps when considering $E \cup F$. This might causes $\mathcal{L}^N(E \cup F)$ to be strictly smaller than $\mathcal{L}^N(E) + \mathcal{L}^N(F)$. What to do then? Well, the solution is the following: let us just restrict the outer Lebesgue measure to sets that behave in a good way. This defines a notion, called *measurability*. It is nowadays common to write condition (11.3) in an equivalent way that does not involve the requirement to take a set F that is disjoint from E.

Definition 11.15. A set $E \subset \mathbb{R}^N$ is called *Lebesgue measurable* (or *$\mathcal{L}^N$-measurable*) if

$$\mathcal{L}^N(F) = \mathcal{L}^N(F \setminus E) + \mathcal{L}^N(E \cap F),$$

for all sets $F \subset \mathbb{R}^N$.

Remark 11.16. Note that $E \subset \mathbb{R}^N$ is $\mathcal{L}^N$-measurable if and only if $\mathbb{R}^N \setminus E$ is $\mathcal{L}^N$-measurable.

Definition 11.17. The outer Lebesgue measure restricted to the family of Lebesgue measurable sets is called *Lebesgue measure*.

First of all, we show that the definition of measurability above is equivalent to a stronger version of (11.3).

Lemma 11.18. *Let $E \subset \mathbb{R}^N$. Then, E is Lebesgue measurable if and only if*

$$\mathcal{L}^N(G \cup F) = \mathcal{L}^N(G) + \mathcal{L}^N(F),$$

for all $G \subset E$, and all $F \subset \mathbb{R}^N$ with $E \cap F = \emptyset$.

Proof. Step 1. Assume E to be Lebesgue measurable. Let $G \subset E$, and $F \subset \mathbb{R}^N$ with $E \cap F = \emptyset$. Then, define $H := G \cup F$. By definition of measurability, we get that

$$\mathcal{L}^N(G \cup F) = \mathcal{L}^N(H) = \mathcal{L}^N(H \cap E) + \mathcal{L}^N(H \setminus E) = \mathcal{L}^N(G) + \mathcal{L}^N(F).$$

Step 2. Assume the additivity condition to hold. Let $A \subset \mathbb{R}^N$. Define

$$G := A \cap E, \qquad F := A \setminus E.$$

Then, G and F satisfies the assumptions, and therefore

$$\mathcal{L}^N(A) = \mathcal{L}^N(G \cup F) = \mathcal{L}^N(G) + \mathcal{L}^N(F) = \mathcal{L}^N(A \cap E) + \mathcal{L}^N(A \setminus E).$$

This concludes the proof. □

Next, we show that the family of $\mathcal{L}^N$-measurable sets behaves nicely with respect to *countable* Boolean operations: indeed, it is closed under complement, countable union, and, in turn, also under countable intersection.

Proposition 11.19. *The followings hold:*

(i) $\emptyset$, and $\mathbb{R}^N$ are Lebesgue measurable;
(ii) If $E, F \subset \mathbb{R}^N$ are Lebesgue measurable, then also $E \setminus F$ is;

(iii) If $(E_i)_{i\in\mathbb{N}} \subset \mathbb{R}^N$ is a sequence of Lebesgue measurable sets, then

$$\bigcup_{i\in\mathbb{N}} E_i, \qquad \bigcap_{i\in\mathbb{N}} E_i,$$

are also Lebesgue measurable;

(iv) If $E \subset \mathbb{R}^N$ is Lebesgue measurable, then $x + E$ is Lebesgue measurable, for all $x \in \mathbb{R}^N$.

Proof. Step 1. It is easy to see that $\emptyset$ is Lebesgue measurable, and that, E is Lebesgue measurable if and only if $\mathbb{R}^N \setminus E$ is. Moreover, also (iv) is easy to verify.

Step 2. Let $E, F \subset \mathbb{R}^N$ be Lebesgue measurable. To prove that $E \setminus F$ is Lebesgue measurable, let $G \subset \mathbb{R}^N$. Then,

$$\begin{aligned}
\mathcal{L}^N(G) &= \mathcal{L}^N(G \cap F) + \mathcal{L}^N(G \setminus F) \\
&= \mathcal{L}^N(G \cap F) + \mathcal{L}^N((G \setminus F) \cap E) + \mathcal{L}^N((G \setminus F) \setminus E) \\
&= \big[\mathcal{L}^N(G \cap F) + \mathcal{L}^N((G \setminus F) \setminus E)\big] + \mathcal{L}^N((G \setminus F) \cap E) \\
&= \big[\mathcal{L}^N((G \setminus (E \setminus F)) \cap F) + \mathcal{L}^N((G \setminus (E \setminus F)) \setminus F)\big] + \mathcal{L}^N(G \cap (E \setminus F)) \\
&= \mathcal{L}^N(G \setminus (E \setminus F)) + \mathcal{L}^N(G \cap (E \setminus F)),
\end{aligned}$$

where we have used the set equalities

$$G \cap F = (G \setminus (E \setminus F)) \cap F, \qquad (G \setminus F) \setminus E = (G \setminus (E \setminus F)) \setminus F).$$

This proves that $E \setminus F$ is Lebesgue measurable.

Step 3. Let $(E_i)_{i\in\mathbb{N}}$ be a sequence of Lebesgue measurable sets. We first assume $\{E_i\}_{i\in\mathbb{N}}$ to be pairwise disjoint. Set

$$U := \bigcup_{i\in\mathbb{N}} E_i.$$

Let $A \subset \mathbb{R}^N$. We want to prove that

$$\mathcal{L}^N(A) \geq \mathcal{L}^N(A \cap U) + \mathcal{L}^N(A \setminus U). \tag{11.4}$$

Let

$$U_n := \bigcup_{i=0}^{n} E_i.$$

We claim that U_n is Lebesgue measurable. In particular, we claim that

$$\mathcal{L}^N(A) \geq \sum_{i=0}^{n} \mathcal{L}^N(A \cap E_i) + \mathcal{L}^N(A \setminus U_n), \tag{11.5}$$

for all $n \in \mathbb{N}$, and all $A \subset \mathbb{R}^N$. This allows to obtain (11.4). Indeed, since

$$A \setminus U \subset A \setminus U_n,$$

from sub-additivity (see Lemma 11.5 (ii)) we get

$$\mathcal{L}^N(A \setminus U) \leq \mathcal{L}^N(A \setminus U_n),$$

for all $n \in \mathbb{N}$. Therefore, from (11.5), we get

$$\mathcal{L}^N(A) \geq \sum_{i=0}^{\infty} \mathcal{L}^N(A \cap E_i) + \mathcal{L}^N(A \setminus U) \geq \mathcal{L}^N(A \cap U) + \mathcal{L}^N(A \setminus U),$$

where last inequality follows from countable sub-additivity (see Lemma 11.5 (iii)), since

$$A \cap U \subset \bigcup_{i\in\mathbb{N}} (A \cap E_i).$$

Thus, we are left with proving (11.4). We prove (11.4) by induction on $n \in \mathbb{N}$. For $n = 1$ it follows from the measurability of E_1. Assume that (11.4) holds for $n \in \mathbb{N}$.] Then, by using the measurability of E_{n+1}, and then that of U_n, we get

$$\begin{aligned}\mathcal{L}^N(A) &= \mathcal{L}^N(A \cap E_{n+1}) + \mathcal{L}^N(A \setminus E_{n+1}) \\ &= \mathcal{L}^N(A \cap E_{n+1}) + \mathcal{L}^N\left((A \setminus E_{n+1}) \cap U_n\right) + \mathcal{L}^N\left((A \setminus E_{n+1}) \setminus U_n\right) \\ &= \mathcal{L}^N(A \cap E_{n+1}) + \mathcal{L}^N\left((A \setminus E_{n+1}) \cap U_n\right) + \mathcal{L}^N\left(A \setminus U_{n+1}\right) \\ &= \mathcal{L}^N(A \cap E_{n+1}) + \mathcal{L}^N(A \cap U_n) + \mathcal{L}^N\left(A \setminus U_{n+1}\right), \end{aligned} \tag{11.6}$$

where, in the previous to last step, we used the definition of U_{n+1} to get that

$$(A \setminus E_{n+1}) \setminus U_n = A \setminus U_{n+1},$$

while last step follows from the identity

$$(A \setminus E_{n+1}) \cap U_n = A \cap U_n,$$

since the sets E_i's are pairwise disjoint. Now, by using the set $A \cap U_n$ in (11.5) instead of A, we get

$$\mathcal{L}^N(A \cap U_n) \geq \sum_{i=0}^{n} \mathcal{L}^N(A \cap E_i), \tag{11.7}$$

since $A \cap U_n = A \cap E_i$, and $(A \cap U_n) \setminus U_n = \emptyset$. Thus, from (11.6) and (11.7), we get

$$\begin{aligned}\mathcal{L}^N(A) &\geq \mathcal{L}^N(A \cap E_{n+1}) + \sum_{i=0}^{n} \mathcal{L}^N(A \cap E_i) + \mathcal{L}^N\left(A \setminus U_{n+1}\right) \\ &= \sum_{i=0}^{n+1} \mathcal{L}^N(A \cap E_i) + \mathcal{L}^N\left(A \setminus U_{n+1}\right). \end{aligned}$$

This concludes the proof of this step.

Step 4. Assume that $(E_i)_{i\in\mathbb{N}} \subset \mathbb{R}^N$ is a sequence of measurable sets. For each $i \in \mathbb{N} \setminus \{0\}$, we can define

$$\widetilde{E}_i := E_i \setminus \bigcup_{j=1}^{i-1} E_j.$$

Note that by Step 2, each $\widetilde{E}_i$ is Lebesgue measurable. Since

$$\bigcup_{i\in\mathbb{N}} E_i = \bigcup_{i\in\mathbb{N}} \widetilde{E}_i,$$

we conclude from step 3 that their union is Lebesgue measurable.

Step 5. Finally, we use Morgan's law to write

$$\bigcap_{i\in\mathbb{N}} E_i = \mathbb{R}^N \setminus \bigcup_{i\in\mathbb{N}} (\mathbb{R}^N \setminus E_i).$$

Since in the previous steps we established that each set in the union on the right-hand side is Lebesgue measurable, we also get that

$$\bigcap_{i\in\mathbb{N}} E_i$$

is Lebesgue measurable. □

Remark 11.20. Note that the family of Peano-Jordan measurable sets is closed under *finite* union and *finite* intersection, but not under countable union or intersection. Moreover, it is closed under complement with respect to a *bounded* rectangle R, not with respect to the entire space $\mathbb{R}^N$ (since Peano-Jordan measurable sets must be bounded).

Families of sets that are closed under complement, countable union, and countable intersection are well studied in mathematics.

Definition 11.21. A family of sets $\mathcal{A} \subset \mathcal{P}(\mathbb{R}^N)$ is called a *σ-algebra* if

(i) $\emptyset \in \mathcal{A}$;
(ii) If $E \in \mathcal{A}$, then $\mathbb{R}^N \setminus E \in \mathcal{A}$;
(iii) If $(E_i)_{i\in\mathbb{N}} \subset \mathcal{A}$, then

$$\bigcup_{i=0}^{\infty} E_i \in \mathcal{A}.$$

Remark 11.22. Note that, if $\mathcal{A} \subset \mathcal{P}(\mathbb{R}^N)$ is a σ-algebra, then

$$\bigcap_{i=0}^{\infty} E_i = \mathbb{R}^N \setminus \bigcup_{i\in\mathbb{N}} (\mathbb{R}^N \setminus E_i) \in \mathcal{A}.$$

whenever $(E_i)_{i\in\mathbb{N}} \subset \mathcal{A}$.

Remark 11.23. Proposition 11.19 states that Lebesgue measurable sets are a σ-algebra.

So far, the only two examples of Lebesgue measurable sets are the empty set and the entire space. Not that many! Luckily, we can prove that a rectangle is Lebesgue measurable, and, by using previous results, obtain that many sets are indeed Lebesgue measurable.

Lemma 11.24. *Let $R \subset \mathbb{R}^N$ be a rectangle. Then, R is Lebesgue measurable.*

Proof. The idea is the following: a rectangle $R \subset \mathbb{R}^N$ has a very nice structure, in the following sense: let $F \subset \mathbb{R}^N$ be any set, and consider the two sets $R \cap F$ and $F \setminus R$. If $\{Q_i\}_{i\in\mathbb{N}}$ is a covering with open cubes of F, it is possible to obtain coverings of $R\cap F$ and $F\setminus R$, respectively by intersecting the cubes Q_i's with R and with $\mathbb{R}^N \setminus R$, respectively. This allows to show that the outer Lebesgue measure of F can be obtained as the sum of the outer Lebesgue measures of $R \cap F$ and $F \setminus R$. Let us make this heuristics more precise.

We need to prove the inequality

$$\mathcal{L}^N(F) \geq \mathcal{L}^N(R \cap F) + \mathcal{L}^N(F \setminus R).$$

Fix $\varepsilon > 0$, and let $\{Q(x_i, r_i)\}_{i\in\mathbb{N}}$ be a covering of F with open cubes such that

$$\mathcal{L}^N(F) + \varepsilon \geq \sum_{i\in\mathbb{N}} r_i^N. \tag{11.8}$$

For each $i \in \mathbb{N}$, let

$$A_i := Q_i \cap R, \qquad B_i := Q_i \setminus R.$$

Then, by using Lemma 11.4, we get that

$$\mathcal{L}^N(A_i) + \mathcal{L}^N(B_i) = \mathcal{L}^N(Q_i) = r_i^N,$$

where last equality follows from Lemma 11.3. Note that

$$F \cap R \subset \bigcup_{i\in\mathbb{N}} A_i, \qquad F \setminus R \subset \bigcup_{i\in\mathbb{N}} B_i. \tag{11.9}$$

The technical issue here is that the sets A_i are not necessarily open. To overcome this, we consider an open pluri-rectangle $\widetilde{A_i} \supset A_i$, with

$$\mathcal{L}^N(\widetilde{A_i}) \leq \mathcal{L}^N(A_i) + \frac{\varepsilon}{2^i}. \tag{11.10}$$

Therefore, by using (11.8), we get

$$\mathcal{L}^N(F) + \varepsilon \geq \sum_{i\in\mathbb{N}} r_i^N$$

$$
\begin{aligned}
&= \sum_{i\in\mathbb{N}} \mathcal{L}^N(\widetilde{A}_i) + \sum_{i\in\mathbb{N}} \mathcal{L}^N(B_i) \\
&\geq \sum_{i\in\mathbb{N}} \mathcal{L}^N(A_i) - \varepsilon + \sum_{i\in\mathbb{N}} \mathcal{L}^N(B_i) \\
&= \mathcal{L}^N\left(\bigcup_{i\in\mathbb{N}} A_i\right) - \varepsilon + \mathcal{L}^N\left(\bigcup_{i\in\mathbb{N}} B_i\right) \\
&\geq \mathcal{L}^N(F\cap R) + \mathcal{L}^N(F\setminus R) - \varepsilon,
\end{aligned}
$$

where in the second inequality we used (11.10), while the third inequality follows from sub-additivity and (11.9). Since $\varepsilon > 0$ is arbitrary, we conclude. □

As a consequence of the above result, we have that the family of Lebesgue measurable sets contains many sets.

Definition 11.25. A set $E \subset \mathbb{R}^N$ is said to be a *Borel set* if it can be written as countable union and countable intersection of open sets.

Lemma 11.26. *Borel sets are Lebesgue measurable.*

Proof. Lemma 11.24 gives us the Lebesgue measurability of rectangles (open, closed, or anything in between). Since by Lemma 11.9 every open set is a countable union of closed cubes, by using Proposition 11.19(iii), we get that every open set is Lebesgue measurable. By using 11.19(i), we get that every closed set is Lebesgue measurable. Thus, from 11.19(iii) again, we get that every countable union of countable intersection of open and closed sets is Lebesgue measurable. □

Remark 11.27. There are sets that are Lebesgue measurable, but not Borel. Their construction is a bit involved.

Moreover, it is easy to see that the following holds.

Lemma 11.28. *Peano-Jordan measurable sets are Lebesgue measurable.*

11.3. **Negligible sets.** Another class of sets of interest that is Lebesgue measurable is that of sets with zero outer Lebesgue measure.

Definition 11.29. We say that a set $E \subset \mathbb{R}^N$ is $\mathcal{L}^N$*-negligible* (or *negligible* with respect to the Lebesgue measure) if $\mathcal{L}^N(E) = 0$.

There is an easy way to determine whether a set is negligible.

Lemma 11.30. *Let $E \subset \mathbb{R}^N$. Then, followings are equivalent*

(i) For every $\varepsilon > 0$, there exists $\{E_i\}_{i\in\mathbb{N}} \subset \mathbb{R}^N$ such that

$$
E \subset \bigcup_{i=0}^{\infty} E_i, \qquad \mathcal{L}^N\left(\bigcup_{i=0}^{\infty} E_i\right) < \varepsilon,
$$

(ii) E is $\mathcal{L}^N$-negligible.

Lemma 11.31. *Let $E \subset \mathbb{R}^N$ be $\mathcal{L}^N$-negligible. Then, E is $\mathcal{L}^N$-measurable.*

The proofs are left as an exercise to the reader. We now introduce some terminology that will be useful later.

Definition 11.32. We say that a property holds for $\mathcal{L}^N$-almost every $x \in \mathbb{R}^N$ ($\mathcal{L}^N$-a.e. in $\mathbb{R}^N$) if it holds for all $x \in \mathbb{R}^N \setminus N$, where $N \subset \mathbb{R}^N$ is $\mathcal{L}^N$-negligible.

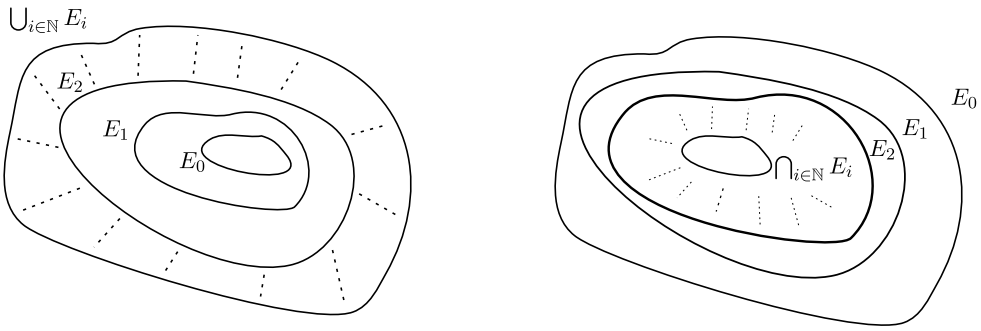

FIGURE 26. The limiting sets for increasing (left) and decreasing (right) sequences of sets.

11.4. **Operations on measurable sets.** In this section we investigate how the Lebesgue measure behaves with respect to limiting operations on Lebesgue measurable sets. Namely, we want to understand what happens when we take *monotone* sequences of sets: for such a sequence it is possible to define a notion of *limiting* set. Indeed, if the sequence of sets $\{E_i\}_{i\in\mathbb{N}} \subset \mathbb{R}^N$ is increasing, namely $E_i \subset E_{i+1}$ for all $i \in \mathbb{N}$, then the limiting set can be written as (see Figure 26 on the left)

$$\bigcup_{i\in\mathbb{N}} E_i.$$

Note that, for each $n \in \mathbb{N}$, by the monotonicity of the sequence, it holds that

$$\bigcup_{i=1}^{n} E_i = E_n.$$

We would like to understand the relation between the two measures

$$\mathcal{L}^N(E_n), \qquad \mathcal{L}^N\left(\bigcup_{i=0}^{\infty} E_i\right).$$

We would expect the first to converge to the latter. This is indeed the case if all the sets E_i's are Lebesgue measurable.

In a similar way, if we have a decreasing sequence $\{E_i\}_{i\in\mathbb{N}} \subset \mathbb{R}^N$, namely $E_{i+1} \subset E_i$ for all $i \in \mathbb{N}$, then the limiting set can be written as (see Figure 26 on the right)

$$\bigcap_{i\in\mathbb{N}} E_i.$$

Note that, for each $n \in \mathbb{N}$, by the monotonicity of the sequence, it holds that

$$\bigcap_{i=1}^{n} E_i = E_n.$$

We would like to understand the relation between the two measures

$$\mathcal{L}^N(E_n), \qquad \mathcal{L}^N\left(\bigcap_{i=0}^{\infty} E_i\right).$$

We would expect the first to converge to the latter. Also for this case, this is indeed true if all the sets E_i's are Lebesgue measurable.

In order to prove these two claims, we need to establish two properties of the Lebesgue measure on Lebesgue measurable sets: it is countably additive on sequences of pairwise disjoint sets (note that measurability is for *finitely* many sets), and it behaves nicely with respect to set difference.

Proposition 11.33. *Let $\{E_i\}_{i\in\mathbb{N}} \subset \mathbb{R}^N$ be $\mathcal{L}^N$-measurable sets. Then:*

(i) If the E_i's are pairwise disjoint, then

$$\mathcal{L}^N\left(\bigcup_{i=0}^{\infty} E_i\right) = \sum_{i=0}^{\infty} \mathcal{L}^N(E_i);$$

(ii) If $E_i \subset E_j$, then $\mathcal{L}^N(E_i) = \mathcal{L}^N(E_j) - \mathcal{L}^N(E_j \setminus E_i)$;

(iii) If $E_0 \subset E_1 \subset E_2 \subset \dots$, then

$$\lim_{i\to\infty} \mathcal{L}^N(E_i) = \mathcal{L}^N\left(\bigcup_{i=0}^{\infty} E_i\right);$$

(iv) If $E_0 \supset E_1 \supset E_2 \supset \dots$ and $\mathcal{L}^N(E_0) < \infty$, then

$$\lim_{i\to\infty} \mathcal{L}^N(E_i) = \mathcal{L}^N\left(\bigcap_{i=0}^{\infty} E_i\right).$$

Proof. First of all, note that all sets in the formulas are Lebesgue measurable, thanks to Proposition 11.19.

Step 1. We prove (i). In step 3 of the proof of Lemma 11.19 we proved that (see (11.5))

$$\mathcal{L}^N(A) \geq \sum_{i=0}^{n} \mathcal{L}^N\left(A \cap \bigcup_{i=0}^{n} E_i\right) + \mathcal{L}^N\left(A \setminus \bigcup_{i=0}^{n} E_i\right),$$

for all $n \in \mathbb{N}$, and all $A \subset \mathbb{R}^N$. By taking $A = \cup_{i=0}^{\infty} E_i$ sending $n \to \infty$, we get

$$\mathcal{L}^N\left(\bigcup_{i=0}^{\infty} E_i\right) \geq \sum_{i=0}^{\infty} \mathcal{L}^N(E_i).$$

The other inequality follows from the sub-additivity.

Step 2. We prove (ii). This follows directly from the measurability of E_j, by taking $A = E_i$.

Step 3. We prove (iii). Let $F_0 \coloneqq E_0$, and, for each $i \in \mathbb{N} \setminus \{0\}$, let

$$F_i \coloneqq E_i \setminus E_{i-1}.$$

Note that

$$\bigcup_{i=0}^{\infty} E_i = \bigcup_{i=0}^{\infty} F_i. \tag{11.11}$$

Then, since the sequence $\{E_i\}_{i\in\mathbb{N}}$ is increasing, we have that the sets F_i's are pairwise disjoint. Therefore, from step 1, we get

$$\mathcal{L}^N\left(\bigcup_{i=0}^{\infty} E_i\right) = \mathcal{L}^N\left(\bigcup_{i=0}^{\infty} F_i\right) = \sum_{i=0}^{\infty} \mathcal{L}^N(F_i) = \mathcal{L}^N(E_0) + \sum_{i=1}^{\infty} [\mathcal{L}^N(E_i) - \mathcal{L}^N(E_{i-i})],$$

where in the last step we used step (ii) together with (11.11). Now, note that, for each $n \in \mathbb{N}$,

$$\mathcal{L}^N(E_0) + \sum_{i=1}^{n} [\mathcal{L}^N(E_i) - \mathcal{L}^N(E_{i-i})] = \mathcal{L}^N(E_n).$$

Therefore,

$$\mathcal{L}^N\left(\bigcup_{i=0}^{\infty} E_i\right) = \lim_{n\to\infty} \mathcal{L}^N(E_n),$$

as desired.

Step 4. We prove (iv). For each $i \in \mathbb{N}$, let

$$F_i := E_0 \setminus E_i.$$

Note that

$$\bigcup_{i\in\mathbb{N}} F_i = E_0 \setminus \bigcap_{i\in\mathbb{N}} E_i, \tag{11.12}$$

and that $\cap_{i\in\mathbb{N}} E_i$ is Lebesgue measurable (see Proposition 11.19 (iii)). Moreover, since each E_i is Lebesgue measurable, we have that each F_i is Lebesgue measurable, and that, thanks to step 2, that

$$\mathcal{L}^N(F_i) = \mathcal{L}^N(E_0) - \mathcal{L}^N(E_i). \tag{11.13}$$

Using the fact that $\{E_i\}_{i\in\mathbb{N}}$ is a decreasing sequence, we get that $\{F_i\}_{i\in\mathbb{N}}$ is an increasing sequence. Therefore, from step 3 we get

$$\mathcal{L}^N\left(\bigcup_{i\in\mathbb{N}} F_i\right) = \lim_{i\to\infty} \mathcal{L}^N(F_i). \tag{11.14}$$

We now write the left-hand side and the right-hand side. From (11.12), we get

$$\mathcal{L}^N\left(\bigcup_{i\in\mathbb{N}} F_i\right) = \mathcal{L}^N\left(E_0 \setminus \bigcap_{i\in\mathbb{N}} E_i\right) = \mathcal{L}^N(E_0) - \mathcal{L}^N\left(\bigcap_{i\in\mathbb{N}} E_i\right),$$

where in the last step we used step 2. Now we write the right-hand side. From (11.13), together with the assumption that $\mathcal{L}^N(E_0) < \infty$, we have

$$\lim_{i\to\infty} \mathcal{L}^N(F_i) = \mathcal{L}^N(E_0) - \lim_{i\to\infty} \mathcal{L}^N(E_i).$$

Thus, (11.14) writes as

$$\mathcal{L}^N(E_0) - \mathcal{L}^N\left(\bigcap_{i=1}^{\infty} E_i\right) = \mathcal{L}^N(E_0) - \lim_{i\to\infty} \mathcal{L}^N(E_i).$$

Again by using the assumption that $\mathcal{L}^N(E_0) < \infty$, we conclude. □

Remark 11.34. Note that in both cases on the right-hand side we have the measure of the *limiting* set. In the second case the assumption $\mathcal{L}^N(E_1) < \infty$ is crucial. Indeed, consider the decreasing sequence of sets $E_i := (i, \infty)$. Then $\mathcal{L}^N(E_i) = \infty$ for each $i \in \mathbb{N}$, but

$$\mathcal{L}^N\left(\bigcap_{i\in\mathbb{N}} E_i\right) = 0,$$

since $\bigcap_{i\in\mathbb{N}} E_i = \emptyset$.

Remark 11.35. In particular, from Proposition 11.33(i), by using the fact that $\mathcal{L}^N(\{x\}) = 0$, for all $x \in \mathbb{R}^N$, we get that $\mathcal{L}^N(\mathbb{Q}^N) = 0$.

Remark 11.36. Note that countably additivity is the best we can hope for in order to get something that makes sense. Indeed, if we were to ask for

$$\mathcal{L}^N\left(\bigcup_{i\in I} E_i\right) = \sum_{i\in I} \mathcal{L}^N(E_i) \tag{11.15}$$

for *any* family of indexes I, even more than countable, we would have something nonsense. Indeed, consider the unit cube $Q = (0,1)^2$ in the plane. From (11.15), we would get

$$1 = \mathcal{L}^2(Q) = \mathcal{L}^2\left(\bigcup_{x\in Q} \{x\}\right) = \sum_{x\in Q} \mathcal{L}^2(\{x\}) = 0,$$

where the last equality follows from the fact that point $x \in Q$ has zero area.

11.5. **Non-measurable sets.** In the previous section we introduced the notion of Lebesgue measurable sets, and proved that the Lebesgue measure behaves well with respect to set operations on this class. The question is now: is there any set that is not Lebesgue measurable? The answer depends on whether or not the Axiom of Choice is accepted. Indeed, it was proved by Solovay in 1970 that only by using the Zermelo-Fraenkel (choice) of axioms for set theory (ZFC), but not the Axiom of Choice, all sets are Lebesgue measurable. On the other hand, if we use ZFC together with the Axiom of Choice, it is possible to construct several examples of sets that are not Lebesgue measurable. The first was obtained by Vitali in 1905.

Definition 11.37 (Axiom of Choice)**.** The *Axiom of Choice* claims the following: given any family of sets $\{E_\alpha\}_\alpha$ (even more than countable!), there exists a set E which has exactly an element from each of the sets E_α.

Theorem 11.38 (Vitali's Theorem)**.** *There exists a set $E \subset [0,1]$ that is not Lebesgue measurable.*

Proof. On $[0,1]$ consider the equivalence relation $\sim$ defined as follows

$$x \sim y \quad \Leftrightarrow \quad x - y \in \mathbb{Q}.$$

Then, thanks to the Axiom of Choice, it is possible to construct a set E by picking an element from each equivalent class of $[0,1]/\sim$. We note two properties of the set E. First of all, since E contains an element from each equivalent class of $[0,1]/\sim$, given any $x \in [0,1]$, there exist $q \in \mathbb{Q} \cap [-1,1]$ and $y \in E$ such that $x + q = y$. This means that

$$[0,1] \subset \bigcup_{q \in \mathbb{Q} \cap [-1,1]} (q + E) \subset [-1,2]. \tag{11.16}$$

where the inclusion on the right follows from the fact that $E \subset [0,1]$.

We claim that E is not Lebesgue measurable. Indeed, assume by contradiction that it is. Then, by using Proposition 11.19(iv), we have that $q + E$ is Lebesgue measurable, and, by translation invariance (see Lemma11.5(iv)), we have that

$$\mathcal{L}^1(q + E) = \mathcal{L}^1(E),$$

for all $q \in \mathbb{Q}$. Therefore, since $q + E \cap s + E = \emptyset$ if $q \neq s$, from Proposition 11.33(i), we have

$$\mathcal{L}^1\left(\bigcup_{q \in \mathbb{Q} \cap [-1,1]} (q + E)\right) = \sum_{q \in \mathbb{Q} \cap [-1,1]} \mathcal{L}^1(q + E) = \sum_{q \in \mathbb{Q} \cap [-1,1]} \mathcal{L}^1(E). \tag{11.17}$$

Now, from the first inclusion in (11.16), together with monotonicity (see Lemma 11.5(i)) we get that

$$\mathcal{L}^1(E) \geq \mathcal{L}^1([0,1]) = 1.$$

Thus, from (11.17), we get that

$$\mathcal{L}^1\left(\bigcup_{q \in \mathbb{Q} \cap [-1,1]} (q + E)\right) = +\infty.$$

This is in contradiction with the second inclusion in (11.16), since it implies that

$$\mathcal{L}^1\left(\bigcup_{q \in \mathbb{Q} \cap [-1,1]} (q + E)\right) \leq \mathcal{L}^1([-1,2]) = 3.$$

This gives the desired contradiction. □

The essence of the above construction is that the set E is so intertwined, that it is not possible to compute the Lebesgue measure of it as the sum of the Lebesgue measure of two disjoint pieces. By using the above result, it is possible to obtain the following.

Lemma 11.39. *Let $E \subset \mathbb{R}^N$ be a set with $\mathcal{L}^N(E) > 0$. Then, there exists $F \subset E$ such that F is not Lebesgue measurable.*

Remark 11.40. In particular, the above example shows that it is *not* possible to extend the Lebesgue measure to *all* sets of $\mathbb{R}^N$ in a way that preserves both countably additivity and translation invariance. It is nevertheless possible to extend the Lebesgue measure to *all* sets in such a way that finitely additivity holds (but not translation invariance). Finally, whether or not it is possible to extend the Lebesgue measure to all sets in such a way that translation invariance holds, is something that is not decidable by using ZFC.

Even more surprisingly, we can even ask ourselves what happens if also consider isometries: does the Lebesgue measure behave nicely with respect to isometries on *all* sets? The answer is no! And in this case we have a mind-blowing example of how badly things can be! This is the so called *Banach-Tarski paradox* of 1924.

Theorem 11.41 (Banach-Tarski's paradox)**.** *Let $E \subset \mathbb{R}^3$ be the unit ball. Then. there exist disjoint sets $E_1, E_2, E_3, E_4, E_5 \subset E$, with*

$$E = \bigcup_{i=1}^{5} E_i,$$

with the following property: it is possible to translate and rotate each E_i to obtain a set $F_i \subset \mathbb{R}^3$, such that

$$\mathcal{L}^3(F_1 \cup F_2 \cup F_3) = \mathcal{L}^3(F_4 \cup F_5) = \mathcal{L}^3(E).$$

11.6. **Vitali's characterization of Riemann integrability.**

Theorem 11.42 (Vitali's Theorem)**.** *Let $D \subset \mathbb{R}^N$ be a rectangle. A function $f : D \to \mathbb{R}^M$ is Riemann integrable if and only if the set of discontinuities of f has Lebesgue measure zero.*

Proof. By Theorem 10.43, we have that f is Riemann integrable if and only if $\mathcal{PJ}^+(\Omega(f, \varepsilon)) = 0$, for all $\varepsilon > 0$, where

$$\Omega(f, \varepsilon) \coloneqq \{ x \in D \,:\, \omega_f(x) \geq \varepsilon \},$$

and the oscillation of f at x is given by

$$\omega_f(x) \coloneqq \inf_{r>0} \{ \sup |f(y) - f(z)| \,:\, y, z \in B(x, r) \}.$$

By Lemma 10.42, the set $\Omega(f, \varepsilon)$ is closed, and thus, by Lemma 11.26, Lebesgue measurable. Moreover, note that the set S of discontinuities of f can be written as

$$S = \bigcap_{n \in \mathbb{N} \setminus \{0\}} \Omega(f, 1/n), \tag{11.18}$$

and that the sequence of sets $\{\Omega(f, 1/n)\}_{n \in \mathbb{N} \setminus \{0\}}$ is increasing.

We are now in position to prove the result. Assume that f is Riemann integrable. Then, $\mathcal{PJ}^+(\Omega(f, 1/n)) = 0$, for all $n \in \mathbb{N} \setminus \{0\}$. Since

$$\mathcal{L}^N(\Omega(f, 1/n)) \leq \mathcal{PJ}^+(\Omega(f, 1/n)),$$

we get that

$$\mathcal{L}^N(\Omega(f, 1/n)) = 0,$$

for all $n \in \mathbb{N} \setminus \{0\}$. Thus, from(11.18) together with Proposition 11.19, we get that $\mathcal{L}^N(S) = 0$.

Assume now that $\mathcal{L}^N(S) = 0$. Since, for all $n \in \mathbb{N} \setminus \{0\}$, $\Omega(f, 1/n) \subset S$, we get that $\mathcal{L}^N(\Omega(f, 1/n)) = 0$. Since $\Omega(f, 1/n) \subset D$, and thus it is bounded, by Theorem 11.13, we get that

$$\mathcal{PJ}^+(\Omega(f, 1/n)) = \mathcal{L}^N(\Omega(f, 1/n)) = 0,$$

where we used the fact that $\Omega(f, 1/n)$ is closed. Thus, from Theorem 10.43 we get that f is Riemann integrable. □

Remark 11.43. In particular, Theorem 11.42 together with Remark 11.35 implies that the Dirichlet function (10.1) is *not* Riemann integrable because its set of discontinuities is $[0, 1]$, and it is not negligible with respect to the Lebesgue measure.

12. Lebesgue integration

Nature laughs at the difficulty of integration.
Pier-Simon Laplace

The Cauchy-Darboux-Riemann notion of integration can be seen as a geometric notion based on the Peano-Jordan content. Indeed, we have seen in Theorem10.38 that a function $f : D \to \mathbb{R}$, where $D \subset \mathbb{R}^N$ is a rectangle, is Riemann integrable if and only if the sets

$$E^+ := \{(x, y) \in D \times [0, \infty) \, : \, 0 \leq y \leq f(x)\},$$

and

$$E^- := \{(x, y) \in D \times (-\infty, 0] \, : \, f(x) \leq y \leq 0\}$$

are Peano-Jordan measurable. In this case, it holds

$$\int_D f(x)\, dx = \mathcal{PJ}(E^+) - \mathcal{PJ}(E^-).$$

This means that the Riemann integral is the *area/volume* under the graph, where the *area/volume* is computed by using the Peano-Jordan content. We have also seen that (see Theorem 10.43 f is Riemann integrable if and only if it is bounded and $\mathcal{PJ}^+(\Omega(f, \varepsilon)) = 0$, for all $\varepsilon > 0$. We recall that

$$\Omega(f, \varepsilon) := \{\, x \in D \, : \, \omega_f(x) \geq \varepsilon \,\}.$$

where, for $x \in D$, the oscillation of f at x is defined as

$$\omega_f(x) = \lim_{r \to 0} \left| \sup_{B(x,r)} f - \inf_{B(x,r)} f \right|.$$

We now want to comment on this condition and to understand how to change the definition of the integral in such a way to overcome the limitation of the Riemann integration. The reason why the above characterization of Riemann integrability holds is the following: the Riemann integral is defined by using approximation by piecewise constant function. In order for this approximation to work, up to a small set (namely, of a set of Peano-Jordan measure zero), we need f to be continuous in *most* of D. This is because the piecewise constant functions used in the approximation are defined as follows:

(i) First, we consider a partition of the domain D;
(ii) Then, we assign values on each component of partition.

For a generic function $f : D \to \mathbb{R}$, there is no reason why such piecewise constant functions should give a good approximation from above and from below of the function itself. How to then build a piecewise constant function that is a good approximation of a given f? Well, we just change the order in which we construct such a piecewise constant function: namely, we first select the values that we want to assign (step (ii)), and then we understand on which set to assign the value (step (i)).

Namely, we first partition the target space $\mathbb{R}$ into intervals $[y_i, y_{i+1})$, for $i \in \mathbb{N}$. Then, we consider the pre-images of each of those intervals

$$E_i := f^{-1}([y_i, y_{i+1})).$$

We define piecewise constant functions $g, h : D \to \mathbb{R}$ as follows

$$g(x) = y_i \text{ on } E_i, \qquad h(x) = y_{i+1} \text{ on } E_i.$$

Thus, we have $g \leq f \leq h$, and the idea is to use these type of piecewise constant functions to define the lower and the upper integral. The integral of such a piecewise constant function will be

$$\int_D g(x)\, dx = \sum_{i \in \mathbb{N}} y_i \mathcal{L}^N(E_i), \qquad \int_D g(x)\, dx = \sum_{i \in \mathbb{N}} y_{i+1} \mathcal{L}^N(E_i).$$

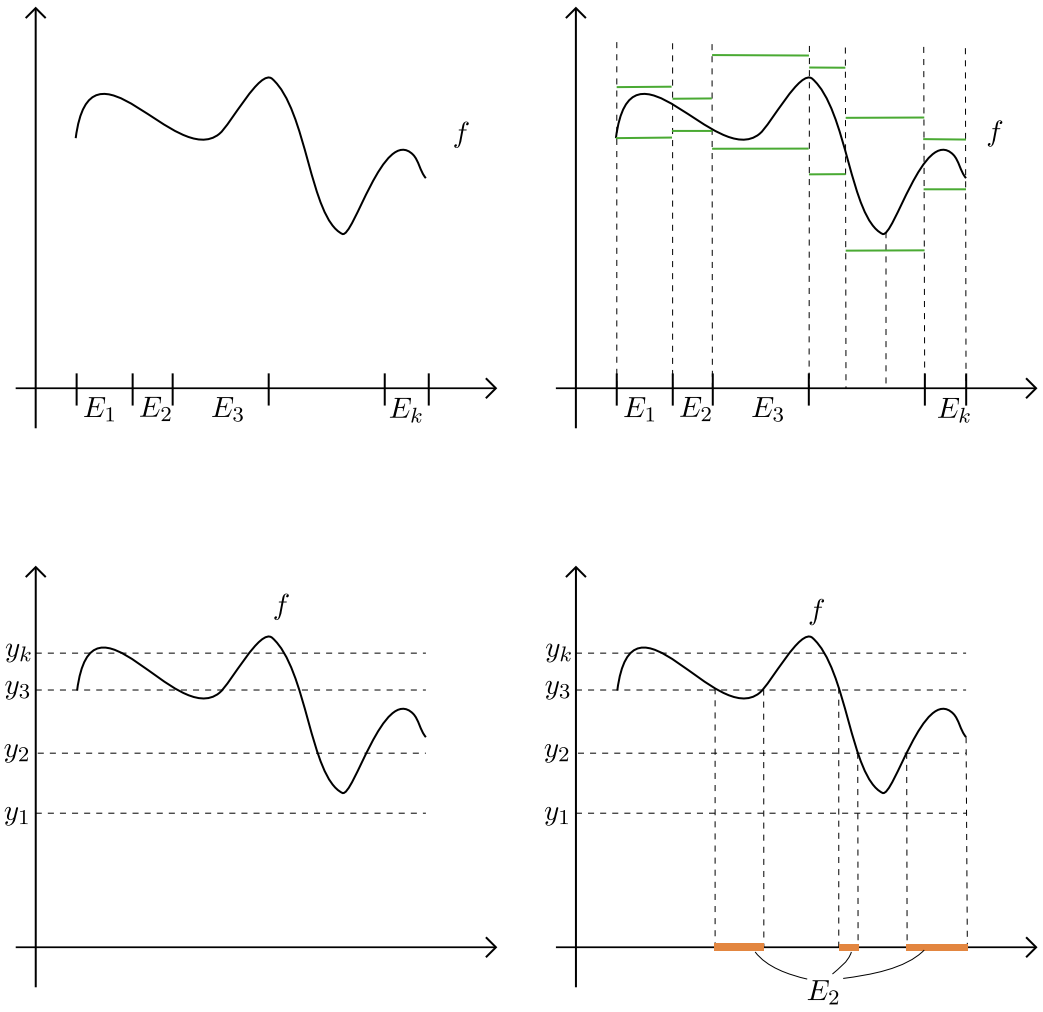

FIGURE 27. The paradigm shift from the Riemann integral (on the top) and the Lebesgue's one (on the bottom). For the Riemann integral, we first partition the domain into set $E_1, \dots, E_k$, and then we assign values on each of those sets. For the Lebesgue integral, instead, we first partition the target space into intervals $[y_i y_{i+1})$, and the pre-images of each of those intervals will be the set where we define the piecewise constant function to be y_i. For instance, in the figure on the right, the set E_2 (depicted in orange) is the set where f is in between y_2 and y_3, and the piecewise constant function is y_2.

So far, it seems that no condition is needed on the function f. Nevertheless, there is a questions of the quantity above to be well defined that requires a bit of care. Indeed, let us suppose that we write a piecewise constant function in two ways

$$f(x) = \sum_{i=1}^{k} \mathbb{1}_{E_i}(x) y_i = \sum_{j=1}^{m} \mathbb{1}_{F_j}(x) z_j.$$

We would of course like to have

$$\sum_{i=1}^{k} y_i \mathcal{L}^N(E_i) = \sum_{i=j}^{m} z_j \mathcal{L}^N(F_j)\,,$$

for *any* choice of the sets F_j's. This is precisely equivalent to require that the sets E_i's are $\mathcal{L}^N$-measurable. Thus, for the above strategy to define the integral to work, we need to make sure that each of the sets

$$E_i = f^{-1}([y_i, y_{i+1}))$$

is Lebesgue measurable. This gives the notion of *Lebesgue measurable functions*.

12.1. **Lebesgue measurable functions.** The theory of Lebesgue is powerful enough also to treat functions that take the value $\pm\infty$. In the following, we will use the following notation

$$\overline{\mathbb{R}} \coloneqq \mathbb{R} \cup \{\pm\infty\}.$$

We introduce the concept of Lebesgue measurable function.

Definition 12.1. A function $f : \mathbb{R}^N \to \overline{\mathbb{R}}$ is said to be *Lebesgue measurable* if

$$f^{-1}((a, +\infty)) \coloneqq \{x \in \mathbb{R}^M : f(x) \in (a, +\infty)\}$$

is Lebesgue measurable, for all $a \in \mathbb{R}$.

Remark 12.2. Note that, given a set $E \subset \mathbb{R}^N$, the function $\mathbb{1}_E$ is Lebesgue measurable if and only if E is Lebesgue measurable.

Exercise 12.3. For each $a \in \mathbb{R}$, write the set $f^{-1}((a, +\infty))$ for the following functions:

(i) $f : \mathbb{R} \to \mathbb{R}$ defined as $f(x) \coloneqq x^2$;
(ii) $f : \mathbb{R}^2 \to \mathbb{R}$ defined as $f(x, y) \coloneqq xy$;
(iii) The Dirichlet function $f : \mathbb{R} \to \mathbb{R}$ defined as $f(x) \coloneqq \mathbb{1}_{\mathbb{Q}}$;
(iv) The function $f \coloneqq \mathbb{1}_{(0,1)}$;
(v) The function $f \coloneqq \mathbb{1}_{[0,1]}$.

Are those sets Lebesgue measurable?

First of all, we show some equivalent conditions for a function to be Lebesgue measurable.

Lemma 12.4. *Let $f : \mathbb{R}^N \to \overline{\mathbb{R}}$. Then, followings are equivalent:*

(i) f is Lebesgue measurable;
(ii) $f^{-1}([a, +\infty))$ is Lebesgue measurable, for all $a \in \mathbb{R}$;
(iii) $f^{-1}((-\infty, a))$ is Lebesgue measurable, for all $a \in \mathbb{R}$;
(iv) $f^{-1}((-\infty, a])$ is Lebesgue measurable, for all $a \in \mathbb{R}$;
(v) $f^{-1}(B)$ is Lebesgue measurable, for all Borel sets $B \subset \mathbb{R}$, and $f^{-1}(\{+\infty\})$, $f^{-1}(\{-\infty\})$ are Lebesgue measurable.

The proof is left as an exercise to the reader, since it boils down to write each of the sets in an equivalent condition as a countable union or countable intersection of sets of another condition, and use the properties of Lebesgue measurable sets (see Proposition 11.19) and of the Lebesgue measure on them (see Proposition 11.33).

Remark 12.5. We want to draw a parallelism between the definition of a Lebesgue measurable function and that of a continuous function. We have seen in Lemma 3.10 that for a continuous function $f : \mathbb{R}^N \to \mathbb{R}$, the sets

$$f^{-1}((a, +\infty))$$

are open, for all $a \in \mathbb{R}$. It is possible to prove that this property characterizes continuity. Thus, Lebesgue measurable functions are functions for which we relax the constrain of what kind of regularity we ask for the sets $f^{-1}((a, +\infty))$.

As a consequence of the above remark, we get that.

Lemma 12.6. *Let $f : \mathbb{R}^N \to \mathbb{R}$ be a continuous function. Then, f is Lebesgue measurable.*

Moreover, we have another class of functions for which it is easy to prove that they are Lebesgue measurable.

Lemma 12.7. *A monotone function $f : \mathbb{R} \to \mathbb{R}$ is Lebesgue measurable.*

Proof. Assume f is increasing. Let $a \in \mathbb{R}$. Then, the set $f^{-1}((-\infty, a))$ is either empty, or of the form $(-\infty, b)$ or $(-\infty, b]$, for some $b \in \mathbb{R}$. All of these sets are Lebesgue measurable. □

So far, we only know that continuous and monotone functions are Lebesgue measurable. Given a function f, obtained by performing countably many operations with functions f_n's, we would like to infer that, if each of the functions f_n's is Lebesgue measurable, also f is Lebesgue measurable. This is what we are going to prove now.

Definition 12.8. Let $f, g : \mathbb{R}^N \to \overline{\mathbb{R}}$. We define $\max\{f, g\}, \min\{f, g\} : \mathbb{R}^N \to \overline{\mathbb{R}}$ as

$$\max\{f, g\}(x) \coloneqq \max\{f(x), g(x)\},$$
$$\min\{f, g\}(x) \coloneqq \min\{f(x), g(x)\}.$$

Remark 12.9. An alternative notation (inspired by *Logic*) for the max and the min between two functions is the following

$$\max\{f, g\} = f \vee g, \qquad \min\{f, g\} = f \wedge g.$$

The symbols $\wedge$ is called *AND*, while the symbol $\vee$ is called *OR*. The reason for the notation is the following: if you assign the value 1 at a statement S_1 if the statement S_2 is true, and the value 0 if the sentence is false, then the statement

$$S_1 \text{ and } S_2,$$

is true if both S_1 and S_2 are true. Thus, the value we assign to such a statement is the minimum between the values assigned to the statements S_1 and S_2. Moreover, if we consider the statement

$$S_1 \text{ or } S_2$$

we get that it is true if at least one of the statements S_1 and S_2 is true. Thus, the value we assign to such a statement is the maximum between the values assigned to the statements S_1 and S_2.

Definition 12.10. Let $(f_n)_{n\in\mathbb{N}}$ be functions $f_n : \mathbb{R}^N \to \overline{\mathbb{R}}$. We define the functions

$$\sup_{n\in\mathbb{N}} f_n, \ \inf_{n\in\mathbb{N}} f_n : \mathbb{R}^N \to \overline{\mathbb{R}}$$

as

$$\sup_{n\in\mathbb{N}} f_n(x) \coloneqq \sup\{f_n(x) : n \in \mathbb{N}\},$$

and

$$\inf_{n\in\mathbb{N}} f_n(x) \coloneqq \inf\{f_n(x) : n \in \mathbb{N}\}.$$

Moreover, we define $\liminf_{n\to\infty} f_n, \limsup_{n\to\infty} f_n : \mathbb{R}^N \to \overline{\mathbb{R}}$ as

$$\limsup_{n\to\infty} f_n(x) \coloneqq \limsup_{n\to\infty} (f_n(x))_{n\in\mathbb{N}} \coloneqq \inf_{i\to\infty} \sup_{j\geq i} f_i(x),$$

and

$$\liminf_{n\to\infty} f_n(x) \coloneqq \liminf_{n\to\infty} (f_n(x))_{n\in\mathbb{N}} \coloneqq \sup_{i\to\infty} \inf_{j\geq i} f_i(x).$$

Definition 12.11. Given a function $f : \mathbb{R}^N \to \overline{\mathbb{R}}$, we define its *positive* and *negative* part as

$$f^+ \coloneqq \max\{f, 0\}, \qquad f^- \coloneqq \max\{-f, 0\},$$

respectively.

Remark 12.12. Note that $f^+, f^- \geq 0$, and that $f = f^+ - f^-$, $|f| = f^+ + f^-$.

Proposition 12.13. *Let $f, g : \mathbb{R}^N \to \overline{\mathbb{R}}$ be Lebesgue measurable functions. Then, the functions*

$$|f|, \qquad f^+, \qquad f^-, \qquad f + g, \qquad fg, \qquad \max\{f, g\}, \qquad \min\{f, g\}$$

are Lebesgue measurable.

Proof. Step 1. We first prove that $f+g$ is Lebesgue measurable. Let $a \in \mathbb{R}$. If $x \in \mathbb{R}$ is such that $f(x)+g(x) > a$, then, there exists $q \in \mathbb{Q}$ such that $f(x)+g(x) > q > a$. Therefore, we can write

$$(f+g)^{-1}((a,+\infty)) = \bigcup_{\substack{r+s>a \\ r,s\in\mathbb{Q}}} \left[f^{-1}((r,+\infty)) \cap g^{-1}((s,+\infty))\right].$$

By the Lebesgue measurability of f and g, we get that

$$f^{-1}((r,+\infty)), \qquad g^{-1}((s,+\infty))$$

are Lebesgue measurable for each $r,s \in \mathbb{Q}$. Thus, by using the properties of intersection and countable union of Lebesgue measurable sets (see Proposition 11.19(iii)), we get that $(f+g)^{-1}((a,+\infty))$ is Lebesgue measurable.

Step 2. Now we prove that, if f is Lebesgue measurable, also f^2 is. This follows by writing, for $a \geq 0$

$$(f^2)^{-1}((a,\infty)) = f^{-1}(a^{\frac{1}{2}},+\infty),$$

while for $a < 0$ it holds $(f^2)^{-1}((a,\infty)) = \mathbb{R}^N$.

Step 3. We now prove that fg is Lebesgue measurable. This follows from the writing

$$fg = \frac{(f+g)^2 - f^2 - g^2}{2},$$

together with step 1, 2.

Step 4. Now, we prove that f^+ and f^- are measurable. Indeed,

$$f^+(x) = f(x)\mathbb{1}_{E^+}, \qquad f^-(x) = -f(x)\mathbb{1}_{E^-}$$

where $E^+ := \{x \in \mathbb{R}^N : f(x) \geq 0\}$, and $E^- := \{x \in \mathbb{R}^N : f(x) \geq 0\}$. These two sets are Lebesgue measurable, and thus also $\mathbb{1}_{E^+}$ and $\mathbb{1}_{E^-}$.

Step 5. Finally

$$|f| = f^+ + f^-, \qquad \max\{f,g\} = (f-g)^+ + g, \qquad \min\{f,g\} = -(f-g)^- + g.$$

Thus, by using the previous steps, they are all Lebesgue measurable. □

Proposition 12.14. *Let $(f_n)_{n\in\mathbb{N}}$ be Lebesgue measurable functions. Then, the functions*

$$\sup_{n\in\mathbb{N}} f_n, \qquad \inf_{n\in\mathbb{N}} f_n, \qquad \liminf_{n\to\infty} f_n, \qquad \limsup_{n\to\infty} f_n$$

are Lebesgue measurable.

Proof. By using the writings

$$\left(\sup_{n\in\mathbb{N}} f_n\right)^{-1}((a,+\infty)) = \bigcup_{n\in\mathbb{N}} f_n^{-1}((a,+\infty)),$$

$$\left(\inf_{n\in\mathbb{N}} f_n\right)^{-1}((a,+\infty)) = \bigcap_{n\in\mathbb{N}} f_n^{-1}((a,+\infty)),$$

for all $a \in \mathbb{R}$, we get that $\sup_{n\in\mathbb{N}} f_n$ and $\inf_{n\in\mathbb{N}} f_n$ are Lebesgue measurable. Moreover, by definition,

$$\limsup_{n\to\infty} f_n = \inf_{i\to\infty} \sup_{j\geq i} f_i(x),$$

and

$$\liminf_{n\to\infty} f_n = \sup_{i\to\infty} \inf_{j\geq i} f_i(x),$$

and thus are Lebesgue measurable. □

Other operations with Lebesgue measurable functions require a bit more care, in particular the composition of functions. What can be surprising at first (but clear if you look at the definition!), is that the composition of Lebesgue measurable functions is not necessarily Lebesgue measurable! Indeed, let $f : \mathbb{R}^N \to \mathbb{R}$, and $g : \mathbb{R} \to \mathbb{R}$ be two Lebesgue measurable functions. We wonder whether or not the composition $g \circ f : \mathbb{R}^N \to \mathbb{R}$ is Lebesgue measurable. Let's take $a \in \mathbb{R}$, and consider the set

$$(g \circ f)^{-1}((a, +\infty)) = \{x \in \mathbb{R}^N : (g \circ f)(x) > a\} = f^{-1}(g^{-1}((a, +\infty))).$$

Now, we know that $g^{-1}((a, +\infty))$ is Lebesgue measurable, thanks to the Lebesgue measurability of g. But we do not know anything about $f^{-1}(g^{-1}((a, +\infty)))$, since $g^{-1}((a, +\infty))$ might not be a Borel set (see Remark 11.27), and thus the Lebesgue measurability of f gives no condition on the regularity of the set $f^{-1}(g^{-1}((a, +\infty)))$. Nevertheless, by using the above argument, together with the discussion of the characterization of continuity in Remark 12.5, we get the following result.

Lemma 12.15. *Let $g : \mathbb{R}^N \to \mathbb{R}$ be a continuous function, and let $f : \mathbb{R} \to \mathbb{R}$ be a Lebesgue measurable function. Then, $g \circ f$ is Lebesgue measurable.*

Moreover, not only sub, super, and level sets of a Lebesgue measurable functions are measurable, but also when comparing two Lebesgue measurable functions.

Lemma 12.16. *Let $f, g : \mathbb{R}^N \to \mathbb{R}$ be Lebesgue measurable functions. Then, the sets*

$$\{x \in \mathbb{R}^N : f(x) > g(x)\}, \qquad \{x \in \mathbb{R}^N : f(x) \geq g(x)\}, \qquad \{x \in \mathbb{R}^N : f(x) = g(x)\}$$

are Lebesgue measurable.

We now study an important class of $\mathcal{L}^N$-measurable functions, that will later be used to define the Lebesgue integral.

Definition 12.17. We say that an $\mathcal{L}^N$-measurable function $f : \mathbb{R}^N \to \mathbb{R}$ is *simple* if its image is finite. Namely, if it is possible to write

$$f(x) = \sum_{i=0}^{k} \mathbb{1}_{E_i}(x) y_i$$

for all $x \in \mathbb{R}^N$, where $E_i \subset \mathbb{R}^N$, and $y_i \in \mathbb{R}$, for all $i = 1, \dots, k$.

Remark 12.18. Note that, in the definition, we are not requiring the sets E_i's to be pairwise disjoint. Nevertheless, it is always possible to write a simple function $f : \mathbb{R}^N \to \mathbb{R}$ as

$$f(x) = \sum_{i=0}^{k} \mathbb{1}_{F_i}(x) z_i,$$

where the sets F_i's are pairwise disjoint (prove it!).

Remark 12.19. Note that simple functions have only *finite* values, namely $\pm\infty$ are not allowed!

It is possible to characterize Lebesgue measurable functions as pointwise limit of simple functions.

Lemma 12.20. *Let $f : \mathbb{R}^N \to \mathbb{R}$. Then, f is Lebesgue measurable if and only if there exists a sequence of simple functions $(f_n)_{n \in \mathbb{N}}$ such that $f_n \to f$ pointwise.*

In particular, if $f \geq 0$, it is possible to construct the sequence in such a way that

$$0 \leq f_1 \leq f_2, \leq \dots \leq f.$$

Finally, note that the convergence is uniform on every set where f is bounded from above.

The proof is left as an exercise to the reader, since the strategy is already presented in the heuristic argument at the beginning of this chapter.

12.2. **The Lebesgue integral for $\mathcal{L}^N$-measurable functions.** In this section we make rigorous the heuristic we made at the introduction of this lecture. There is here an asymmetry with the theory of Riemann integration: it is possible to define the Lebesgue integral also for functions that are not Lebesgue measurable. This will be done by approximating it from above and from below by (countably-simple) $\mathcal{L}^N$-measurable functions. For $\mathcal{L}^N$-measurable functions this procedure is equivalent to a simpler definition. Thus, in this section we will only consider $\mathcal{L}^N$-measurable functions, while in next section we extend the notion of integral to non $\mathcal{L}^N$-measurable functions.

The definition of the Lebesgue integral for $\mathcal{L}^N$-measurable functions is by approximating *from below* with countably-simple functions.

Definition 12.21 (Integral of a positive simple function)**.** Let $f : \mathbb{R}^N \to [0, +\infty]$ be a simple function

$$f(x) = \sum_{i=0}^{k} \mathbb{1}_{E_i}(x) y_i \,.$$

We define the *Lebesgue integral* of f as

$$\int_{\mathbb{R}^N} f(x) \, \mathrm{d}\mathcal{L}^N(x) \coloneqq \sum_{i=0}^{k} \mathcal{L}^N(E_i) y_i,$$

with the convention that if $y_i = 0$ and $\mathcal{L}^N(E_i) = \infty$, then $y_i \mathcal{L}^N(E_i) = 0$.

Remark 12.22. Note that the $\mathcal{L}^N$-measurability of the simple function (that holds by definition!) is needed in order to have a well-defined object. Indeed, if we write a simple function f in two different ways,

$$f(x) = \sum_{i=1}^{k} \mathbb{1}_{E_i}(x) y_i = \sum_{j=1}^{m} \mathbb{1}_{F_j}(x) z_j,$$

then, we would like the integral to be well defined. Namely, that

$$\sum_{i=1}^{k} \mathcal{L}^N(E_i) y_i = \sum_{i=j}^{m} \mathcal{L}^N(F_j) z_j \,,$$

for *any* choice of the sets F_j's. This requirement is equivalent to having the sets E_i's $\mathcal{L}^N$-measurable.

Remark 12.23. An important fact to highlight is that the Riemann and the Lebesgue integral coincide on the class of functions that are piecewise constant in the sense of Definition 10.29.

Definition 12.24 (Integral of a positive function)**.** Let $f : \mathbb{R}^N \to [0, \infty]$ be $\mathcal{L}^N$-measurable. We define the *Lebesgue integral* of f as

$$\int_{\mathbb{R}^N} f(x) \, \mathrm{d}\mathcal{L}^N(x) \coloneqq \sup \left\{ \int_{\mathbb{R}^N} g(x) \, \mathrm{d}\mathcal{L}^N(x) \, : \, g \text{ simple}, \; g \leq f \quad \mathcal{L}^N - \text{a.e.} \right\}.$$

Remark 12.25. Note that we could have given an equivalent definition by requiring that $g \leq f$ *everywhere*. Indeed, since $f \geq 0$, given a simple function $g : \mathbb{R}^N \to [0, \infty)$ such that $g \leq f$ $\mathcal{L}^N$-a.e., we can construct a simple function $\widetilde{g} : \mathbb{R}^N \to [0, \infty)$ such that

$$g \leq f \text{ everywhere}, \qquad \int_{\mathbb{R}^N} g(x) \, \mathcal{L}^N(x) = \int_{\mathbb{R}^N} \widetilde{g}(x) \, \mathcal{L}^N(x).$$

Indeed, set $\widetilde{g} \coloneqq g \mathbb{1}_N$, where

$$N \coloneqq \{x \in \mathbb{R}^N \, : \, g(x) > f(x)\}.$$

By assumption, $\mathcal{L}^N(N) = 0$. Thus, by Lemma 11.31, N is Lebesgue measurable, and thus $\widetilde{g}$ is a simple function satisfying the required properties.

First of all, note that this definition is compatible with that previously given for countably-simple functions. The proof is left as an exercise to the reader.

Lemma 12.26. *Let $f : \mathbb{R}^N \to [0, +\infty]$ be a simple function*

$$f(x) = \sum_{i=0}^{k} \mathbb{1}_{E_i}(x) y_i.$$

Then,

$$\sum_{i=0}^{k} \mathcal{L}^N(E_i) y_i = \sup \left\{ \int_{\mathbb{R}^N} g(x) \, \mathrm{d}\mathcal{L}^N(x) \, : \, g \text{ simple}, \; g \le f \;\; \mathcal{L}^N - a.e. \right\}.$$

We now extend the definition of the Lebesgue integral to a generic $\mathcal{L}^N$-measurable function.

Definition 12.27 (Integral of a generic $\mathcal{L}^N$-measurable function)**.** Let $f : \mathbb{R}^N \to \overline{\mathbb{R}}$ be an $\mathcal{L}^N$-measurable function. Assume that

$$\int_{\mathbb{R}^N} f^+ \, \mathrm{d}\mathcal{L}^N(x) < \infty, \qquad \text{or} \qquad \int_{\mathbb{R}^N} f^- \, \mathrm{d}\mathcal{L}^N(x) < \infty. \tag{12.1}$$

We define the *Lebesgue integral* of f as

$$\int_{\mathbb{R}^N} f(x) \, \mathrm{d}\mathcal{L}^N(x) := \int_{\mathbb{R}^N} f^+(x) \, \mathrm{d}\mathcal{L}^N(x) - \int_{\mathbb{R}^N} f^-(x) \, \mathrm{d}\mathcal{L}^N(x).$$

A function satisfying condition (12.1) is called *integrable*.

Remark 12.28. Assumption (12.1) is in order to avoid $+\infty - \infty$ in the definition of the integral.

Finally, we define the Lebesgue integral of an $\mathcal{L}^N$-measurable function over a generic $\mathcal{L}^N$-measurable set.

Definition 12.29. Let $f : \mathbb{R}^N \to \overline{\mathbb{R}}$ be an $\mathcal{L}^N$-measurable function, and let $E \subset \mathbb{R}^N$ be $\mathcal{L}^N$-measurable. We define

$$\int_E f(x) \, \mathrm{d}\mathcal{L}^N(x) := \int_{\mathbb{R}^N} \mathbb{1}_E(x) f(x) \, \mathrm{d}\mathcal{L}^N(x).$$

Note that the function on the right-hand side is $\mathcal{L}^N$-measurable.

Remark 12.30. In the following we will write the results by considering $\mathcal{L}^N$-measurable functions $f : \mathbb{R}^N \to \mathbb{R}$ integrated in the entire space $\mathbb{R}^N$. Thanks to the above definition, the results hold true also when integrating over a measurable set $E \subset \mathbb{R}^N$.

We now start investigating several properties of the Lebesgue integral. First of all, we state that the Lebesgue integral extends the notion of the Riemann integral. The proof of this result requires theorems that will be proved next class, and therefore we will postpone it.

Theorem 12.31. *Let $f : R \to \mathbb{R}$ be Riemann integrable, where $R \subset \mathbb{R}^N$ is a rectangle. Then, f is Lebesgue integrable, and the two integrals coincide.*

An important property is that the Lebesgue integral does not *see* the difference of two functions on $\mathcal{L}^N$-negligible sets. The proof follows directly from the definition of the Lebesgue integral, since we are requiring the simple functions to approximate from below $\mathcal{L}^N$-almost everywhere.

Lemma 12.32. *Let $E \subset \mathbb{R}^N$ be an $\mathcal{L}^N$-negligible set, and let $f : \mathbb{R}^N \to \mathbb{R}$ be a Lebesgue measurable function. Then,*

$$\int_E f(x) \, \mathrm{d}\mathcal{L}^N(x) = 0.$$

In particular, if $g : \mathbb{R}^N \to \overline{\mathbb{R}}$ is an $\mathcal{L}^N$-measurable function such that

$$\mathcal{L}^N\left(\{x \in \mathbb{R}^N : f(x) \neq g(x)\}\right) = 0,$$

then,

$$\int_{\mathbb{R}^N} f(x)\, d\mathcal{L}^N(x) = \int_{\mathbb{R}^N} g(x)\, d\mathcal{L}^N(x).$$

In the same spirit, it is possible to prove (left as an exercise to the reader), that a non-negative Lebesgue measurable function with zero integral must be zero $\mathcal{L}^N$-almost everywhere.

Lemma 12.33. *Let $f : \mathbb{R}^N \to [0, \infty]$ be a Lebesgue measurable function such that*

$$\int_{\mathbb{R}^N} f(x)\, d\mathcal{L}^N(x) = 0.$$

Then, $f(x) = 0$ for $\mathcal{L}^N$-almost every $x \in \mathbb{R}^N$.

The Lebesgue integral satisfies some basic properties of monotonicity and homogeneity. The proof follows directly from the definition.

Lemma 12.34. *Let $f, g : \mathbb{R}^N \to \overline{\mathbb{R}}$ be $\mathcal{L}^N$-integrable. Then,*

$$\int_{\mathbb{R}^N} af(x)\ \mathrm{d}\mathcal{L}^N(x) = a \int_{\mathbb{R}^N} f(x)\ \mathrm{d}\mathcal{L}^N(x),$$

for all $a \in \mathbb{R}$. Moreover, if $f \leq g$ $\mathcal{L}^N$-a.e., then

$$\int_{\mathbb{R}^N} f(x)\ \mathrm{d}\mathcal{L}^N(x) \leq \int_{\mathbb{R}^N} g(x)\ \mathrm{d}\mathcal{L}^N(x),.$$

Finally, if $A \subset B$ are two Lebesgue measurable sets, and $f \geq 0$, then

$$\int_A f(x)\ \mathrm{d}\mathcal{L}^N(x) \leq \int_B f(x)\ \mathrm{d}\mathcal{L}^N(x).$$

Now, we consider what happens to the Lebesgue integral of a monotone sequence of simple functions converging to a limiting function (see Lemma 12.20). Such a result will later be extended to a general sequence of monotone functions (this is indeed called the *Lebesgue Monotone Convergence Theorem*, see Theorem 13.4). The reason why we use this path is because it allows to see the proof of such result in two different ways.

Proposition 12.35. *Let $f : \mathbb{R}^N \to [0, +\infty]$ be a Lebesgue measurable, and let $(f_n)_{n\in\mathbb{N}}$ be a sequence of simple functions such that $f_n \to f$ pointwise. Then,*

$$\lim_{n\to\infty} \int_{\mathbb{R}^N} f_n(x)\ \mathrm{d}\mathcal{L}^N(x) = \int_{\mathbb{R}^N} f(x)\ \mathrm{d}\mathcal{L}^N(x).$$

Proof. First of all, note that since the sequence f_n is increasing, then, by using the monotonicity of the Lebesgue integral (see Lemma 12.34), we get that

$$\int_{\mathbb{R}^N} f_{n+1}(x)\ \mathrm{d}\mathcal{L}^N(x) \geq \int_{\mathbb{R}^N} f_n(x)\ \mathrm{d}\mathcal{L}^N(x),$$

for all $n \in \mathbb{N}$. Therefore, the limit

$$\lim_{n\to\infty} \int_{\mathbb{R}^N} f_n(x)\ \mathrm{d}\mathcal{L}^N(x) = \sup_{n\in\mathbb{N}} \int_{\mathbb{R}^N} f_n(x)\ \mathrm{d}\mathcal{L}^N(x)$$

exists.

Since $f_n \leq f$ for all $n \in \mathbb{N}$, by using again the monotonicity of the Lebesgue integral (see Lemma 12.34), we get that

$$\lim_{n\to\infty} \int_{\mathbb{R}^N} f_n(x)\ \mathrm{d}\mathcal{L}^N(x) = \int_{\mathbb{R}^N} f(x)\ \mathrm{d}\mathcal{L}^N(x).$$

To prove the opposite inequality, the idea is the following. Let $g : \mathbb{R}^N \to \mathbb{R}$ be a simple function with $g \leq f$, whose Lebesgue integral is close to the Lebesgue integral of f (in case this latter is $+\infty$, we mean that it is very large). Since $(f_n)_{n\in\mathbb{N}}$ converges to f pointwise, for n large enough, we expect $f_n \geq g$. Since this happens for all such functions g, we get the desired inequality. The

only technical issue to take care of, is that it might be that $g = f$ on some regions, and thus, it might be that $f_n < g$ in that region. Thus, we need to lower a bit g, in order to avoid such situation.

Fix $\lambda \in (0,1)$. Let $g : \mathbb{R}^N \to \mathbb{R}$ be a simple function

$$g(x) = \sum_{i=1}^{k} \mathbb{1}_{E_i}(x) y_i,$$

with $g \leq f$. For $m \in \mathbb{N}$, we want to consider the sets where $\lambda g \leq f_m$, namely

$$F_m := \{x \in \mathbb{R}^N \,:\, \lambda g(x) \leq f_m(x)\}.$$

Note that the set F_m is Lebesgue measurable, thanks to Lemma 12.16. Set

$$g_m := \sum_{i=1}^{k} \mathbb{1}_{E_i \cap F_m}(x) \lambda y_i.$$

Note that, since the sequence $(f_n)_{n\in\mathbb{N}}$ is increasing, we have that $g_m \leq f_n$, for all $n \geq m$. In particular, by using the monotonicity of the Lebesgue integral (see Lemma 12.34), we get that

$$\int_{\mathbb{R}^N} g_m(x) \, \mathrm{d}\mathcal{L}^N(x) \leq \lim_{n\to\infty} \int_{\mathbb{R}^N} f_n(x) \, \mathrm{d}\mathcal{L}^N(x). \tag{12.2}$$

Now, we note that, for m large enough, the set $E_i \cap F_m \neq \emptyset$, for all $i = 1, \dots, k$. This is where we use the fact that $g \leq f$, that $\lambda \in (0,1)$, and that f_n is increasing to f. Moreover, the sequence of Lebesgue measurable sets $E_i \cap F_m$ is increasing to the set E_i, for each $i = 1, \dots, k$. In particular, thanks to Proposition 11.33(ii), it holds

$$\lim_{m\to\infty} \mathcal{L}^N(E_i \cap F_m) = \mathcal{L}^N(E_i). \tag{12.3}$$

Therefore, by using the homogeneity of the Lebesgue integral (see Lemma 12.34), we obtain

$$\begin{aligned}
\lambda \int_{\mathbb{R}^N} g(x) \, \mathrm{d}\mathcal{L}^N(x) &= \int_{\mathbb{R}^N} \lambda g(x) \, \mathrm{d}\mathcal{L}^N(x) \\
&= \sum_{i=1}^{k} \lambda y_i \mathcal{L}^N(E_i) \\
&= \lim_{m\to\infty} \sum_{i=1}^{k} \lambda y_i \mathcal{L}^N(E_i \cap F_m) \\
&= \lim_{m\to\infty} \int_{\mathbb{R}^N} g_m(x) \, \mathrm{d}\mathcal{L}^N(x) \\
&\leq \lim_{n\to\infty} \int_{\mathbb{R}^N} f_n(x) \, \mathrm{d}\mathcal{L}^N(x),
\end{aligned}$$

where in the third equality we used (12.3), while last equality follows from (12.2). Now, since $\lambda \in (0,1)$ is arbitrary, by taking the limit as $\lambda \to 1$ we get

$$\int_{\mathbb{R}^N} g(x) \, \mathrm{d}\mathcal{L}^N(x) \leq \lim_{n\to\infty} \int_{\mathbb{R}^N} f_n(x) \, \mathrm{d}\mathcal{L}^N(x),$$

for all simple function $g : \mathbb{R}^N \to \mathbb{R}$ with $g \leq f$. Thus, the result follows from the definition of the Lebesgue integral of f. □

Remark 12.36. Note that the core strategy of the proof of the above result is to reduce to the case of a increasing sequence of Lebesgue measurable sets, and use the monotonicity properties of the Lebesgue measure.

We are now in position to prove that the Lebesgue integral is linear. We first establish the result for non-negative functions, since for general functions we might get into trouble with $+\infty - \infty$.

Lemma 12.37. *Let $f, g : \mathbb{R}^N \to [0, +\infty]$ be $\mathcal{L}^N$-measurable function. Then,*

$$\int_{\mathbb{R}^N} [f(x) + g(x)] \, \mathrm{d}\mathcal{L}^N(x) = \int_{\mathbb{R}^N} f(x) \, \mathrm{d}\mathcal{L}^N(x) + \int_{\mathbb{R}^N} g(x) \, \mathrm{d}\mathcal{L}^N(x).$$

Proof. Step 1. Assume f and g are both simple functions

$$f(x) = \sum_{i=1}^k \mathbb{1}_{E_i}(x) y_i, \qquad g(x) = \sum_{i=1}^k \mathbb{1}_{F_i}(x) z_i.$$

Note that there is no loss of generality in assuming that the number of sets is the same. Then,

$$f(x) + g(x) = \sum_{i=1}^k \mathbb{1}_{E_i}(x)(y_i + z_i).$$

Therefore, we get that

$$\begin{aligned} \int_{\mathbb{R}^N} [f(x) + g(x)] \, \mathrm{d}\mathcal{L}^N(x) &= \sum_{i=1}^k \mathcal{L}^N(E_i)(y_i + z_i) \\ &= \sum_{i=1}^k \mathcal{L}^N(E_i) y_i + \sum_{i=1}^k \mathcal{L}^N(E_i) z_i \\ &= \int_{\mathbb{R}^N} f(x) \, \mathrm{d}\mathcal{L}^N(x) + \int_{\mathbb{R}^N} g(x) \, \mathrm{d}\mathcal{L}^N(x). \end{aligned}$$

Step 2. In the general case, we use Lemma 12.20 to find increasing sequences $(f_n)_{n\in\mathbb{N}}$ and $(g_n)_{n\in\mathbb{N}}$ of non-negative simple functions such that $f_n \to f$ and $g_n \to g$. Note that $\{f_n + g_n\}_{n\in\mathbb{N}}$ is an increasing sequence of non-negative simple functions such that $f_n + g_n \to f + g$. Then, from step 1, for each $n \in \mathbb{N}$, we get that

$$\int_{\mathbb{R}^N} [f_n(x) + g_n(x)] \, \mathrm{d}\mathcal{L}^N(x) = \int_{\mathbb{R}^N} f_n(x) \, \mathrm{d}\mathcal{L}^N(x) + \int_{\mathbb{R}^N} g_n(x) \, \mathrm{d}\mathcal{L}^N(x).$$

Thus, by using Proposition 12.35, we take the limit on both sides, and get the desired result. □

We now want to focus on those functions for which the Lebesgue integral is finite.

Definition 12.38. We say that an $\mathcal{L}^N$-integrable function $f : \mathbb{R}^N \to \overline{\mathbb{R}}$ belongs to the *Lebesgue space* $L^1(\mathbb{R}^N)$, if

$$\int_{\mathbb{R}^N} |f(x)| \, d\mathcal{L}^N(x) = \int_{\mathbb{R}^N} f^+(x) \, d\mathcal{L}^N(x) + \int_{\mathbb{R}^N} f^-(x) \, d\mathcal{L}^N(x) < \infty.$$

In such a case, we say that f is *absolutely* Lebesgue integrable.

Remark 12.39. Note that this definition is similar to that given for series: a converging series, and an absolute converging series. This is not accidental: indeed, by the theory of abstract integration that you'll see in *Measure Theory*, it is possible to see a series as an integral with respect to a certain measure!

Lemma 12.40. *Let $f : \mathbb{R}^N \to \overline{\mathbb{R}}$ be such that $f \in L^1(\mathbb{R}^N)$. Then,*

$$\mathcal{L}^N\left(\{x \in \mathbb{R}^N : f(x) \in \{\pm\infty\}\}\right) = 0.$$

Namely, f is finite $\mathcal{L}^N$-almost everywhere.

Moreover, it is possible to extend the linearity of the Lebesgue integral to the case of general Lebesgue measurable functions $f, g : \mathbb{R}^N \to \overline{\mathbb{R}}$, provided one of the two is absolutely Lebesgue integrable.

Lemma 12.41. *Let $f, g : \mathbb{R}^N \to \overline{\mathbb{R}}$ be $\mathcal{L}^N$-measurable function, and assume $f \in L^1(\mathbb{R}^N)$. Then,*

$$\int_{\mathbb{R}^N} [f(x) + g(x)] \, \mathrm{d}\mathcal{L}^N(x) = \int_{\mathbb{R}^N} f(x) \, \mathrm{d}\mathcal{L}^N(x) + \int_{\mathbb{R}^N} g(x) \, \mathrm{d}\mathcal{L}^N(x).$$

The proof is left as an exercise to the reader.

The Lebesgue integral satisfies a sort of *continuum* version of the triangle inequality.

Lemma 12.42. *Let $f : \mathbb{R}^N \to \mathbb{R}$ be $\mathcal{L}^N$-measurable. Then,*

$$\left| \int_{\mathbb{R}^N} f(x) \, \mathrm{d}\mathcal{L}^N(x) \right| \leq \int_{\mathbb{R}^N} |f(x)| \, \mathrm{d}\mathcal{L}^N(x).$$

Proof. We have that

$$\begin{aligned} \left| \int_{\mathbb{R}^N} f(x) \, \mathrm{d}\mathcal{L}^N(x) \right| &= \left| \int_{\mathbb{R}^N} f^+(x) \, \mathrm{d}\mathcal{L}^N(x) - \int_{\mathbb{R}^N} f^-(x) \, \mathrm{d}\mathcal{L}^N(x) \right| \\ &\leq \int_{\mathbb{R}^N} f^+(x) \, \mathrm{d}\mathcal{L}^N(x) + \int_{\mathbb{R}^N} f^-(x) \, \mathrm{d}\mathcal{L}^N(x) \\ &= \int_{\mathbb{R}^N} |f(x)| \, \mathrm{d}\mathcal{L}^N(x), \end{aligned}$$

where in the second step we used the inequality $|a - b| \leq a + b$, for $a, b \geq 0$. □

12.3. **The Lebesgue integral for general functions.** Finally, we extend the notion of Lebesgue integral to functions that are not necessarily $\mathcal{L}^N$-integrable. In this case, we need to approximate from above and from below the function with countably-simple functions.

Definition 12.43. Let $f : \mathbb{R}^N \to \overline{\mathbb{R}}$. We define the *lower Lebesgue integral* as

$$\int_{*\,\mathbb{R}^N} f(x) \, \mathrm{d}\mathcal{L}^N(x) \coloneqq \sup \left\{ \int_{\mathbb{R}^N} g(x) \, \mathrm{d}\mathcal{L}^N(x) \, : \, g \text{ simple}, \; g \leq f \;\; \mathcal{L}^N - \text{a.e.} \right\}.$$

Moreover, we define the *upper Lebesgue integral* as

$$\int_{\mathbb{R}^N}^{*} f(x) \, \mathrm{d}\mathcal{L}^N(x) \coloneqq \inf \left\{ \int_{\mathbb{R}^N} g(x) \, \mathrm{d}\mathcal{L}^N(x) \, : \, g \text{ simple}, \, g \geq f \;\; \mathcal{L}^N - \text{a.e.} \right\}.$$

We say that f is *Lebesgue integrable* if the lower and the upper Lebesgue integrals coincide. In this case, we denote the common value by

$$\int_{\mathbb{R}^N} f(x) \, \mathrm{d}\mathcal{L}^N(x).$$

The above definition is consistent with that given for Lebesgue measurable functions. The proof of this statement requires technical results that will be proved next class, and therefore it will be postponed.

Lemma 12.44. *Let $f : \mathbb{R}^N \to \overline{\mathbb{R}}$ be Lebesgue measurable. Then,*

$$\int_{*\,\mathbb{R}^N} f(x) \, \mathrm{d}\mathcal{L}^N(x) = \int_{\mathbb{R}^N}^{*} f(x) \, \mathrm{d}\mathcal{L}^N(x).$$

The same result holds for functions that are non-negative (and not necessarily Lebesgue measurable).

Lemma 12.45. *Let $f : \mathbb{R}^N \to [0, +\infty]$. Then,*

$$\int_{*\,\mathbb{R}^N} f(x) \, \mathrm{d}\mathcal{L}^N(x) = \int_{\mathbb{R}^N}^{*} f(x) \, \mathrm{d}\mathcal{L}^N(x).$$

Proof. If

$$\int_{*\,\mathbb{R}^N} f(x) \, \mathrm{d}\mathcal{L}^N(x) = +\infty,$$

then also

$$\int_{\mathbb{R}^N}^{*} f(x) \, \mathrm{d}\mathcal{L}^N(x) = +\infty,$$

and the result trivially holds.

Therefore, assume the lower Lebesgue integral to be finite. In particular, we have that

$$\mathcal{L}^N\left(\{x \in \mathbb{R}^N : f(x) = +\infty\}\right) = 0.$$

Let $t > 1$, and consider the countably simple function

$$g_t(x) \coloneqq \sum_{k \in \mathbb{Z}} t^k \mathbb{1}_{E_k}(x).$$

where, for each $k \in \mathbb{Z}$, we set

$$E_k \coloneqq \left\{x \in \mathbb{R}^N \,:\, t^k \leq f(x) < t^{k+1}\right\}.$$

By definition, we have that

$$g_t \leq f < t g_t.$$

Therefore, by monotonicity of the lower and the upper Lebesgue integral, we get that

$$\int_{*\ \mathbb{R}^N} f(x)\ \mathrm{d}\mathcal{L}^N(x) \leq \int_{*\ \mathbb{R}^N} t g_t(x)\ \mathrm{d}\mathcal{L}^N(x) = t \int_{*\ \mathbb{R}^N} g_t(x)\ \mathrm{d}\mathcal{L}^N(x) \leq t \int_{\mathbb{R}^N}^* f(x)\ \mathrm{d}\mathcal{L}^N(x),$$

where in the last inequality we used the definition of the lower integral. Since $t > 1$ is arbitrary, by sending $t \to 1$ in the above chain of inequalities we obtain that

$$\int_{*\ \mathbb{R}^N} f(x)\ \mathrm{d}\mathcal{L}^N(x) \leq \int_{\mathbb{R}^N}^* f(x)\ \mathrm{d}\mathcal{L}^N(x).$$

This concludes the proof, since the other inequality holds by definition. □

Remark 12.46. What prevents us from extending the above result to a generic (namely, not necessarily Lebesgue measurable) function $f : \mathbb{R}^N \to \overline{\mathbb{R}}$ is that we do not know whether or not the sets

$$\{f \geq 0\}, \qquad \{f < 0\}$$

are Lebesgue measurable.

We now have extended the Riemann integral to a wider class of functions by using the Lebesgue integral. This enjoys several properties that are extremely useful when dealing with problems in Analysis. We will see next class how to use these properties to prove important results relating the limit of the integral of a sequence of functions, with the integral of the limit of the sequence of functions.

Lebesgue laughs at the difficulty of integration.
Nature

13. Limiting theorems

As we saw, one of the main issues of the Riemann integral was the lack of good behavior with respect to sequences of functions converging pointwise. We already saw in Proposition 12.35 that the Lebesgue integral does not have such an issue when considering an increasing sequence of simple functions. In this section, we investigate deeper the relation between

$$\int_{\mathbb{R}^N} f(x)\,d\mathcal{L}^N(x), \qquad \int_{\mathbb{R}^N} f_n(x)\,d\mathcal{L}^N(x),$$

where f is the liminf, the limsup, or the limit of the sequence $(f_n)_{n\in\mathbb{N}}$. First of all, we want to see when things go wrong. There are three cases where the above two objects are not related by an equality (in the limit). Considering such situations, will allow us to understand what result to expect in a general case, and what additional assumptions to ask on the sequence $(f_n)_{n\in\mathbb{N}}$ in order to get equality.

Consider the following sequences of functions $f_n, g_n, h_n : \mathbb{R} \to [0,\infty)$ defined as

$$f_n \coloneqq \mathbb{1}_{[n,n+1]}, \qquad g_n \coloneqq \frac{1}{n}\mathbb{1}_{[0,n]}, \qquad h_n \coloneqq n\mathbb{1}_{(0,1/n)}.$$

All three functions converge to $f \equiv 0$ pointwise. In particular, g_n converges to f uniformly. Nevertheless, we have that

$$\int_{\mathbb{R}^N} f_n(x)\,d\mathcal{L}^N(x) = \int_{\mathbb{R}^N} g_n(x)\,d\mathcal{L}^N(x) = \int_{\mathbb{R}^N} h_n(x)\,d\mathcal{L}^N(x) = 1,$$

for all $n \in \mathbb{N}$. Therefore, since

$$\int_{\mathbb{R}^N} f(x)\,d\mathcal{L}^N(x) = 0.,$$

there is a *loss of mass* in the limit. What is it due to? Well, in the first case, it is because the mass *goes to infinity horizontally*. In the second case, because it *spreads out horizontally*, while in the latter case because it *concentrates* to a single point. Therefore, it seems that, for a *general sequence of non-negative* functions $(f_n)_{n\in\mathbb{N}}$, we could only expect

$$\int_{\mathbb{R}^N} f(x)\,d\mathcal{L}^N(x) \le \liminf_{n\to\infty} \int_{\mathbb{R}^N} f_n(x)\,d\mathcal{L}^N(x).$$

This is indeed the case. The result is called Fatou's Lemma (see Theorem 13.1). There are two ways to avoid the loss of mass. One is to have an increasing sequence of functions(see Theorem 13.4). The other is to ask the sequence to be bounded from above by an integrable function (see Theorem 13.6). These two results are known as Lebesgue Monotone Convergence, and Lebesgue Dominated Convergence Theorem, respectively.

Before proving the three results, we would like to comment that it is possible to start from any of them, and then prove the other two, in any order. We choose to start from the Fatou's lemma, that allows to get the Lebesgue Monotone and Dominated Convergence Theorems as simple consequences.

13.1. Fatou's Lemma.

We start by investigating the case of a general sequence of functions, for which the pointwise limit does not need to exist. Nevertheless, we can get bounds for the asymptotic behavior of the the sequence of integrals.

Theorem 13.1 (Fatou's lemma)**.** *Let $(f_n)_{n\in\mathbb{N}}$ be a sequence of $\mathcal{L}^N$-measurable functions. If $f_n \ge g$ for all $n \in \mathbb{N}$, where $g \in L^1(\mathbb{R}^N)$, then*

$$\int_{\mathbb{R}^N} \liminf_{n\to\infty} f_n(x)\,d\mathcal{L}^N(x) \le \liminf_{n\to\infty} \int_{\mathbb{R}^N} f_n(x)\,d\mathcal{L}^N(x)\,.$$

If $f_n \le g$ for all $n \in \mathbb{N}$, where $g \in L^1(\mathbb{R}^N)$, then

$$\limsup_{n\to\infty} \int_{\mathbb{R}^N} f_n(x)\,d\mathcal{L}^N(x) \le \int_{\mathbb{R}^N} \limsup_{n\to\infty} f_n(x)\,d\mathcal{L}^N(x).$$

Proof. The strategy we use is the same as that employed in the proof of Proposition 12.35: to consider simple functions, and to reduce to the case of a monotone sequence of Lebesgue measurable sets. We only prove the first part, for

$$f \coloneqq \liminf_{n\to\infty} f_n.$$

The other case follows by using the result for the liminf applied to the sequence of functions $(g - f_n)_{n\in\mathbb{N}}$.

For each $n \in \mathbb{N}$, let $\widetilde{f}_n \coloneqq f_n - g$. Note that $\widetilde{f}_n \geq 0$, and that

$$\widetilde{f} \coloneqq \liminf_{n\to\infty} \widetilde{f}_n = \liminf_{n\to\infty} f_n - g = f - g.$$

Now, by definition we have that

$$\liminf_{n\to\infty} \widetilde{f}_n(x) = \sup_{n\in\mathbb{N}} \inf_{k\geq n} \widetilde{f}_k(x).$$

For each $n \in \mathbb{N}$, define $\varphi_n \coloneqq \inf_{k\geq n} \widetilde{f}_k$. Then, the sequence $(\varphi_n)_{n\in\mathbb{N}}$ is increasing, and converging to $\widetilde{f}$. Fix $\lambda \in (0,1)$. Let $h : \mathbb{R}^N \to \mathbb{R}$ be a simple function

$$h(x) = \sum_{i=1}^{k} \mathbb{1}_{E_i}(x) y_i,$$

with $0 \leq h \leq \widetilde{f}$. Then, by using the fact that $\widetilde{f} \geq 0$, we get that $\lambda h \leq \widetilde{f}$, with equality if and only if they are both zero. For $m \in \mathbb{N}$, set

$$F_m \coloneqq \{x \in \mathbb{R}^N \,:\, \lambda h(x) \leq \varphi_m(x)\}.$$

Note that the set F_m is Lebesgue measurable thanks to Lemma 12.16. Since the sequence $(\varphi_n)_{n\in\mathbb{N}}$ is increasing, we get that $\varphi_n \geq \lambda h$ on F_m, for all $n \geq m$. Moreover, since $\lambda h \leq \widetilde{f}$, we get that $(F_m)_{m\in\mathbb{N}}$ is increasing to $\mathbb{R}^N$. Set

$$h_m \coloneqq \sum_{i=1}^{k} \mathbb{1}_{E_i\cap F_m}(x) \lambda y_i.$$

Note that $h_m \leq \varphi$, for all $n \geq m$. In particular, by using the monotonicity of the Lebesgue integral (see Lemma 12.34), we get that

$$\int_{\mathbb{R}^N} h_m(x)\, \mathrm{d}\mathcal{L}^N(x) \leq \liminf_{n\to\infty} \int_{\mathbb{R}^N} \varphi_n(x)\, \mathrm{d}\mathcal{L}^N(x). \tag{13.1}$$

and that, for each $n \in \mathbb{N}$,

$$\liminf_{n\to\infty} \int_{\mathbb{R}^N} \varphi_n(x)\, \mathrm{d}\mathcal{L}^N(x) \leq \liminf_{n\to\infty} \int_{\mathbb{R}^N} \widetilde{f}_n(x)\, \mathrm{d}\mathcal{L}^N(x). \tag{13.2}$$

This last inequality follows from the definition of the function φ_n, since it is the infimum of a tail of the $\widetilde{f}_k$. Moreover, the sequence of Lebesgue measurable sets $E_i \cap F_m$ is increasing to the set E_i, for each $i = 1, \dots, k$. In particular, thanks to Proposition 11.33(ii), it holds

$$\lim_{m\to\infty} \mathcal{L}^N(E_i \cap F_m) = \mathcal{L}^N(E_i). \tag{13.3}$$

Therefore, by using (13.3), we obtain

$$\begin{aligned}
\lambda \int_{\mathbb{R}^N} h(x)\, \mathrm{d}\mathcal{L}^N(x) &= \int_{\mathbb{R}^N} \lambda h(x)\, \mathrm{d}\mathcal{L}^N(x) \\
&= \sum_{i=1}^{k} \lambda y_i \mathcal{L}^N(E_i) \\
&= \lim_{m\to\infty} \sum_{i=1}^{k} \lambda y_i \mathcal{L}^N(E_i \cap F_m)
\end{aligned}$$

$$= \lim_{m\to\infty} \int_{\mathbb{R}^N} h_m(x)\ \mathrm{d}\mathcal{L}^N(x)$$
$$\leq \liminf_{n\to\infty} \int_{\mathbb{R}^N} \varphi_n(x)\ \mathrm{d}\mathcal{L}^N(x)$$
$$\leq \liminf_{n\to\infty} \int_{\mathbb{R}^N} \widetilde{f}_n(x)\ \mathrm{d}\mathcal{L}^N(x),$$

where the previous to last equality follows from (13.1), while last step from (13.2). Now, since $\lambda \in (0,1)$ is arbitrary, by taking the limit as $\lambda \to 1$ we get

$$\int_{\mathbb{R}^N} h(x)\ \mathrm{d}\mathcal{L}^N(x) \leq \liminf_{n\to\infty} \int_{\mathbb{R}^N} \widetilde{f}_n(x)\ \mathrm{d}\mathcal{L}^N(x)$$
$$= \liminf_{n\to\infty} \int_{\mathbb{R}^N} f_n(x)\ \mathrm{d}\mathcal{L}^N(x) - \int_{\mathbb{R}^N} g(x)\ \mathrm{d}\mathcal{L}^N(x),$$

where last step follows from the linearity of the integral. This inequality holds for all simple function $h : \mathbb{R}^N \to \mathbb{R}$ with $0 \leq h \leq \widetilde{f}$. Thus, from the definition of the Lebesgue integral of $\widetilde{f}$, together with Lemma 12.41, we get

$$\int_{\mathbb{R}^N} f(x)\ \mathrm{d}\mathcal{L}^N(x) - \int_{\mathbb{R}^N} g(x)\ \mathrm{d}\mathcal{L}^N(x) = \int_{\mathbb{R}^N} \widetilde{f}(x)\ \mathrm{d}\mathcal{L}^N(x)$$
$$\leq \liminf_{n\to\infty} \int_{\mathbb{R}^N} f_n(x)\ \mathrm{d}\mathcal{L}^N(x) - \int_{\mathbb{R}^N} g(x)\ \mathrm{d}\mathcal{L}^N(x).$$

Since $g \in L^1(\mathbb{R}^N)$, by simplifying the same terms on both sides, we get the desired result. □

Remark 13.2. It could be that the above inequalities are strict even if $\lim_{n\to\infty} f_n$ exists, as it was shown in the introduction.

Remark 13.3. In particular, if $(f_n)_{n\in\mathbb{N}}$ is a sequence of $\mathcal{L}^N$-measurable functions such that $g \leq f_n \leq h$, for some $g, h \in L^1(\mathbb{R}^N)$, then

$$\int_{\mathbb{R}^N} \liminf_{n\to\infty} f_n(x)\, d\mathcal{L}^N(x) \leq \liminf_{n\to\infty} \int_{\mathbb{R}^N} f_n(x)\, d\mathcal{L}^N(x)$$
$$\leq \limsup_{n\to\infty} \int_{\mathbb{R}^N} f_n(x)\, d\mathcal{L}^N(x) \leq \int_{\mathbb{R}^N} \limsup_{n\to\infty} f_n(x)\, d\mathcal{L}^N(x).$$

This means that the sequence

$$\int_{\mathbb{R}^N} f_n(x)\, d\mathcal{L}^N(x)$$

has bounds from below and from above given by the integrals of $\liminf_{n\to\infty} f_n$ and $\limsup_{n\to\infty} f_n$, respectively.

13.2. **Lebesgue Monotone Convergence Theorem.** A special case of sequences, is that of a monotone sequence. In such a case, the pointwise limit $\lim_{n\to\infty} f_n(x)$ exists for all (or $\mathcal{L}^N$-almost all) $x \in \mathbb{R}^N$.

Theorem 13.4 (Lebesgue's Monotone Convergence Theorem)**.** *Let $(f_n)_{n\in\mathbb{N}}$ be an increasing sequence of $\mathcal{L}^N$-measurable functions such that $f_n \geq g$, for some $g \in L^1(\mathbb{R}^N)$. Then,*

$$\lim_{n\to\infty} \int_{\mathbb{R}^N} f_n(x)\, d\mathcal{L}^N(x) = \int_{\mathbb{R}^N} \lim_{n\to\infty} f_n(x)\, d\mathcal{L}^N(x).$$

The same result holds in the case of a decreasing sequence $(f_n)_{n\in\mathbb{N}}$ of $\mathcal{L}^N$-measurable functions such that $f_n \leq g$ for some $g \in L^1(\mathbb{R}^N)$.

Proof. Assume that the sequence $(f_n)_{n\in\mathbb{N}}$ is increasing. The case where the sequence is decreasing follows by using similar arguments.

We first prove that

$$\lim_{n\to\infty}\int_{\mathbb{R}^N} f_n(x)\,d\mathcal{L}^N(x) \leq \int_{\mathbb{R}^N}\lim_{n\to\infty} f_n(x)\,d\mathcal{L}^N(x).$$

First of all, note that, since the sequence $(f_n)_{n\in\mathbb{N}}$ is increasing, we get that the limit

$$\lim_{n\to\infty} f_n = \sup_{n\in\mathbb{N}} f_n, \tag{13.4}$$

exists, as well as the limit

$$\lim_{n\to\infty}\int_{\mathbb{R}^N} f_n(x)\,d\mathcal{L}^N(x) = \sup_{n\in\mathbb{N}}\int_{\mathbb{R}^N} f_n(x)\,d\mathcal{L}^N(x). \tag{13.5}$$

Then, since

$$f_m \leq \lim_{n\to\infty} f_n = \sup_{n\in\mathbb{N}} f_n,$$

for all $n\in\mathbb{N}$, by using the monotonicity of the Lebesgue integral, we get that

$$\int_{\mathbb{R}^N} f_m(x)\,d\mathcal{L}^N(x) \leq \int_{\mathbb{R}^N}\lim_{n\to\infty} f_n(x)\,d\mathcal{L}^N(x),$$

for each $m\in\mathbb{N}$. Thus,

$$\lim_{n\to\infty}\int_{\mathbb{R}^N} f_n(x)\,d\mathcal{L}^N(x) \leq \int_{\mathbb{R}^N}\lim_{n\to\infty} f_n(x)\,d\mathcal{L}^N(x).$$

To prove the other inequality, we reason as follows. The assumption $f_n \geq g$ with $g\in L^1(\mathbb{R}^N)$ ensures that we can apply the first part of Fatou's Lemma, yielding

$$\begin{aligned}\int_{\mathbb{R}^N}\lim_{n\to\infty} f_n(x)\,d\mathcal{L}^N(x) &= \int_{\mathbb{R}^N}\liminf_{n\to\infty} f_n(x)\,d\mathcal{L}^N(x)\\ &\leq \liminf_{n\to\infty}\int_{\mathbb{R}^N} f_n(x)\,d\mathcal{L}^N(x)\\ &= \lim_{n\to\infty}\int_{\mathbb{R}^N} f_n(x)\,d\mathcal{L}^N(x),\end{aligned}$$

where the first equality follows from (13.4), while last by (13.5). This concludes the proof of the theorem. □

An important application of Lebesgue's Monotone Convergence Theorem is to series of functions.

Corollary 13.5. *Let $(f_n)_{n\in\mathbb{N}}$ be an increasing sequence of $\mathcal{L}^N$-measurable functions $f_n:\mathbb{R}^N\to[0,+\infty)$. Then,*

$$\int_{\mathbb{R}^N}\sum_{n\in\mathbb{N}} f_n(x)\,d\mathcal{L}^N(x) = \sum_{n\in\mathbb{N}}\int_{\mathbb{R}^N} f_n(x)\,d\mathcal{L}^N(x).$$

The proof is left as an exercise to the reader.

13.3. **Lebesgue Dominated Convergence Theorem.** Finally, if the pointwise limit of a sequence of functions is known, but the sequence is not monotone, we wonder whether or not this translates into convergence of the integrals of that sequence. Next important result gives a positive answer under very mild assumptions.

Theorem 13.6 (Lebesgue's Dominated Convergence Theorem)**.** *Let $(f_n)_{n\in\mathbb{N}}$ be a sequence of $\mathcal{L}^N$-measurable functions such that*

$$f_n(x)\to f(x)$$

for $\mathcal{L}^N$-almost every $x\in\mathbb{R}^N$. Assume that

$$|f_n(x)|\leq g(x)$$

for $\mathcal{L}^N$-almost every $x \in \mathbb{R}^N$, where $g \in L^1(\mathbb{R}^N)$. Then, $f \in L^1(\mathbb{R}^N)$, and

$$\lim_{n\to\infty} \int_{\mathbb{R}^N} |f_n(x) - f(x)|\, d\mathcal{L}^N(x) = 0.$$

In particular,

$$\lim_{n\to\infty} \int_{\mathbb{R}^N} f_n(x)\, d\mathcal{L}^N(x) = \int_{\mathbb{R}^N} f(x)\, d\mathcal{L}^N(x).$$

Proof. The proof of the last fact is already contained in Remark 13.3 since, by assumption $-g \le f_n \le g$. On the other hand, we want to prove something more. The idea is to apply the same argument to the sequence of functions $h_n := 2g - |f_n - f|$. Note that, by assumption, $0 \le h_n \le 2g$, and that $h_n \to 2g$ pointwise $\mathcal{L}^N$-almost everywhere. Therefore, by Remark 13.3, we get that

$$\begin{aligned}
\int_{\mathbb{R}^N} 2g(x)\, d\mathcal{L}^N(x) &= \lim_{n\to\infty} \int_{\mathbb{R}^N} h_n(x)\, d\mathcal{L}^N(x) \\
&= \lim_{n\to\infty} \int_{\mathbb{R}^N} [2g(x) - |f_n(x) - f(x)|]\, d\mathcal{L}^N(x) \\
&= \lim_{n\to\infty} \left[\int_{\mathbb{R}^N} 2g(x) - \int_{\mathbb{R}^N} |f_n(x) - f(x)|\, d\mathcal{L}^N(x) \right] \\
&= \int_{\mathbb{R}^N} 2g(x)\, d\mathcal{L}^N(x) - \lim_{n\to\infty} \int_{\mathbb{R}^N} |f_n(x) - f(x)|\, d\mathcal{L}^N(x),
\end{aligned}$$

where in the third step we used the linearity of the integral, since $2g \in L^1(\mathbb{R}^N)$ (see Lemma 12.41). This gives the desired result, since the term on the left-hand side is finite, and thus we can simplify it with that on the right-hand side. □

Remark 13.7. Note that if $(f_n)_{n\in\mathbb{N}}$ is a sequence of Lebesgue measurable functions, the fact that

$$\lim_{n\to\infty} \int_{\mathbb{R}^N} f_n(x)\, d\mathcal{L}^N(x) = \int_{\mathbb{R}^N} f(x)\, d\mathcal{L}^N(x)$$

does *not* imply that

$$\lim_{n\to\infty} \int_{\mathbb{R}^N} |f_n(x) - f(x)|\, d\mathcal{L}^N(x) = 0,$$

not even if $f_n \to f$ pointwise. Find a counterexample!

Remark 13.8. What the above theorem is saying is that the pointwise convergence of a sequence of Lebesgue measurable functions uniformly bounded by an L^1 function implies that the sequence actually converge in the L^1 norm (the 1-Minkowski norm for functions, see Example 1.13).

In view of the Fatou's Lemma, and the Lebesgue Dominated Convergence Theorem, it is interesting to understand how to quantify the loss of mass.

Lemma 13.9. *Let $(f_n)_{n\in\mathbb{N}}$ be a sequence of functions $f_n \in L^1(\mathbb{R}^N)$ with $f_n \ge g$ $\mathcal{L}^N$-almost everywhere for all $n \in \mathbb{N}$, where $g \in L^1(\mathbb{R}^N)$. Assume that $f_n \to f$ pointwise $\mathcal{L}^N$-almost everywhere, where $f \in L^1(\mathbb{R}^N)$. Then,*

$$\lim_{n\to\infty} \left[\int_{\mathbb{R}^N} f_n(x)\, d\mathcal{L}^N(x) - \int_{\mathbb{R}^N} f(x)\, d\mathcal{L}^N(x) - \int_{\mathbb{R}^N} |f_n(x) - f(x)|\, d\mathcal{L}^N(x) \right] = 0.$$

The proof is left as an exercise to the reader.

13.4. **Lebesgue integral as extension of the Riemann integral.** We now have the technical tools to prove that the Lebesgue integral extends the Riemann integral.

Proof of Theorem 12.31. By assumption, the function $f : R \to \mathbb{R}$ is Riemann integrable. Therefore, it is possible to find an increasing sequence of functions $(l_n)_{n\in\mathbb{N}}$, and a decreasing sequence of functions $(u_n)_{n\in\mathbb{N}}$, with $l_n \leq f \leq u_n$, each u_n and each l_n are constants on pluri-rectangles, and

$$\lim_{n\to\infty} \int_R u_n(x)\,dx = \lim_{n\to\infty} \int_R l_n(x)\,dx = \int_R f(x)\,dx, \tag{13.6}$$

where the integrals are the Riemann integrals (note the notation dx). Each function u_n and l_n is Lebesgue measurable. Let

$$u := \inf_{n\in\mathbb{N}} u_n, \qquad l := \sup_{n\in\mathbb{N}} l_n.$$

By using Proposition 12.14, we get that u and l are Lebesgue measurable. Since the sequence $(u_n - l_n)_{n\in\mathbb{N}}$ is decreasing, by using the Monotone Convergence Theorem (see Theorem 13.4), we get that

$$\begin{aligned}\int_R (u(x) - l(x))\,d\mathcal{L}^N(x) &= \lim_{n\to\infty} \int_R (u_n(x) - l_n(x))\,d\mathcal{L}^N(x)\\ &= \lim_{n\to\infty} \int_R (u_n(x) - l_n(x))\,dx\\ &= 0,\end{aligned}$$

where the second step follows from the fact that the Riemann and the Lebesgue integral coincides for functions that are constants on pluri-rectangles, while last equality follows from (13.6). Since $u - l \geq 0$, by Lemma 12.33 we get that $u(x) = l(x)$ for $\mathcal{L}^N$-almost every $x \in R$. Since $l_n \leq f \leq u_n$, this implies that $u(x) = l(x) = f(x)$ for $\mathcal{L}^N$-almost every $x \in R$. Thus, f is Lebesgue measurable, since it coincides $\mathcal{L}^N$-almost everywhere with a Lebesgue measurable function. Thus, using again Theorem 13.4, we get

$$\int_R f(x)\,d\mathcal{L}^N(x) = \lim_{n\to\infty} \int_R u_n(x)\,d\mathcal{L}^N(x) = \lim_{n\to\infty} \int_R u_n(x)\,dx = \int_R f(x)\,dx,$$

where last equality follows from (13.6). □

Remark 13.10. We recall that a function that is Lebesgue integrable is not necessarily Riemann integrable (as an example, the Dirichlet function).

We end this section with a question. As we saw in Chapter 10, the Riemann integral was originally introduced as the anti-derivative. Then, Darboux gave it a geometric interpretation, that allowed to extend the Riemann integral to functions defined on higher dimensional spaces. Moreover, that was the turning point in order to understand how to better extended the elementary notion of area/volume to build a geometric theory of integration that overcomes the limitations of the Riemann integral when dealing with limits of sequences of functions. Now is the question: what happened to the idea of the integral as anti-derivative? Namely, does the Fundamental Theorem of Calculus holds for the Lebesgue integral? So, is it true that

$$\lim_{h\to 0} \frac{1}{h} \int_{[a,x_0+h]} f(x)\,d\mathcal{L}^1(x) = f(x_0), \tag{13.7}$$

whenever $f : \mathbb{R} \to \mathbb{R}$ is Lebesgue integrable? Well, that for sure does not hold for *all* points! Indeed, consider the function $f := [0, 1]$, and $x_0 = 0$. Then, the above limit does not exist, since for $h < 0$ the above quantity is zero, while for $h > 0$ it is one. Even more dramatically, assume that we can find a function $f : \mathbb{R} \to \mathbb{R}$ and a point x_0 such that (13.7) holds. Then, by Lemma 12.32, we know that, if we change f on a set of measure zero, the integral does not change. In particular, let

$$g(x) := \begin{cases} f(x) & \text{if } x \neq x_0,\\ y_0 & \text{if } x = x_0,\end{cases}$$

where $y_0 \neq f(x_0)$. Then,

$$f(x_0) = \lim_{h\to 0} \frac{1}{h} \int_{[a,x_0+h]} f(x)\, d\mathcal{L}^1(x) = \lim_{h\to 0} \frac{1}{h} \int_{[a,x_0+h]} g(x)\, d\mathcal{L}^1(x) \neq g(x_0).$$

Thus, the choice of a function that is equal $\mathcal{L}^N$-almost everywhere to f might change the validity of (13.7). This is something really annoying, since the left-hand side of (13.7) would be the same. Therefore, it seems that everything is lost, since it is not even possible to talk about the *pointwise value* of a function $f \in L^1(\mathbb{R})$.

Luckily for us, Lebesgue comes to the rescue once again! The results can be interpreted both as a Fundamental Theorem of Calculus in the case of $f : \mathbb{R} \to \mathbb{R}$, as well as a way to define the *pointwise value* of a function $f \in \mathbb{R}^N \to \mathbb{R}$ for $\mathcal{L}^N$-almost every point.

Theorem 13.11 (Lebesgue Differentiation Theorem)**.** *Let $f \in L^1(\mathbb{R}^N)$. Then, the limit*

$$\lim_{r\to 0} \frac{1}{\mathcal{L}^N(B(x,r))} \int_{B(x,r)} f(y)\, d\mathcal{L}^N(y) = f(x), \tag{13.8}$$

for $\mathcal{L}^N$-almost every $x \in \mathbb{R}^N$.

The proof of such a result requires sophisticated techniques that are beyond the scope of the course.

Remark 13.12. Note that the right-hand side of (13.8) is the *average* of f in the ball $B(x,r)$. The Lebesgue Differentiation Theorem says that the pointwise average exists for $\mathcal{L}^N$-almost every $x \in \mathbb{R}^N$, and that it is equal to the value of the function at that point. Since the left-hand side is the same for any $g \in L^1(\mathbb{R}^N)$ with $g = f$ $\mathcal{L}^N$-almost everywhere, this is a way to define the *pointwise value* of f $\mathcal{L}^N$-almost everywhere. Namely, we have a good representative of f that behaves well with respect to pointwise averages.